Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–29

A Tree-based Bayesian Accelerated Failure Time Cure Model for Estimating Heterogeneous Treatment Effect[*]

Rongqian Sun[†] and Xinyuan Song[‡,§]

Abstract. Estimating heterogeneous treatment effects has drawn increasing attention in medical studies, considering that patients with divergent features can undergo a different progression of disease even with identical treatment. Such heterogeneity can co-occur with a cured fraction for biomedical studies with a time-to-event outcome and further complicates the quantification of treatment effects. This study considers a joint framework of Bayesian causal forest and accelerated failure time cure model to capture the cured proportion and treatment effect heterogeneity through three separate Bayesian additive regression trees. Under the potential outcomes framework, conditional and sample average treatment effects within the uncured subgroup are derived on the scale of log survival time subject to right-censoring, and treatment effects on the scale of survival probability are derived for each individual. Bayesian backfitting Markov chain Monte Carlo algorithm with the Gibbs sampler is conducted to estimate the causal effects. Simulation studies show the satisfactory performance of the proposed method. The proposed model is then applied to a breast cancer dataset extracted from the SEER database to demonstrate its usage in detecting heterogeneous treatment effects and cured subgroups. Combined with popular mitigation strategies, the proposed method can also alleviate confounding induced by immortal time bias.

Keywords: Bayesian additive regression trees, cured subgroup, heterogeneous treatment effect, nonparametric methods, right-censored survival outcome.

1 Introduction

Clinical trials and observational studies with a time-to-event outcome have gained rapid development in terms of statistical analysis from at least two aspects: causal inference and prediction models. In the former case, the treatment effect on the scale of hazard, survival time, or survival probability has been constructed with causal interpretations through regression models or G-formula under the counterfactual framework, while mainly focusing on the average treatment effect (ATE) with a default homogeneity assumption embedded (Andersen et al., 2017; Gran et al., 2010). Estimation of nuisance functions is often sacrificed for the facility of interpretation. For the latter case, on the contrary, flexible nonparametric and machine learning approaches have been developed

[*] This research was fully supported by GRF Grants (14303622, 14302220) from Research Grant Council of the Hong Kong Special Administration Region.

[†] Department of Statistics, Chinese University of Hong Kong, Hong Kong, sunrq@link.cuhk.edu.hk

[‡] Department of Statistics, Chinese University of Hong Kong, Hong Kong, xysong@sta.cuhk.edu.hk

[§] To whom correspondence should be addressed.

https://doi.org/10.1214/23-BA1402

to capture the unknown relationship between risk factors and the time-to-event outcome and reach satisfactory prediction performance without any parametric or linear assumptions (Ishwaran et al., 2008; Sparapani et al., 2016; Steingrimsson and Morrison, 2020). The two directions were almost developed in parallel, while the increasing popularity of personalized medicine in recent years is drawing them together to address heterogeneity in treatment effects.

Considering that patients with divergent features can go through a different progression of disease even with identical treatment, a rising number of causal ensemble methods have been proposed to explore individualized treatment effects under the context of survival analysis, including regression-tree based subgroup analysis (Foster et al., 2011), causal survival forest (Cui et al., 2020), and Bayesian additive regression trees (BART, Chipman et al., 2010)-based approaches. BART is a nonparametric sum-of-trees prestigious for its flexibility in fitting a complex nonlinear regression surface, where each single tree is penalized as a weak learner through regularization priors and thereby avoids overfitting. It requires neither a prespecified functional form nor rescaling of the predictors but possesses excellent out-of-sample prediction performance and automatically ranks the importance of predictors. BART was first introduced to causal inference to address the heterogeneity in treatment effect on a continuous outcome by flexibly estimating the covariate-specific conditional average treatment effect (CATE) (Hill, 2011), and consecutively found as the best-performing methods in the Atlantic causal inference data analysis challenge (Dorie et al., 2019). It was further extended to survival analysis with a binary outcome for decision making (Logan et al., 2019). A recent work of Henderson et al. (2020) innovatively combined the accelerated failure time (AFT) model (Wei, 1992) with BART to construct individualized treatment effect on the scale of survival time and has been found to outperform a series of popular black-box models according to a recent overview on causal machine learning for survival analysis (Hu et al., 2021). However, this approach regards the treatment indicator as just another covariate and is subject to two inherent weaknesses in heterogeneous treatment effect estimation as pointed by Hahn et al. (2020): strong confounding and vague regularization imposed on treatment effect. The former one can easily happen in the existence of high-dimensional nuisance parameters, where regularization may falsely attribute certain confounding to the causal effect (Hahn et al., 2018). Targeted selection, e.g., doctors may assign patients with better prognoses to a treatment group, can also lead to this problem. The latter one follows from the fact that the original BART prior is not tailored for CATE estimation despite its efficacious for good out-of-sample prediction. Although the first problem can be tackled easily by including the estimated propensity score as an additional covariate, the second is intrinsic for BART.

Besides, biomedical studies with specific interventions and a time-to-event outcome sometimes confront particular problems such as patient compliance and the existence of a cured fraction, which may lead to bias in treatment effect estimation. For the former, units who disobey the treatment assignment fail to measure the actual efficacy of the treatment, towards which Yu et al. (2015) proposed a solution by constructing treatment effect within the compliers subgroup through a semiparametric transformation model with mixture components. For the latter, there can exist a proportion of “long-term survivors” or non-susceptible subjects who will never experience the event of interest, like disease-caused death or recurrence of cancer, even if they are followed abundantly long (see, e.g., Conlon et al., 2014; Lambert et al., 2007; Rutqvist et al., 1984; Othus et al., 2012). A common way to determine whether such a proportion might exist in a survival dataset is to check the survival curves. If the curve levels off after a certain time point and reaches a plateau at the end of the follow-up, a cure model may be appropriate for analysis. Such data can be viewed as a mixture of cured and uncured subjects, where the cured fraction corresponds to the proportion of long-term survivors. Handling the data in a conventional way that assumes all subjects eventually experience the event once the follow-up is adequately long can result in severe bias in parameter estimation (Othus et al., 2017). To derive treatment effect under such circumstances, Gao and Zheng (2017) considered the difference of cured rate between treatment and control groups through a semiparametric transformation model. Zhou and Song (2021) proposed a multiple-mediator structure with Cox mixture cure model and regarded the cure group label as a special mediator. However, both studies assumed homogeneous treatment effects, although heterogeneity and the existence of a cured fraction can co-occur in real-world medical studies.

Another source of confounding faced by observational studies with survival data in general is the immortal time bias (ITB), which occurs when there is a period of time during which patients assigned to the treated group cannot experience the event of interest. For instance, in pharmacoepidemiologic studies, treatment may be prescribed with a delay after diagnosis, and subjects must remain event-free until the actual start of treatment to be identified as treated (Suissa, 2008). Ignoring such an unexposed period in study design can create an artificial survival advantage for the treated group and lead to overestimation of the treatment effect. The issue of ITB has been recognized in the survival literature since the 1970s and been accounted for within the causal inference framework in the past decade. Various mitigation strategies have been developed to alleviate this specific source of confounding. A naïve approach is to exclude the immortal time by redefining time zero as the actual start of treatment for subjects in the treated group (Liu et al., 2012; Mi et al., 2013). Zhou et al. (2005) proposed imputing the missing immortal times for subjects in the control group based on the observed ones in the treated group to ensure the same distribution of immortal times across treatment arms. More recent advances include modeling time-dependent treatment indicator and covariates for longitudinal survival data (Andersen et al., 2021; Karim et al., 2016); emulating target trials to align the eligibility criteria, treatment assignment, and time zero for the exposure of interest (Hernán et al., 2016); and using sequential approaches to stratify the time intervals, include only unexposed subjects within the corresponding interval into analysis, and pool the results of each interval to estimate the overall treatment effect (Mansournia et al., 2021). Nevertheless, as it is inherently challenging to address all sources of confounding simultaneously, these mitigation strategies focus primarily on ITB and do not fully consider treatment effect heterogeneity raised by individual-specific features.

Motivated by these problems, we propose an adaptation of Bayesian causal forest (BCF, Hahn et al., 2020) combined with AFT mixture cure model in this study to accommodate heterogeneous treatment effect estimation for time-to-event data in the presence of a cured fraction. Three separate BART are introduced to model the individual-specific cured probability, the unknown confounding raised by pre-treatment covariates imbalance, and the targeted CATE for the uncured subgroup. Combining with the probit BART that captures the cured rate, the BCF-based AFT model addresses the aforementioned problems through a regularization prior tailored for CATE within the uncured subgroup and enables a straightforward derivation of treatment effect on the scale of log survival time and survival probability. By implementing the Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm (Hastie and Tibshirani, 2000), we obtain the posterior distribution of the causal estimands directly instead of altering the treatment indicator and running the iteration again as the current BART-based approach does, which turns out more efficient. Once the patient-specific CATE within the uncured subgroup is obtained, we further explore their posterior distributions to check the proportion of patients who possess a CATE that deviates a lot from the average level and patients with extremely high or low chance to benefit from the treatment. The detailed pattern of how certain covariates induce such heterogeneity is further manifested through partial dependence plot (Friedman, 2001), a popular visualization tool for causal interpretations of black-box predictive models (Zhao and Hastie, 2021). The proposed method contributes primarily to quantifying heterogeneous treatment effect on a survival outcome in the presence of confounding raised by targeted selection, covariate imbalance, and ignorance of a cured fraction. However, in cases where ITB is likely to induce additional confounding, it is also straightforward to combine the proposed model with some matching-based strategies (see, e.g., Wang et al., 2022) to mitigate ITB in a two-stage way.

The rest of this article is organized as follows. Section 2 introduces the proposed model and derivation of the causal estimands. Section 3 presents the Bayesian estimation procedure with prior specification and posterior inference. Section 4 evaluates the empirical performance of the proposed model through simulation studies. Section 5 applies the proposed method to a dataset extracted from the Surveillance, Epidemiology, and End Results (SEER) database to further demonstrate its usage in detecting heterogeneous treatment effects with a possible cured fraction, and in mitigating the impact of ITB through a two-stage approach. Section 6 concludes the article. Technical details and additional simulation results are presented in the Supplementary Material (Sun and Song, 2023).

2 Model description

2.1 Overview of BART

Let $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_p)^T$ be a $p \times 1$ vector of explanatory variables and Y be the response variable. BART treats the unknown regression function of Y on $\mathbf{x}$ as an ensemble of J binary trees as follows:

$$E(Y \mid \mathbf{x}) \triangleq f(\mathbf{x}) = \sum_{j=1}^{J} g(\mathbf{x}; \mathcal{T}_j, \mathcal{M}_j), \tag{2.1}$$

where $f(\mathbf{x})$ is the true unknown regression function, $\mathcal{T}_j$ denotes the structure of the jth binary tree composed of a set of internal nodes with splitting rules and b_j terminal

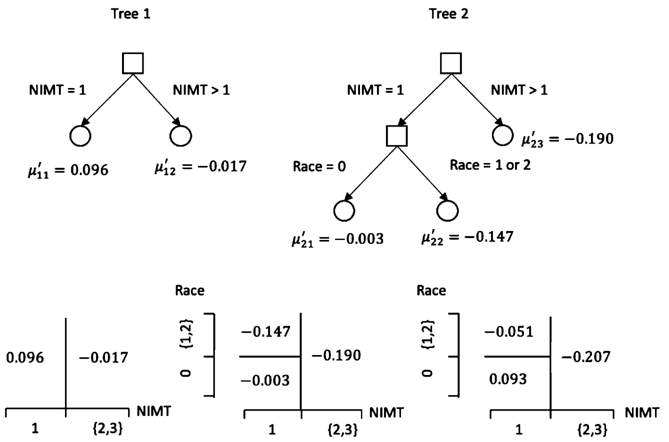


Figure 1: Two binary trees (top left and right) and how each of them as well as their sum partitions the covariate space (bottom left, middle, and right, respectively) in one iteration of MCMC steps for $\tau(\mathbf{x}_i)$ in the analysis of SEER breast cancer data. The two splitting variables are total number of in situ/malignant tumors (NIMT $\in \{1, 2, 3\}$) and race (0 = white, 1 = black, 2 = others).

nodes, and $\mathcal{M}_j = (\mu_{j1}, \ldots, \mu_{jb_j})^T$ denotes the vector of parameter values assigned to the terminal nodes. The splitting rules of each internal node are of the form $\{\mathrm{x}_k \leq c\}$ vs. $\{\mathrm{x}_k > c\}$ with x_k ($k \in \{1, \ldots, p\}$) being the kth component of $\mathbf{x}$, while the top-down sequence of all splitting rules of $\mathcal{T}_j$, as a whole, partitions the original covariate space $\mathbb{R}^p$ into b_j subsets represented by the terminal nodes. $g(\mathbf{x}; \mathcal{T}_j, \mathcal{M}_j)$ denotes the function that assigns node parameter $\mu_{jl} \in \mathcal{M}_j$ to $\mathbf{x}$ suppose that it is allocated to the lth ($l = 1, \ldots, b_j$) terminal node of $\mathcal{T}_j$ according to the above rules. Note that a specific $\mathbf{x}$ can be placed to one unique terminal node within each tree $\mathcal{T}_j$, and $E(Y \mid \mathbf{x})$ is thereby expressed as a sum of the corresponding μ_{jl}s over the J trees. Omitting the intercept term from (2.1) requires some form of centering for the response but is typical for a majority of BART-based algorithms. Figure 1 visualizes the structure of two simple binary trees and the way covariate space is partitioned by their sum. Through such partitioning, the tree-based regression repetitively splits the original dataset into more homogeneous subsets and then fits a piecewise constant function with respect to this partition. To put it to the extreme, a single binary tree can even fit the training data perfectly through a one-observation-per-leaf structure as long as it grows deep enough, but a severe overfitting problem follows. BART, instead, turns this single deep tree into a sum of shallow trees with carefully designed regularization (or roughly speaking, substitutes the single complex piecewise constant function with a sum of simpler ones).

2.2 Tree-based Bayesian accelerated failure time cure model

We consider an observational study with two-arm treatment and right-censored time-to-event outcome with a cured fraction. For subject $i = 1, \ldots, n$, let $\mathbf{x}_i$ be the vector of pre-treatment covariates, T_i and C_i be the failure time of interest and censoring time, respectively, $\delta_i = I(T_i < C_i)$ indicate whether the ith subject fails or is censored, and $Y_i = \min(T_i, C_i)$ be the observed time. Let $A_i \in \{0, 1\}$ be the treatment indicator that takes value 1 if subject i is enrolled into the treatment group and 0 otherwise, of which the effect is of our primary interest. Let $G_i \in \{0, 1\}$ be the partially observed group label indicating whether the ith subject belongs to the uncured ($G_i = 1$) or cured ($G_i = 0$) subgroup. For subjects who experience the event, $G_i = 1$ is observed together with $\delta_i = 1$; while for censored subjects with $\delta_i = 0$, G_i remains unobservable. Subjects in the uncured group are viewed as susceptible to the event of interest such that $T_i < \infty$, while subjects in the cured group are viewed as non-susceptible with $T_i = \infty$. The cured rate for subject i is modeled through BART with a probit link function as follows:

$$Pr(G_i = 1 \mid \mathbf{x}_i, A_i) = \Phi(\upsilon_c(\mathbf{x}_i, A_i)) = \Phi\left(\sum_{j_c=1}^{J_c} g_c(\mathbf{x}_i, A_i; \widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c})\right), \tag{2.2}$$

where $\upsilon_c(\mathbf{x}_i, A_i) = \sum_{j_c=1}^{J_c} g_c(\mathbf{x}_i, A_i; \widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c})$ is the mean of the underlying normal random variable modeled as the sum of J_c binary trees $\{(\widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c}), j_c = 1, \dots, J_c\}$ and $g_c(\mathbf{x}_i, A_i; \widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c})$ reflects the implementation of the corresponding partition rules.

For the uncured subjects, an AFT model combined with BCF is further defined to explore the treatment effect on survival time and the evidence of underlying heterogeneity. The distribution of the event time given $G_i = 1$ is determined by

$$\begin{aligned}\log(T_i \mid \mathbf{x}_i, A_i, G_i = 1) &= \upsilon(\mathbf{x}_i) + \tau(\mathbf{w}_i)A_i + \epsilon_i \\ &= \sum_{j=1}^{J} g_1(\mathbf{x}_i; \mathcal{T}_j, \mathcal{M}_j) + \sum_{h=1}^{H} g_2(\mathbf{w}_i; \mathcal{T}'_h, \mathcal{M}'_h)A_i + \epsilon_i,\end{aligned} \tag{2.3}$$

where $\mathbf{w}_i$ is a subvector of $\mathbf{x}_i$ representing covariates that possibly raise heterogeneity of the treatment effect, i.e., the "effect modifiers"; $\upsilon(\mathbf{x}_i) = \sum_{j=1}^{J} g_1(\mathbf{x}_i; \mathcal{T}_j, \mathcal{M}_j)$ and $\tau(\mathbf{w}_i) = \sum_{h=1}^{H} g_2(\mathbf{w}_i; \mathcal{T}'_h, \mathcal{M}'_h)$ are two separate BART that capture the confounding induced by covariate imbalance (also referred to as prognostic effects) and the treatment effect modified by $\mathbf{w}_i$, respectively; $\{(\mathcal{T}_j, \mathcal{M}_j), j = 1, \dots, J\}$ denote the J binary trees comprising $\upsilon(\mathbf{x}_i)$, $\{(\mathcal{T}'_h, \mathcal{M}'_h), h = 1, \dots, H\}$ denote the H binary trees comprising $\tau(\mathbf{w}_i)$, and $g_1(\mathbf{x}; \mathcal{T}_j, \mathcal{M}_j)$ and $g_2(\mathbf{w}; \mathcal{T}'_h, \mathcal{M}'_h)$ are the corresponding node-parameter-allocation functions; ϵ_i is the residual term satisfying $E(\epsilon_i) = 0$.

This BCF structure on the right-hand side of (2.3) directly models CATE for the uncured subjects, explicitly controls the regularization imposed on it, and takes account of the fact that not every covariate serves as confounder and modifier simultaneously. Following the default settings of Chipman et al. (2010) and Hahn et al. (2020), which have been proven to endow effective implementation of BART-based approaches, we fulfill specification of the proposed model through the following priors on $\{(\widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c}), j_c = 1, \dots, J_c\}$, $\{(\mathcal{T}_j, \mathcal{M}_j), j = 1, \dots, J\}$, and $\{(\mathcal{T}'_h, \mathcal{M}'_h), h = 1, \dots, H\}$: (i) the probability that a node at depth d ($d \in \{0, 1, 2, \dots\}$) continues splitting is assumed as $\alpha(1 + d)^{-\beta}$, where $\alpha \in (0, 1)$ and $\beta > 0$ are pre-specified hyperparameters controlling scope of each individual tree; (ii) the splitting variable at each internal node is uniformly chosen from the discrete set of all available variables; (iii) the splitting value at each internal node with known splitting variable is uniformly selected from the discrete set of all available splitting values constructed from the interpolated sample quantiles, and (iv) given $\widetilde{\mathcal{T}}_{j_c}$, $\mathcal{T}_j$, and $\mathcal{T}'_h$, the terminal node parameters $\widetilde{\mu}_{j_c l} \in \widetilde{\mathcal{M}}_{j_c}$, $\mu_{jl} \in \mathcal{M}_j$, and $\mu'_{hl} \in \mathcal{M}'_h$ are assumed with normal prior distributions $N(0, \frac{c_c^2}{J_c k_c^2})$, $N(0, \frac{c_\mu^2}{4Jk_1^2})$, and $N(0, \frac{c_\mu^2}{4Hk_2^2})$, respectively, where $\{c_c, c_\mu,\ k_c,\ k_1,\ k_2\}$ are pre-defined hyperparameters such that substantial probability is assigned within a desirable range. The above regularization priors construct each binary tree as a weak learner that comprises only a small portion of the overall fit and thereby efficiently circumvent overfitting, with the strength and scope of such regularization determined through the hyperparameters $\{\alpha, \beta, c_c, c_\mu, k_c, k_1, k_2\}$. Full details of the original setting can be found in Chipman et al. (2010), and our choices of the hyperparameters are described in Section 3.

Consistent with the default setting of Bayesian tree ensembles in Chipman et al. (2010) and Hahn et al. (2020), $J_c = J = 200$ trees are set for $\upsilon_c(\mathbf{x}_i, A_i)$ and $\upsilon(\mathbf{x}_i)$ while $H = 50$ trees are set for $\tau(\mathbf{w}_i)$, considering that the pattern of treatment effect heterogeneity is usually much simpler than the pattern of confounding. In this way, stronger regularization is imposed on $\tau(\mathbf{w}_i)$, which is indeed the CATE for the uncured subgroup, as will be shown in the next section. Different choices for J_c, J, H, and the hyperparameters are also considered in Section 4 to check the stability of the proposed method under different prior regularizations on heterogeneity. Finally, considering that all elements of $\mathbf{x}_i$ can serve as modifiers of the treatment effect with an equitable chance when lacking domain knowledge, for notation simplicity, we use $\tau(\mathbf{x}_i)$ in substitute of $\tau(\mathbf{w}_i)$ in the rest of the article.

Up to now, the proposed method was introduced under an ideal circumstance that enrollment and treatment initiation concur at certain time zero, but ITB can serve as another source of confounding in causal survival analysis. For example, cancer patients may have to wait for a period of time after diagnosis to receive the actual therapy, and bias can be induced if this period of immortal time is considered as being exposed. Figure S1(a) of the Supplementary Material illustrates the ideal case under which the proposed method would work smoothly, while Figure S1(b) gives a more complex but commonly encountered case with ITB. Given that the proposed method focuses primarily on confounding raised by targeted selection and covariate imbalance and that solving all sources of confounding at once is intrinsically challenging, a tailored solution to ITB is beyond the scope of this article. Nonetheless, it is feasible to combine the proposed method with some popular mitigation strategies to alleviate ITB in a two-stage way (see, Wang et al., 2022, for an overview). Two specific strategies considered in this article are directly excluding the immortal time and Prescription Time Distribution Matching (PTDM, Zhou et al., 2005). For the former, the immortal time is first subtracted from Y_i for each subject i with $A_i = 1$, after which the proposed model is implemented for treatment effect estimation. For the latter, the missing immortal times for subjects with $A_i = 0$ are imputed from the observed ones in the treated group through random sampling with replacement or propensity score matching. The immortal times are then subtracted from the Y_is for every subject in the sample, followed by the implementation of the proposed model. We provide a demonstration of the two-stage approach using the SEER breast cancer data at the end of Section 5.

2.3 Assumptions and causal estimands

We construct causal estimands under the potential outcomes framework (Rubin, 1974, 2005, 1978; Splawa-Neyman et al., 1990). Denote $T_i(1)$ and $T_i(0)$ as the potential survival time of subject i suppose that he/she had been assigned to the treatment and control groups, respectively. Similarly, we let $C_i(1)$ and $C_i(0)$ denote the corresponding potential censoring times, and $G_i(1)$ and $G_i(0)$ denote the potential group indicators under treatment and control, respectively. We focus on three causal estimands based on these counterfactuals: the conditional average survival probability (CASP), CASP$(t, \boldsymbol{x})$,

$$Pr\big(T_i(1) > t \mid \mathbf{x}_i = \boldsymbol{x}\big) - Pr\big(T_i(0) > t \mid \mathbf{x}_i = \boldsymbol{x}\big), \tag{2.4}$$

the uncured conditional average treatment effect (UCATE), UCATE$(\boldsymbol{x})$,

$$E\big\{\log(T_i(1)) \mid \mathbf{x}_i = \boldsymbol{x}, G_i(1) = 1\big\} - E\big\{\log(T_i(0)) \mid \mathbf{x}_i = \boldsymbol{x}, G_i(0) = 1\big\} \tag{2.5}$$

for uncured subjects with $G_i = 1$, and the uncured sample average treatment effect (USATE)

$$\frac{1}{n_1}\sum_{i:G_i=1}\Big[E\big\{\log(T_i(1)) \mid \mathbf{x}_i = \boldsymbol{x}_i, G_i(1) = 1\big\} - E\big\{\log(T_i(0)) \mid \mathbf{x}_i = \boldsymbol{x}_i, G_i(0) = 1\big\}\Big] \tag{2.6}$$

for uncured subjects with $G_i = 1$, where n_1 is the total number of subjects identified as uncured.

The causal estimands defined above characterize treatment effect from two aspects. First, the construction of UCATE and USATE within the uncured group borrows the idea of complier average causal effect (Yu et al., 2015) in a sense that only uncured subjects with finite event time reflect the treatment effect in terms of extending event time, just as only the complier subgroup reflects the true treatment effect. However, unlike the complier class, which can be viewed as independent of the treatment received, the uncured group can vary with a different cured rate under different treatment arms and thus produces four possible cases: (i) $G_i(1) = G_i(0) = 1$; (ii) $G_i(1) = G_i(0) = 0$; (iii) $G_i(1) = 0, G_i(0) = 1$; and (iv) $G_i(1) = 1, G_i(0) = 0$. UCATE and USATE treat uncured subjects with $G_i = 1$ as case (i) to formulate the (conditional) average treatment effect on the scale of the logarithm of survival time and thereby serve as answers to the question "How long will the treatment extend the survival time of a subject if he/she remains susceptible to the event?" It is also worth noticing that case (ii) contributes nothing to quantifying treatment effect on the scale of survival time with $T_i(1) = T_i(0) = \infty$, while case (iv) is comparatively implausible unless the treatment jeopardizes survival by turning subjects non-susceptible to the event into susceptible ones. Second, by excluding case (iv) as suggested by the assumptions below, CASP covers cases (i)–(iii) and captures treatment effect on the scale of the conditional probability of surviving over certain time t with

$$\begin{aligned} Pr\big(T_i(a) > t \mid \mathbf{x}_i = \boldsymbol{x}\big) &= Pr\big(T_i(a) > t \mid \mathbf{x}_i = \boldsymbol{x}, G_i(a) = 1\big) Pr\big(G_i(a) = 1 \mid \mathbf{x}_i = \boldsymbol{x}\big) \\ &\quad + 1 \cdot Pr\big(G_i(a) = 0 \mid \mathbf{x}_i = \boldsymbol{x}\big), \quad a = 0, 1, \end{aligned}$$

which accommodates the treatment effect on improving the cured rate and serves as an answer to a second question "How much more chance could the subject have to survive more than t months/years from the event if he/she receives the treatment?". Causal estimands proposed in this work are tailored for the two questions that subjects are most concerned about in practice. The following common assumptions of the potential outcomes framework (Imbens and Rubin, 1997; Pearl, 1995; Rosenbaum, 1984; Stone, 1993; Yu et al., 2015) are required for identification of the causal estimands:

Assumption 1 (SUTVA)**.** $T_i = A_i T_i(1) + (1 - A_i) T_i(0)$, i.e, treatment assignments for each unit do not interfere and each treatment level defines a unique outcome for each unit. Similarly, $C_i = A_i C_i(1) + (1 - A_i) C_i(0)$, and $G_i = A_i G_i(1) + (1 - A_i) G_i(0)$.

Assumption 2 (Positivity)**.** $0 < Pr(A_i = 1 \mid \mathbf{x}_i) < 1$, i.e, each unit has a positive probability of allocation to either arm of the treatment. Similarly, $0 < Pr(G_i = 1 \mid \mathbf{x}_i) < 1$, i.e., each subject has a positive probability of being cured.

Assumption 3 (Weak unconfoundedness)**.** $G_i(a) \perp\!\!\!\perp A_i \mid \mathbf{x}_i$ for $a = 0, 1$, i.e., there is no unmeasured confounding between treatment and cure rate. "A $\perp\!\!\!\perp$ B" denotes independence between A and B. $T_i(a) \perp\!\!\!\perp A_i \mid \mathbf{x}_i, G_i(a) = 1$ for $a = 0, 1$, i.e, there is no unmeasured confounding between treatment and survival time for uncured subjects.

Assumption 4 (Independent censoring)**.** $C_i(a) \perp\!\!\!\perp T_i(a), G_i(a) \mid \mathbf{x}_i$ for $a \in \{0, 1\}$.

Assumption 5 (Exclusion restriction)**.** $Pr\big(G_i(1) \leq G_i(0) \mid \mathbf{x}_i\big) = 1$.

Under the above assumptions, it is straightforward to see that by viewing uncured subjects as case (i), we have $G_i(1) = G_i(0) = G_i = 1$ and $E\big\{\log(T_i(a)) \mid \mathbf{x}_i = \boldsymbol{x}, G_i(a) = 1\big\} = E\big\{\log(T_i) \mid \mathbf{x}_i = \boldsymbol{x}, A_i = a, G_i = 1\big\}$ for $a \in \{0, 1\}$. The UCATE defined in (2.5) can thereby be expressed as

$$E\big\{\log(T_i) \mid \boldsymbol{x}, 1, 1\big\} - E\big\{\log(T_i) \mid \boldsymbol{x}, 0, 1\big\} = \upsilon(\boldsymbol{x}) + \tau(\boldsymbol{x}) - \upsilon(\boldsymbol{x}) = \tau(\boldsymbol{x}),$$

while the USATE is derived as $\bar{\tau} = \frac{1}{n_1} \sum_{i:G_i=1} \tau(\boldsymbol{x}_i)$. Similarly, with $Pr\big(T_i(a) > t \mid \mathbf{x}_i = \boldsymbol{x}, G_i(a) = 1\big) = Pr(T_i > t \mid \mathbf{x}_i = \boldsymbol{x}, A_i = a, G_i = 1)$ and $Pr\big(G_i(a) = z \mid \mathbf{x}_i\big) = Pr(G_i = z \mid \mathbf{x}_i = \boldsymbol{x}, A_i = a)$ for $a, z \in \{0, 1\}$, CASP defined in (2.4) can be expressed as

$$F_\epsilon\big(\log(t) - \upsilon(\boldsymbol{x})\big)\Phi\big(\upsilon_c(\boldsymbol{x}, 0)\big) - F_\epsilon\big(\log(t) - \upsilon(\boldsymbol{x}) - \tau(\boldsymbol{x})\big)\Phi\big(\upsilon_c(\boldsymbol{x}, 1)\big),$$

where F_ϵ denotes the cumulative distribution function of residual ϵ_i. Considering that the nonparametric tree ensembles are fully capable of capturing nonlinear interrelationships among variables, we assume that $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ for simplicity. In this way, the causal estimands are identified with the regression surface fitted by Models (2.2)–(2.3).

While a more flexible alternative is to model the residual distribution as a location mixture of Gaussian distributions (see, e.g., Henderson et al., 2020; Yang et al., 2010), we found through a pilot simulation study that such a combination of BCF and Dirichlet process prior did not work synergistically in improving point or interval estimation of the treatment effect on the time-to-event outcome. The unnecessary model complexity brought about thereby makes it less preferable to the normal assumption on residual distribution.

3 Bayesian analysis

3.1 Prior specification

Assuming independence among the individual trees and residual variance σ^2, the prior distribution for the proposed model is formulated as

$$
\begin{aligned}
& p\left(\{(\widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c})\}_{j_c=1}^{J_c}, \{(\mathcal{T}_j, \mathcal{M}_j)\}_{j=1}^{J}, \{(\mathcal{T}'_h, \mathcal{M}'_h)\}_{h=1}^{H}, \sigma^2\right) \\
&= \prod_{j_c=1}^{J_c} p(\widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c}) \prod_{j=1}^{J} p(\mathcal{T}_j, \mathcal{M}_j) \prod_{h=1}^{H} p(\mathcal{T}'_h, \mathcal{M}'_h) p(\sigma^2) \\
&= \prod_{j_c=1}^{J_c} \prod_{l=1}^{b_{j_c}} p(\mu_{j_c l} \mid \widetilde{\mathcal{T}}_{j_c}) p(\widetilde{\mathcal{T}}_{j_c}) \prod_{j=1}^{J} \prod_{l=1}^{b_j} p(\mu_{jl} \mid \mathcal{T}_j) p(\mathcal{T}_j) \prod_{h=1}^{H} \prod_{l=1}^{b_l} p(\mu'_{hl} \mid \mathcal{T}'_h) p(\mathcal{T}'_h) p(\sigma^2).
\end{aligned}
$$

Among the hyperparameters $\{\alpha, \beta, c_c, c_\mu, k_c, k_1, k_2\}$, α and β control the height of each individual binary tree through $p(\widetilde{\mathcal{T}}_{j_c})$, $p(\mathcal{T}_j)$, or $p(\mathcal{T}'_h)$, while c_c, c_μ, k_c, k_1, and k_2 confine the prior probability for $\upsilon_c(\mathbf{x})$, $\upsilon(\mathbf{x})$, and $\tau(\mathbf{x})$ as the sum of trees through $p(\mu_{j_c l} \mid \widetilde{\mathcal{T}}_{j_c})$, $p(\mu_{jl} \mid \mathcal{T}_j)$, and $p(\mu'_{hl} \mid \mathcal{T}'_h)$, respectively. Following the default choices suggested by Chipman et al. (2010), of which remarkable efficacy has been shown under a variety of cases, we set $\alpha = 0.95$ and $\beta = 2$ for $\upsilon_c(\mathbf{x})$ and $\upsilon(\mathbf{x})$ such that shallow trees with 2 or 3 terminal nodes are preferred. A stronger regularization with $\alpha' = 0.25$ and $\beta' = 3$ suggested by Hahn et al. (2020) is considered for $\tau(\mathbf{x})$ to evade misidentified heterogeneity in treatment effects. The default choices are also applied to check the stability of the estimation results under different prior knowledge on heterogeneity.

To assign plausible prior distributions for the prognostic and treatment effects, we follow the procedure of Henderson et al. (2020) to first center the observed time Y_i by $Y_i^* = Y_i \exp(-\hat{\mu}_{AFT})$ and set $c_\mu = 4\hat{\sigma}_{AFT}$, where $\hat{\mu}_{AFT}$ and $\hat{\sigma}_{AFT}$ are the intercept and scale estimates of the parametric AFT model fitted with intercept only and log-normal residuals. As a sum of J terminal node parameters, the prognostic effect $\upsilon(\mathbf{x})$ is thereby assigned normal prior $N(0, \frac{4\hat{\sigma}^2_{AFT}}{k_1^2})$ such that the interval $(-\frac{4\hat{\sigma}_{AFT}}{k_1}, \frac{4\hat{\sigma}_{AFT}}{k_1})$ covers around 95% of its prior probability. Similarly, the UCATE $\tau(\mathbf{x})$ is concentrated within $(-\frac{4\hat{\sigma}_{AFT}}{k_2}, \frac{4\hat{\sigma}_{AFT}}{k_2})$ with a prior probability of around 95%, and the choice of $k_1 = 2$ and $k_2 = 4$ assigns approximately 99.6% prior probability of the expected logarithm of survival time under treatment within $\hat{\mu}_{AFT} \pm 3\hat{\sigma}_{AFT}$. For the probit cure model, we follow the suggestion of Tan and Roy (2019) to set $c_c = 3$ and $k_c = 2$, resulting in a plausible range $(-3, 3)$ with a prior probability of 95% for the normal random variable underlying G_i.

For variance of the residual terms, an inverse-chi-squared prior is imposed as $\sigma^2 \sim \kappa\nu/\chi^2_\nu$, where κ and ν are pre-determined hyperparameters. According to the common practice, κ is chosen such that the prior probability of σ^2 larger than some known rough estimate $\hat{\sigma}^2_r$ approximately equals $1-q$, where $q \in (0, 1)$ is a pre-specified constant that adjusts the magnitude of σ^2 relative to $\hat{\sigma}^2_r$. We take the default setting of Chipman et al. (2010) to set $\nu = 3$ and let $\hat{\sigma}_r$ be the scale estimates of the corresponding parametric AFT model with the same set of covariates. Considering that such a parametric model neglects the cured fraction and can thus lead to an overestimated $\hat{\sigma}^2_r$, we set $q = 0.99$ in the sense that σ^2 for the uncured subgroup is very likely no larger than $\hat{\sigma}^2_r$. Besides, a smaller choice of $q = 0.95$ is also applied to check the stability of the estimation results under different prior knowledge on the cured fraction.

3.2 Posterior inference

Considering that Model (2.2) is designed for the cured rate of every subject across the sample while Model (2.3) is proposed for the event time of only the uncured subgroup, we aim to derive the posterior distribution of binary trees $\{(\widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c})\}_{j_c=1}^{J_c}$ based on the entire sample, and that of $\{(\mathcal{T}_j, \mathcal{M}_j)\}_{j=1}^{J}$ and $\{(\mathcal{T}'_h, \mathcal{M}'_h)\}_{h=1}^{H}$ based on the uncured subgroup (across the iterations). By augmenting the observed data with the latent failure times and group labels, full conditional distribution of the binary trees can be derived with the complete-data likelihood and prior distributions in Section 3.1. The Bayesian backfitting MCMC algorithm is employed to sequentially update the individual trees for each BART, combined with the Gibbs sampler to obtain the full conditional distributions of the latent components. Derivation of the posterior distributions and implementation of the algorithm is provided in Web Appendix A of the Supplementary Material.

After discarding the first N_0 burn-in iterations, we collect N posterior draws of $\{(\widetilde{\mathcal{T}}_{j_c}, \widetilde{\mathcal{M}}_{j_c})\}_{j_c=1}^{J_c}$, $\{(\mathcal{T}_j, \mathcal{M}_j)\}_{j=1}^{J}$, and $\{(\mathcal{T}'_h, \mathcal{M}'_h)\}_{h=1}^{H}$ and obtain the Bayesian estimate of the causal estimands and their 95% credible interval. The convergence of the algorithm can be easily checked from the trace plot of σ^2. For every individual, the estimated CASP is given by

$$\begin{aligned}\frac{1}{N}\sum_{itr=N_0+1}^{N_0+N}\Bigg[&\Phi\left(\frac{\log(t)-\upsilon^{(itr)}(\mathbf{x}_i)}{\sigma}\right)\Phi\left(\upsilon_c^{(itr)}(\mathbf{x}_i,0)\right)\\ &-\Phi\left(\frac{\log(t)-\upsilon^{(itr)}(\mathbf{x}_i)-\tau^{(itr)}(\mathbf{x}_i)}{\sigma}\right)\Phi\left(\upsilon_c^{(itr)}(\mathbf{x}_i,1)\right)\Bigg];\end{aligned} \tag{3.1}$$

while for the uncured subgroup, the estimated UCATE is given by

$$\hat{\tau}(\mathbf{x}_i) = \frac{1}{N}\sum_{itr=N_0+1}^{N_0+N}\tau^{(itr)}(\mathbf{x}_i) = \frac{1}{N}\sum_{itr=N_0+1}^{N_0+N} g_2\left(\mathbf{x}_i; \mathcal{T}_h'^{(itr)}, \mathcal{M}_h'^{(itr)}\right),$$

and the estimated USATE is accordingly calculated as

$$\hat{\hat{\tau}} = \frac{\sum_{i=1}^{n} I(\hat{G}_i = 1)\hat{\tau}(\mathbf{x}_i)}{\sum_{i=1}^{n} I(\hat{G}_i = 1)}.$$

Evidence of heterogeneity in treatment effects can thus be verified through assessing the CASP and UCATE for each uncured subject, as Figures 3(a) and 3(b) depict. An alternative approach to quantifying the degree of heterogeneity is the posterior probability of differential treatment effect (Henderson et al., 2020) defined as

$$D_i = Pr\left(\tau(\mathbf{x}_i) \leq \hat{\hat{\tau}} \mid \mathcal{D}\right), \quad D_i^* = \max\{1 - 2D_i, 2D_i - 1\},$$

where either a too large or too small D_i indicates a fair deviation from the USATE for the ith uncured subject. Henderson et al. (2020) suggests $D_i^* > 0.95$ (i.e., $D_i > 0.95$ or $D_i < 0.05$) as strong evidence of individual-specific differential treatment effect while $D_i^* > 0.8$ (i.e., $D_i > 0.90$ or $D_i < 0.10$) as mild evidence, and further regards the proportion of subjects with such strong/mild evidence as a summary measure of heterogeneity. We adopt this procedure in the following simulation and real data analyses to get a sight of the possible heterogeneity. Similarly, deviation from the sample average CASP at certain fixed time t can also be assessed for each individual.

In addition, the effect modifiers will be of major interest once a considerable level of heterogeneity is found in the above way. Based on the collected posterior draws for each binary tree, it is easy to check the frequency of usage for each component of $\mathbf{x}$ in $\upsilon(\mathbf{x})$, $\tau(\mathbf{x})$, and $\upsilon_c(\mathbf{x})$, and thus get the set of "top" confounders and effect modifiers of survival time and modifiers of the cured rate. With the effect modifiers recognized, a natural question arises as to the explicit way they give rise to heterogeneity in the treatment effects. A popular tool to visualize this mechanism is the partial dependence plot, which has been used to obtain causal interpretation for black-box predictive models, including BART with a time-to-event outcome (Henderson et al., 2020). Under Model (2.3), the partial effect of a specific modifier x_l (the lth component of $\mathbf{x}$) on UCATE is defined by

$$\tau_l^{PE}(x) = \frac{1}{n_1} \sum_{i:G_i=1} \tau(x, \mathbf{x}_{i,-l}), \quad l = 1, \ldots, p,$$

i.e., averaging the UCATE over the sample with x_l fixed at certain value x. The estimated partial effect of each recognized modifier is directly obtained through the posterior draws $\{(\mathcal{T}_h', \mathcal{M}_h')\}_{h=1}^{H}$. For a discrete x_l, the posterior density plot of the partial effect at different categories can serve as a visualization tool of the induced heterogeneity, as shown in Figures 4(a) and 4(b).

4 Simulation study

This simulation evaluates the empirical performance of the proposed methodology. We followed the data generating process of Hahn et al. (2020), the groundbreaking work on BCF structure, to achieve a fair comparison between the proposed model accommodating cured fraction and the conventional one-BART structure under the survival context.

To mimic the dimensionality of covariates in the real data analysis, datasets were generated with $p = 8$ covariates, with the first three, the sixth, and the seventh drawn as standard normal random variables, the fourth drawn as a categorical random variable with three levels $\{1, 2, 3\}$, the fifth drawn from Bernoulli distribution with a success probability of 0.5, and the last drawn from the uniform distribution on $(0, 1)$. The true group label was generated by $G_i \sim Bernoulli\big(0.8\Phi(-0.5\mathrm{x}_7+\mathrm{x}_4 A)+0.05+0.1\mathrm{x}_8\big)$ with a cured proportion around 35% or $G_i \sim Bernoulli\big(0.8\Phi(-1-0.5\mathrm{x}_7+\mathrm{x}_4 A)+0.05+0.1\mathrm{x}_8\big)$ with a cured proportion around 50%. For the uncured subgroup, both a homogeneous treatment effect of $\tau(\mathbf{x}) = 1.5$ and a heterogeneous one as $\tau(\mathbf{x}) = 1 + 0.5\mathrm{x}_2 - 0.25\mathrm{x}_2\mathrm{x}_5$ were considered, whereas the prognostic function was assumed with a nonlinear form of $\upsilon(\mathbf{x}) = -1.2 + 0.4I(\mathrm{x}_4 = 1) - 0.2I(\mathrm{x}_4 = 2) - 0.8I(\mathrm{x}_4 = 3) + 0.2|\mathrm{x}_3 - 1|$. The true propensity score for each subject was generated by

$$\pi(\mathbf{x}_i) \triangleq Pr(A_i = 1 \mid \mathbf{x}_i) = 0.8\Phi\left(\frac{3\upsilon(\mathbf{x}_i)}{s_\upsilon} - 0.5\mathrm{x}_{i1}\right) + 0.05 + 0.1\zeta_i, \quad i = 1, \ldots, n,$$

where s_υ is the sample standard deviation of $\upsilon(\mathbf{x})$ and $\zeta_i \overset{i.i.d}{\sim} \text{Uniform}(0, 1)$. Selection bias exists under such a setting since $\pi(\mathbf{x}_i)$ is monotone in $\upsilon(\mathbf{x}_i)$, where x_1, x_3, and x_4 serve as confounders. Contrastingly, x_2 and x_5 are true modifiers of the treatment effect. Finally, we considered a normal residual distribution $\epsilon_i \overset{i.i.d}{\sim} N(0, 0.25)$ and then generated the true failure times for uncured subjects with $G_i = 1$ based on Model (2.3). The censoring time C_i was independently generated from the exponential distribution $\exp(\lambda)$, where λ was selected to keep a censoring rate of around 70% or 80% corresponding to the cured proportion of 35% and 50%, respectively. Three sample sizes $n = 500, 1{,}000$, and $2{,}000$ were considered.

We fitted the parametric AFT model with intercept only and log-normal residuals to obtain the intercept estimates $\hat{\mu}_{AFT}$ and scale estimates $\hat{\sigma}_{AFT}$ based on the generated data. The parametric AFT model with the full set of covariates was also fitted to obtain a rough scale estimate $\hat{\sigma}_r$. The estimated propensity scores were included as a covariate to circumvent regularization-induced confounding. The following prior inputs were considered as specified in Section 3.1:

$$\begin{aligned}\text{Prior (I)} \quad & c_c = 3,\ c_\mu = 4\hat{\sigma}_{AFT},\ k_c = 2,\ k_1 = 2,\ k_2 = 4,\ \nu = 3,\ q = 0.99, \\ & \tilde{\alpha} = \alpha = 0.95,\ \tilde{\beta} = \beta = 2,\ \alpha' = 0.25,\ \beta' = 3.\end{aligned} \tag{4.1}$$

The algorithm converged within 200 iterations, and 2,000 posterior samples were collected by taking every two iterations after a burn-in stage of 1,000 iterations to obtain Bayesian estimates of the causal estimands. Table 1 summarizes the average root mean square error (RMSE) of the estimated UCATE, USATE, and CASP as well as the coverage rate (Cover.) and average length (Len.) of their 95% credible intervals based on 100 replications with the cured proportion being 35%. As expected, performance improves in terms of RMSE and interval coverage rate/length for both UCATE and USATE when the sample size increases, despite that interval coverage rate for CASP being lower than the nominal level. Such a lower coverage rate is possibly attributed to an inevitable misjudgment on G_i for a certain proportion of the censored subjects. We compared the performance of the proposed model in terms of the estimated UCATE, USATE, and CASP with that of the AFTrees approach proposed by Henderson et al. (2020), which outperforms a series of machine learning approaches while still taking no account of the cured fraction or regularization tailored for CATE. Simulation results reported in Table 1 suggest that our proposed model possesses smaller RMSE and shorter credible intervals with higher (or at least comparable) coverage probability for the estimated UCATE and USATE compared with the results of AFTrees. In terms of the estimated CASP, the proposed method brings smaller RMSE and slightly wider intervals with a notably higher coverage rate than the competing methods. Such strength holds with a similar pattern when the sample size increases and the cured proportion increases to 50%, while the coverage rate for USATE turned lower due to the high censoring rate. Detailed estimation results can be found in Table S1 of the Supplementary Material.

Considering that one key advantage of Model (2.3) is the controllable regularization priors imposed directly on the treatment effect through the BCF structure, we assessed how the Bayesian estimates are affected by changes in the strength and scope of such regularization imposed on $\tau(\mathbf{x}_i)$ through varying choices of hyperparameters. The varying conditions we considered are: (II) $\alpha' = \alpha = 0.95$, $\beta' = \beta = 2$, such that a weaker shrinkage toward homogeneity relative to Prior (I) was imposed on the UCATE; (III) $J_c = J = H = 200$; (IV) $J_c = J = H = 100$; (V) $J_c = J = H = 50$; and (VI) $q = 0.95$ such that the prior probability $Pr(\sigma^2 > \hat{\sigma}_r^2)$ is slightly higher. All other hyperparameters remain unchanged. The Bayesian estimates at sample size $n = 500$ and 1,000 with a cured proportion of 35% are given in Table S2 of the Supplementary Material, whereas the estimates obtained under Prior (I) are also presented in the first line for the ease of comparison. Overall, the average RMSE, coverage rate, and interval length for each causal estimand remain stable under the above settings, which is possibly due to the relatively large sample size of 1,000 we chose to mimic the size of SEER breast cancer data in Section 5. When the true treatment effect is indeed heterogeneous, alleviating the shrinkage to homogeneity through the BART prior on $\tau(\mathbf{x})$ as setting (II) slightly improves interval coverage rate for the estimated CASP, but at the cost of a mild increase in interval length and RMSE of UCATE and USATE. If the true treatment effect is homogeneous, on the contrary, the additional chances of misidentified heterogeneity allowed by setting (II) lead to increased RMSE and slightly decreased coverage rate for UCATE and USATE, which are likely due to nuisances learned by the comparatively deeper trees in $\tau(\mathbf{x})$. Varying the number of binary trees inside each BART or the residual-variance related hyperparameter q also makes little difference to the estimation results, somehow concurring with the cross-validation result obtained in Chipman et al. (2010). Similar patterns were obtained with $n = 500$, except that the increase in RMSE of UCATE and USATE gets even larger under settings (II)–(III) with less shrinkage to homogeneity and more noises captured by the redundant trees (e.g., $H = 200$). Overall, Bayesian estimates obtained are insensitive to prior inputs and almost the best under the default Prior (I).

For the ease of conveying the common causal assumptions and defining the causal estimands as in Section 2.3, we have adopted the most frequently used binary treatment indicator A_i with fixed coding one for the treated subjects and zero for the controls.

Heterogeneous: $\tau(\mathbf{x}_i) = 1 + 0.5x_{i2} - 0.25x_{i2}x_{i5}$

n	Method	UCATE			USATE			CASP		
		RMSE	Cover.	Len.	RMSE	Cover.	Len.	RMSE	Cover.	Len.
500	TBAFTcure	0.364	0.940	1.385	0.143	0.94	0.560	0.179	0.879	0.505
	AFTrees:np	0.383	0.951	1.592	0.183	0.96	0.665	0.210	0.619	0.293
	AFTrees:sp	0.513	0.931	1.931	0.348	0.82	0.948	0.219	0.674	0.343
1000	TBAFTcure	0.283	0.955	1.159	0.094	0.96	0.374	0.155	0.889	0.437
	AFTrees:np	0.319	0.947	1.277	0.157	0.80	0.426	0.199	0.582	0.239
	AFTrees:sp	0.546	0.877	1.712	0.401	0.40	0.690	0.203	0.631	0.301
2000	TBAFTcure	0.229	0.959	0.937	0.062	0.96	0.261	0.143	0.873	0.367
	AFTrees:np	0.274	0.940	1.025	0.147	0.53	0.288	0.194	0.523	0.194
	AFTrees:sp	0.574	0.799	1.533	0.417	0.05	0.509	0.185	0.568	0.263

Homogeneous: $\tau(\mathbf{x}_i) = 1.5$

n	Method	UCATE			USATE			CASP		
		RMSE	Cover.	Len.	RMSE	Cover.	Len.	RMSE	Cover.	Len.
500	TBAFTcure	0.206	0.999	1.431	0.146	1	0.614	0.152	0.917	0.496
	AFTrees:np	0.216	0.999	1.741	0.137	1	0.761	0.167	0.691	0.328
	AFTrees:sp	0.298	0.999	2.143	0.208	1	1.106	0.177	0.675	0.339
1000	TBAFTcure	0.187	0.995	1.147	0.123	0.93	0.422	0.146	0.904	0.429
	AFTrees:np	0.203	0.998	1.380	0.123	0.95	0.496	0.165	0.643	0.277
	AFTrees:sp	0.349	0.986	1.895	0.216	0.91	0.810	0.171	0.621	0.291
2000	TBAFTcure	0.144	0.993	0.886	0.082	0.94	0.291	0.139	0.874	0.361
	AFTrees:np	0.188	0.994	1.073	0.118	0.84	0.331	0.157	0.627	0.244
	AFTrees:sp	0.409	0.947	1.662	0.189	0.92	0.604	0.164	0.541	0.248

* np denotes nonparametric AFTrees with error distribution treated as a location mixture of Gaussian distributions; sp denotes semiparametric AFTrees with error distribution treated as normal.

Table 1: Average RMSE, coverage rate and interval length of the 95% credible interval for UCATE, USATE, and CASP estimated by the proposed TBAFTcure approach and AFTrees in Henderson et al. (2020) under $n = 500, 1000, 2000$ with $\epsilon_i \overset{i.i.d}{\sim} N(0, 0.25)$, cured rate around 35%, and censoring rate around 70%.

Nonetheless, it is trouble-free to incorporate the invariant parameterization of Hahn et al. (2020), which is a data-adaptive way of treatment coding, to the proposed model. By substituting $A_i = 1$ for the treated subjects with a parameter b_1 and $A_i = 0$ for the controls with a parameter b_0, the AFT model in (2.3) becomes

$$\begin{aligned}\log(T_i \mid \mathbf{x}_i, A_i, G_i = 1) &= \upsilon(\mathbf{x}_i) + \tau(\mathbf{w}_i) b_{A_i} + \epsilon_i \\ &= \sum_{j=1}^{J} g_1(\mathbf{x}_i; \mathcal{T}_j, \mathcal{M}_j) + \sum_{h=1}^{H} g_2(\mathbf{w}_i; \mathcal{T}'_h, \mathcal{M}'_h) b_{A_i} + \epsilon_i, \end{aligned} \tag{4.2}$$

but identification of the causal estimands remains the same and only minor adjustments are required in the posterior inference. Due to space limitations, details are provided in Web Appendix A. To compare model performance under the two choices of treatment coding, we conducted an additional simulation and followed the original work of Hahn et al. (2020) to assign a normal prior for the scale parameter b_{A_i}, i.e., $b_{A_i} \sim N(0, \frac{1}{2})$. J_c, J, and H were set as default and all other hyperparameters were set as Prior (I). Bayesian estimates of the causal estimands under the settings of $n = 500$ and 1,000 with a cured proportion of 35% are reported as condition (VII) in Table S2. The average RMSE, coverage rate, and interval length for the estimated USATE and CASP are relatively stable under the two different ways of treatment coding, while for the estimated UCATE, it turns out that invariant parameterization favours the case when the true treatment effect is homogeneous, with lower average RMSE and shorter credible intervals. When the true treatment effect is heterogeneous, slightly higher RMSE and lower coverage rate are observed for UCATE estimated under invariant parameterization, suggesting that stronger regularization to homogeneity can be induced by treatment coding in the AFT cure model. Overall, such discrepancies diminish as sample size increases to 2000 and above, and strengths of the proposed method over the AFTrees hold regardless of treatment coding when a cured fraction does exist.

We also evaluated robustness of the proposed method by varying (i) the level of residual variance σ^2 and (ii) the censoring and cured proportion under the setting of $n = 1{,}000$ and default Prior (I). For the former, Bayesian estimates of the causal estimands are given in Table S3 of the Supplementary Material. An increase in residual variance degrades estimation accuracy by inducing larger RMSE and lower coverage rate for the UCATE and USATE, but the proposed model still performs better than the AFTrees in both aspects, except that coverage rate for the UCATE obtained under invariant parameterization is slightly lower under the heterogeneous setting. Estimation of the CASP is not weakened for both methods, possibly because that the induced bias in the numerator of (3.1) is counteracted by the increasing σ in the denominator. For the latter, Table S4 provides estimation results under two relatively lower censoring rates, 25% and 55%, and a lower cured proportion of around 12%. Detailed settings are provided in Web Appendix B of the Supplementary Material. As the proportion of uncured subjects goes up (and the cured proportion goes down), the UCATE and USATE are estimated with better accuracy by both the proposed method and the AFTrees approach for comparison, including smaller average RMSE and shorter credible intervals with improved coverage rate. This is within expectation since more information can be acquired for the AFT model within the uncured subgroup. On the contrary, less information regarding the censored/cured subgroup can deteriorate the accuracy of estimating unobserved G_is by affecting performance of the probit cure model. However, the CASP is still estimated with better RMSE and coverage rate by the proposed method as shown in Table S4, while little improvement is obtained by the AFTrees approach even with the censoring rate decreased to 25% and the cured proportion as low as 12%. Across 100 replications, the average proportion of correctly identified G_is for the censored subjects remains relatively stable at around 85% as the censoring/cured rate decreases, somehow supporting robustness of the proposed method to different levels of complexity in estimating the unobserved cured group labels.

To further verify the chance of misidentifying a cured subgroup for the proposed method, we conducted an additional simulation under the setting of $n = 1{,}000$, heterogeneous treatment effect, without a cured fraction, and a censoring rate around 50%. Across the 100 replications, the number of misidentified cured subjects was consistently no larger than 2 and the estimated sample average cured rate (i.e., $\frac{1}{nN}\sum_{i=1}^{n}\sum_{itr=N_0+1}^{N_0+N} I(G_i^{(itr)} = 0)$) was no larger than 5%, indicating a very slight chance of misidentifying a nonexistent cured fraction. Table S5 of the Supplementary Material presents the corresponding causal estimands and those obtained based on the AFTrees approach. The proposed model performs comparably to the AFTrees approach, confirming a reliable performance of the proposed method regardless of the presence or absence of a cured subgroup. Considering that non-normal residuals from heavy-tailed distributions violate the model assumption and may hinder identification of the unobserved group label G_i, we also assessed performance of the proposed method under the setting of $n = 1{,}000$, heterogeneous treatment effect, a censoring rate around 50% without a cured fraction, and residuals from t or Gamma distribution. The causal estimands were presented in the middle and lower panel of Table S5. The proposed model and the semiparametric AFTrees performed alike, with slightly larger RMSE and lower coverage rate of the credible intervals compared with the nonparametric AFTrees tailored for handling non-normal residuals. Besides, the coverage rate of credible intervals for the CASP decreased for both approaches when the residuals turned non-normal. Despite that, the chance of misidentifying G_i remained low. The number of misidentified cured subjects was no larger than 7 across the 100 replications, with the estimated sample average cured rate again no larger than 5%.

Additionally, to gain some insights into how the proposed model scale to an increasing number of covariates, particularly irrelevant ones, we ran a simulation by adding eight or 24 irrelevant covariates in addition to the aforementioned x_1–x_8, leading to $p = 16$ or $p = 32$, respectively. Table S6 of the Supplementary Material summarizes the estimation results under the setup with $n = 1{,}000$, $\epsilon_i \overset{i.i.d}{\sim} N(0, 0.25)$, a cured rate around 35%, and a censoring rate around 70%. As the number of irrelevant covariates goes up, the performance of the proposed method gradually deteriorates, with slightly increased RMSE and decreased interval coverage rates obtained for each causal estimand compared to those reported in Table 1. These results are not surprising. Given that the proposed method involves pre-treatment covariates as splitting variables in the tree ensembles uniformly and randomly, accurately identifying the true prognostic factors and effect modifiers can be challenging as the number of irrelevant covariates increases. Detailed setups and discussions can be found in Web Appendix B of the Supplementary Material for the sake of space.

To obtain the results reported in Table 1 with $n = 1{,}000$ using the proposed method, it takes approximately 2 minutes of computing time per replication on a Linux machine running R with a CPU block speed of 2.60 GHz. The code for implementing the proceeding analysis is written in R with Rcpp and will be freely available at `https://github.com/roxiesun/TBAFTcure`.

5 Analysis of SEER breast cancer data

The proposed methodology was applied to a dataset regarding inflammatory breast cancer among U.S. females to further demonstrate its usage in estimating heterogeneous treatment effects on the time-to-event outcome in the existence of a cured fraction. The dataset is extracted from the SEER 17 Registries database (November 2021 submission), containing demographic and clinical records of 3,065 females aged between 15 and 85 years diagnosed with inflammatory breast cancer between 2001 and 2008. More detailed information can be found at the official website (https://seer.cancer.gov/). Inflammatory breast cancer is a relatively rare and aggressive type of breast cancer with an incidence of around 1% to 5% in the United States (Levine et al., 1985), well known for its fast-growing at onset and poor prognosis than other types of breast cancer. Neither surgery, radiation therapy, nor hormonal therapy alone was proved efficient in altering the natural history of this disease. However, developments in combined modality treatment over the past decades suggested evidence in improving disease-free survival of the patients (Jaiyesimi et al., 1992; Low et al., 2004). Such combined modality strategy usually includes chemotherapy followed by cancer-directed surgeries and radiation therapy. Notably, the prognosis of patients with the disease can vary across age, tumor stage, and overall health conditions. Moreover, over half of the patients collected did not take cancer-directed surgery with radiation therapy. Therefore, we focused on 1,382 patients with complete records and clearly defined tumor stage to investigate the possibly heterogeneous treatment effect of such cancer-directed surgery with radiation therapy on their lifespan.

The treatment indicator A_i was coded as 1 if the ith patient took cancer-directed surgery with radiation therapy and 0 otherwise. Eight pre-treatment covariates, including age at diagnosis, race (0 = white, 1 = black, 2 = others), Hispanic (1 = yes), tumor stage (0 = regional, 1 = distant), marital status (0 = single, 1 = married, 2 = others including divorced/widowed/separated), chemotherapy (1 = yes), total number of in situ/malignant tumors, and total number of benign/borderline tumors were considered as possible confounders or effect modifiers. The propensity scores estimated through *pbart* function of R package *BART* (Sparapani et al., 2021) were also included as a covariate. We examined the common support of the propensity score distributions under each treatment arm in Figure S2 of the Supplementary Material to check if the positivity assumption holds for the extracted dataset. Positivity violation is usually indicated by lack of sufficient overlap in the propensity score distribution, while the overlapped range observed in Figure S2 suggests only a slim chance of such violation. In general, a common solution to positivity violation is to truncate the propensity score distribution by using only subjects whose propensity scores fall within the common support for analysis and discarding those with extreme propensity score values of zero or one (see, Ju et al., 2019; Kang et al., 2016; Petersen et al., 2012). Nonetheless, recent discussions have highlighted that truncating the propensity score distribution to address positivity violation can come at a cost of changing the target population of interest and inducing extra bias or residual confounding. Such strategies should therefore be used with caution, where situations with high-dimensional covariates and rare outcomes are more likely to truly benefit from them (Shiba and Kawahara, 2021). Considering the overlapped range in Figure S2, we chose not to truncate the propensity

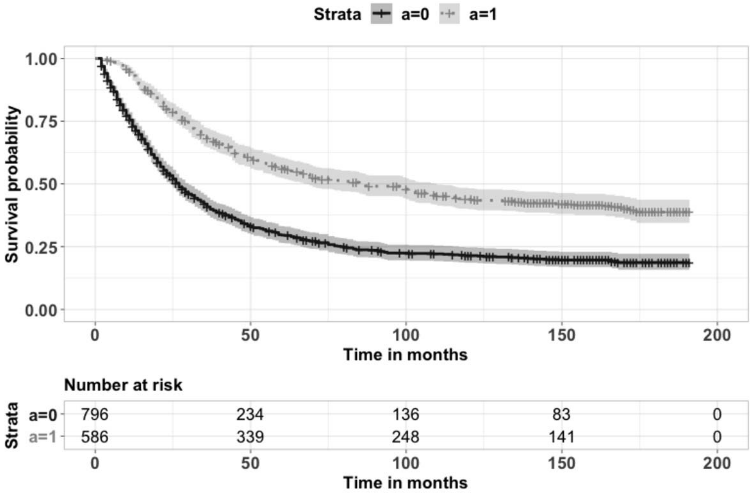
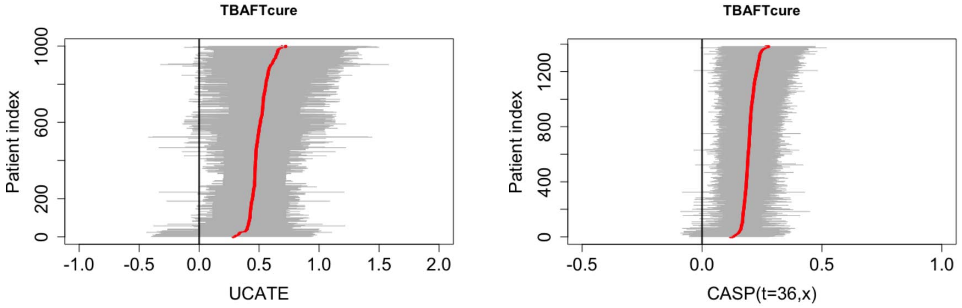


Figure 2: The Kaplan-Meier curves for patients who took cancer-directed surgery with radiation (Treatment = 1) and who not (Treatment = 0) in the inflammatory breast cancer data. For both group, the curves level off after 100 months although the patients were followed as long as 15 years and such proportion is higher in the treatment group, thereby indicating possible existence of a cured subgroup.

score distributions to avoid inducing any additional bias. The survival times obtained from the SEER database were reported in months. The event of interest was defined as death caused by this cancer and recorded with a median observed time of 36 months and a censoring rate of 33.4%. Figure 2 depicts the Kaplan-Meier curves for patients under the treatment group ($n = 586$) and control group ($n = 796$) with obvious leveling off after 100 months even if patients were followed as long as 15 years, suggesting a possible cured subgroup. Therefore, we implemented the proposed model to verify the existence of the cured subgroup and delve into whether patients with a divergent level of pathological and demographic traits benefit differently from being treated. The hyperparameters specified in (4.1) were adopted for analysis. The convergence of the algorithm was checked by trace plots of σ^2 in Figure S3 of the Supplementary Material, where chains with different initial values mixed well very soon. We collected 5,000 iterations by keeping every two after discarding 1,000 burn-in iterations to derive point estimates and credible intervals of the causal estimands. The results are summarized as follows.

Among the 462 censored subjects for whom the group label is unobservable, 384 were estimated by the proposed method with $\hat{G}_i = 0$ and thus identified as cured, while the rest were estimated with $\hat{G}_i = 1$ and recognised as uncured and just censored. The overall cured proportion was therefore calculated as $384/1382 = 27.8\%$. The estimated USATE for the uncured subgroup is 0.498 with a 95% credible interval of $(0.334, 0.667)$, indicating that patients within the uncured subgroup could benefit from the treatment with a survival time which is on average 1.645 (exp(0.498)) times of that if they were not



(a)

(b)

Figure 3: (a) Sorted UCATE for the 998 uncured subjects estimated by the proposed TBAFTcure model and (b) Sorted CASP with $t = 36$, the median observed time, for every subject. Red dots are the posterior means of UCATE/CASP for each (uncured) patient and the grey lines are the corresponding 95% credible intervals.

treated. Figure 3(a) depicts the estimated UCATE for each of the 998 uncured subjects along with the 95% credible intervals. In contrast, Figure 3(b) presents the estimated CASP$(t = 36, \mathbf{x}_i)$ for every subject in the dataset, which is the probability of surviving over the median observed time, 36 months, increased by the treatment. The median observed time is chosen for illustration purpose, and other values of t are completely feasible. It is worth noticing that although the estimated USATE is significantly positive, from both the plotted UCATE and CASP, we found a small portion of patients for whom the treatment effect is indeed nonsignificant in terms of a credible interval containing zero, thereby indicating a possible evidence of heterogeneity for the treatment effect. In other words, although cancer-directed surgery with radiation therapy decelerated the disease-caused death for 1.645 times averagely inside the uncured subgroup and increased the probability of surviving over 36 months by 0.198 (s.e. 0.024) on average, patients with certain features can have large or slim hope of benefiting more than that. Besides, estimation results under invariant parameterization was almost the same with those under 0/1 coding, including an overall cured proportion of 27.6% (381/1382), an estimated USATE of 0.487 with a 95% credible interval of (0.325, 0.651), and almost identical posterior probabilities as reported in Table 2.

To further quantify the degree of heterogeneity, in Table 2 we summarized the posterior probability of differential treatment effect as well as treatment benefit using the estimated UCATE. Results obtained by the AFTrees approach, which omits possible cured fraction, were also listed on the right column for a fair comparison. The proposed TBAFTcure found no strong or mild evidence of heterogeneity according to the $D_i^\star$, nor did the AFTrees approach for comparison. However, it is worth noticing that the cured subgroup identified by TBAFTcure can also be viewed as benefiting far beyond the average level from the treatment and serve as evidence of heterogeneity. In other words, we can capture heterogeneity from two aspects: patients for whom the decelerated event time brought by treatment is limited but deviates from the sample average level and

	TBAFTcure	AFTrees
$D_i^* > 0.95$	0%	0%
$D_i^* > 0.8$	0%	0%
$Pr(\tau(\mathbf{x}_i) > 0 \mid \mathcal{D}) \in (0.95, 1]$	94.3%	100%
$Pr(\tau(\mathbf{x}_i) > 0 \mid \mathcal{D}) \in (0.75, 0.95]$	5.7%	0%
$Pr(\tau(\mathbf{x}_i) > 0 \mid \mathcal{D}) \in (0.25, 0.75]$	0%	0%
$Pr(\tau(\mathbf{x}_i) > 0 \mid \mathcal{D}) \in [0, 0.25]$	0%	0%

* $D_i^* > 0.95/0.8$ suggests strong/mild evidence of differential treatment effect according to Henderson et al. (2020).

Table 2: The posterior probabilities of differential treatment effect and treatment benefit using UCATE estimated by the proposed TBAFTcure approach and CATE estimated by AFTrees approach in Henderson et al. (2020) for patients with inflammatory breast cancer.

patients who become cured and possess an event time of infinity, which in total takes up 27.8% of the whole sample. Although barely any evidence of heterogeneity was found by both approaches, we can still look in to some possible origins of heterogeneity in the treatment effect. The top three covariates used by $\tau(\mathbf{x})$ of the proposed model for splitting were Hispanic, race, and chemotherapy; while the top three splitting variables for $v_c(\mathbf{x}, A)$ were chemotherapy, tumor stage, and age. In contrast, the top variables other than treatment used by the AFTrees were Hispanic, tumor stage, chemotherapy, and the number of benign/borderline tumors. Unlike our method, however, the AFTrees could not shed light on whether these covariates were selected because of modifying the treatment effect or interpreting mixture components of the population, thereby failing to differentiate between confounders and treatment modifiers. Finally, Figure 4(a) shows the posterior density of the estimated UCATE under different levels of Hispanic, and Figure 4(b) gives the partial dependence plot for the number of in situ/malignant tumors. It is straightforward to see that Hispanic females benefit more from receiving

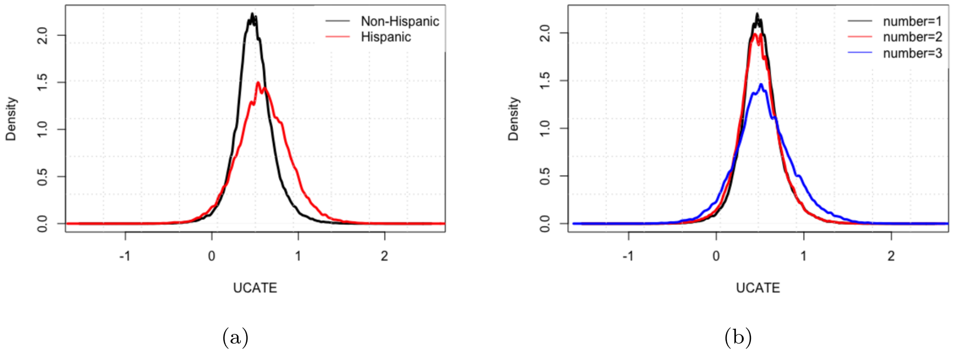


Figure 4: The partial effect of (a) Hispanic and (b) number of in situ/malignant tumors on the estimated UCATE for patients with inflammatory breast cancer.

cancer-directed surgery with radiation therapy, while the number of in situ/malignant tumors is less likely an effect modifier. This is also consistent with previous findings that the incidence and death rate of invasive breast cancer are relatively lower for Hispanic individuals than for Non-Hispanic White (see, e.g., Miller et al., 2021) and the fact that the number of in situ/malignant tumors was not among the top frequently used splitting variables for $\tau(\mathbf{x})$ across the MCMC iterations.

The above estimation results were obtained by implicitly assuming no ITB. The reasons are twofold: (i) although the SEER database contains the variable "Months from diagnosis to treatment" (MDT), it cannot serve as a rigorously defined immortal time. Patients subject to inflammatory breast cancer were usually treated with a combined modality strategy including chemotherapy followed by cancer-directed surgeries and radiation, while MDT was calculated with respect to the initial treatment, which, without clear specification, was more likely the chemotherapy status; (ii) ITB-related confounding can be trivial given that MDT is no larger than 2 months for around 95% of the subjects while over 95% of the treated subjects survived over 10 months as shown in Figure S4 of the Supplementary Material. Despite that, we adopted MDT as the immortal time to verify reliability of the above causal estimates and demonstrate how the proposed model can be used in conjunction with two popular mitigation strategies, i.e, excluding immortal time and PTDM. For the former, we simply subtracted the immortal time from Y_i for patients in the treated group, such that time zero for the treated is the time of actual exposure and time zero for the control is the time of enrollment. The proposed method gave an overall cured proportion of 27.7%, a slightly smaller USATE of 0.450 with a 95% credible interval of (0.282, 0.618), and almost identical posterior probabilities as those reported in Table 2. With the latter strategy, we first randomly sampled the missing immortal time for each subject in the control group with replacement from the observed immortal times in the treated group, and then subtracted the immortal times from the corresponding observed times for every subject in the sample. The same distribution of immortal time was thus ensured for the two treatment arms. If the sampled immortal time was larger than the observed time for a subject in the control group, they would be excluded from further analysis. Combined with PTDM, the proposed method produced distinct results with no cured proportion, a smaller USATE of 0.368 with a 95% credible interval of $(0.187, 0.545)$ and obviously different posterior probabilities. Possible reasons include the intrinsic drawbacks of PTDM, the potential conflicts between causal assumptions and altered distribution of the survival time, and the fact that MDT is not a well-defined immortal time. The estimated causal estimands are presented in Figure S5. Further details and discussion are provided in Web Appendix C of the Supplementary Material due to space limitations. It is worth noting that the two-stage analysis presented in this section is for demonstration purposes only, as the SEER breast cancer dataset does not cover rigorously defined immortal times. Once thorough information on immortal times and time-varying covariates can be acquired, more advanced mitigation strategies can be employed to improve the two-stage procedure.

6 Discussion

In this article, we proposed a tree-based AFT cure model to achieve estimation, detection, and causal interpretation of the possibly heterogeneous treatment effect on the scale of survival time and probability, accommodating the existence of a cured fraction. In cancer survival analysis, many studies focus on progression-free survival since death cannot be "cured", and cure models are useful to explore the heterogeneity among patients who are long-term survivors and those who are not. The proposed methodology can also be suitable in such cases to quantify heterogeneity raised by such cured fraction and effect modifiers. Compared with the current BART-based causal inference for survival data that regards the treatment indicator as just another covariate, the BCF structure adopted in our model enables controllable regularization imposed directly on the UCATE through a separate BART and selects the top-used effect modifiers straightforwardly with easy visualization of the induced heterogeneity. The probit BART component differentiates the uncured subjects for whom treatment effects on extending event time are further assessed. Meanwhile, it selects covariates that explain variation in cured rate for each subject. We developed a Bayesian approach with backfitting MCMC algorithm and the Gibbs sampler to estimate the causal estimands efficiently. Simulation results showed the proposed model outperformed one of the most advanced causal machine learning approaches in survival analysis when a cured fraction did exist. Finally, an application to SEER breast cancer data further manifested the usage of the proposed method.

This study has several limitations. First, we chose the AFT model for its straightforward interpretation of the treatment effect on decelerating/accelerating the event time. But this led to ill-defined CATE on the scale of survival time for cured subjects and thereby a combination of causal estimands with different scales. A joint framework of the BCF and Cox mixture cure model is likely to unify the scale of causal estimands for both the cured and uncured subgroup in terms of the hazard ratio. Other than the logarithm of survival time, modeling the hazard function or survival function with the nonparametric ensemble of trees may improve estimation accuracy of the CASP and facilitate treatment effect characterization on the scale of survival or cured probability. But with link functions or transformation functions involved in this way, how to achieve the straightforward linkage between the target causal estimand and the "modifier" BART $\tau(\mathbf{x})$ efficiently through reparametrization remains a future direction. Second, the proposed method was applied with two naïve mitigation strategies to alleviate ITB-related confounding, considering that the SEER breast cancer dataset does not possess thorough information on ITB. Once sufficient knowledge on enrollment, treatment initiation, and immortal time is accessible, future investigation with emulated trial or more advanced mitigation strategies including landmark analysis and stratified Cox model with time-varying treatment (Wang et al., 2022) are promising and of great interest. But it is also worth noticing that such mitigation strategies can more or less change the original data distribution and distort the underlying causal relationship of interest. Further investigation is required to connect ITB-targeted approaches to causal inference in a more flawless way. Third, the effect modifiers were selected based on merely the frequency of usage in the "modifier" BART $\tau(\mathbf{x})$ rather than rigorous variable selection procedures and thereby cannot guarantee consistency or adapt sparsity, especially in high-dimensional circumstances. Introducing a sparsity-inducing Dirichlet hyperprior (Caron et al., 2022; Linero, 2018) or spike-and-tree prior (Ročková and van der Pas, 2020) on splitting rules of the proposed model, or a permutation-based variable selection approach (Bleich et al., 2014) alternatively, are promising solutions to this problem. Despite the black-box nature as a causal machine learning approach, tree-based ensemble methods have experienced rapid development in theories on posterior consistency in recent years. Extending the theoretical results to the proposed model with time-to-event outcome and a cured fraction is important and worthy of consideration. The fulfillment of these extensions is of great interest and requires further investigation in the future.

Supplementary Material

Supplementary Material for "A tree-based Bayesian accelerated failure time cure model for estimating heterogeneous treatment effect" (DOI: 10.1214/23-BA1402SUPP; .pdf).

References

Andersen, P. K., Perme, M. P., van Houwelingen, H. C., Cook, R. J., Joly, P., Martinussen, T., Taylor, J. M., Abrahamowicz, M., and Therneau, T. M. (2021). "Analysis of time-to-event for observational studies: Guidance to the use of intensity models." *Statistics in Medicine*, 40(1): 185–211. MR4194578. doi: `https://doi.org/10.1002/sim.8757`. 3

Andersen, P. K., Syriopoulou, E., and Parner, E. T. (2017). "Causal inference in survival analysis using pseudo-observations." *Statistics in Medicine*, 36(17): 2669–2681. MR3670384. doi: `https://doi.org/10.1002/sim.7297`. 1

Bleich, J., Kapelner, A., George, E. I., and Jensen, S. T. (2014). "Variable selection for BART: An application to gene regulation." *The Annals of Applied Statistics*, 8(3): 1750–1781. MR3271352. doi: `https://doi.org/10.1214/14-AOAS755`. 24

Caron, A., Baio, G., and Manolopoulou, I. (2022). "Shrinkage Bayesian causal forests for heterogeneous treatment effects estimation." *Journal of Computational and Graphical Statistics*, 31(4): 1202–1214. MR4513381. doi: `https://doi.org/10.1080/10618600.2022.2067549`. 24

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 4(1): 266–298. MR2758172. doi: `https://doi.org/10.1214/09-AOAS285`. 2, 6, 7, 10, 11, 14

Conlon, A., Taylor, J., and Sargent, D. J. (2014). "Multi-state models for colon cancer recurrence and death with a cured fraction." *Statistics in Medicine*, 33(10): 1750–1766. MR3246693. doi: `https://doi.org/10.1002/sim.6056`. 3

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2020). "Estimating heterogeneous treatment effects with right-censored data via causal survival forests." *arXiv preprint arXiv:2001.09887*. 2

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition." *Statistical Science*, 34(1): 43–68. MR3938963. doi: https://doi.org/10.1214/18-STS667. 2

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). "Subgroup identification from randomized clinical trial data." *Statistics in Medicine*, 30(24): 2867–2880. MR2844689. doi: https://doi.org/10.1002/sim.4322. 2

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine." *The Annals of Statistics*, 29(5): 1189–1232. MR1873328. doi: https://doi.org/10.1214/aos/1013203451. 4

Gao, X. and Zheng, M. (2017). "Estimating the causal effects in randomized trials for survival data with a cure fraction and non compliance." *Communications in Statistics-Theory and Methods*, 46(8): 4065–4087. MR3590856. doi: https://doi.org/10.1080/03610926.2015.1076481. 3

Gran, J. M., Røysland, K., Wolbers, M., Didelez, V., Sterne, J. A., Ledergerber, B., Furrer, H., Von Wyl, V., and Aalen, O. O. (2010). "A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study." *Statistics in Medicine*, 29(26): 2757–2768. MR2757022. doi: https://doi.org/10.1002/sim.4048. 1

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). "Regularization and confounding in linear regression for treatment effect estimation." *Bayesian Analysis*, 13(1): 163–182. MR3737947. doi: https://doi.org/10.1214/16-BA1044. 2

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion)." *Bayesian Analysis*, 15(3): 965–1056. MR4154846. doi: https://doi.org/10.1214/19-BA1195. 2, 3, 6, 7, 10, 12, 15, 16

Hastie, T. and Tibshirani, R. (2000). "Bayesian backfitting (with comments and a rejoinder by the authors." *Statistical Science*, 15(3): 196–223. MR1820768. doi: https://doi.org/10.1214/ss/1009212815. 4

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). "Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models." *Biostatistics*, 21(1): 50–68. MR4043845. doi: https://doi.org/10.1093/biostatistics/kxy028. 2, 10, 12, 14, 15, 21

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., and Shrier, I. (2016). "Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses." *Journal of Clinical Epidemiology*, 79: 70–75. 3

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240. MR2816546. doi: https://doi.org/10.1198/jcgs.2010.08162. 2

Hu, L., Ji, J., and Li, F. (2021). "Estimating heterogeneous survival treatment effect in observational data using machine learning." *Statistics in Medicine*, 40(21): 4691–4713. MR4315446. doi: `https://doi.org/10.1002/sim.9090`. 2

Imbens, G. W. and Rubin, D. B. (1997). "Bayesian inference for causal effects in randomized experiments with noncompliance." *The Annals of Statistics*, 25(1): 305–327. MR1429927. doi: `https://doi.org/10.1214/aos/1034276631`. 9

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. MR2516796. doi: `https://doi.org/10.1214/08-AOAS169`. 2

Jaiyesimi, I. A., Buzdar, A. U., and Hortobagyi, G. (1992). "Inflammatory breast cancer: a review." *Journal of Clinical Oncology*, 10(6): 1014–1024. 18

Ju, C., Schwab, J., and van der Laan, M. J. (2019). "On adaptive propensity score truncation in causal inference." *Statistical Methods in Medical Research*, 28(6): 1741–1760. MR3961963. doi: `https://doi.org/10.1177/0962280218774817`. 18

Kang, J., Chan, W., Kim, M.-O., and Steiner, P. M. (2016). "Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores." *Communications for Statistical Applications and Methods*, 23(1): 1. 18

Karim, M. E., Gustafson, P., Petkau, J., Tremlett, H., Benefits, L.-T., of Beta-Interferon for Multiple Sclerosis (BeAMS) Study Group, A. E., Ehsanul Karim, M., Gustafson, P., Petkau, J., Tremlett, H., Shirani, A., et al. (2016). "Comparison of statistical approaches for dealing with immortal time bias in drug effectiveness studies." *American Journal of Epidemiology*, 184(4): 325–335. 3

Lambert, P. C., Thompson, J. R., Weston, C. L., and Dickman, P. W. (2007). "Estimating and modeling the cure fraction in population-based cancer survival analysis." *Biostatistics*, 8(3): 576–594. 3

Levine, P. H., Steinhorn, S. C., Ries, L. G., and Aron, J. L. (1985). "Inflammatory breast cancer: the experience of the Surveillance, Epidemiology, and End Results (SEER) program." *Journal of the National Cancer Institute*, 74(2): 291–297. 18

Linero, A. R. (2018). "Bayesian regression trees for high-dimensional prediction and variable selection." *Journal of the American Statistical Association*, 113(522): 626–636. MR3832214. doi: `https://doi.org/10.1080/01621459.2016.1264957`. 24

Liu, J., Weinhandl, E. D., Gilbertson, D. T., Collins, A. J., and St Peter, W. L. (2012). "Issues regarding 'immortal time' in the analysis of the treatment effects in observational studies." *Kidney International*, 81(4): 341–350. 3

Logan, B. R., Sparapani, R., McCulloch, R. E., and Laud, P. W. (2019). "Decision making and uncertainty quantification for individualized treatments using Bayesian Additive Regression Trees." *Statistical Methods in Medical Research*, 28(4): 1079–1093. MR3934636. doi: `https://doi.org/10.1177/0962280217746191`. 2

Low, J. A., Berman, A. W., Steinberg, S. M., Danforth, D. N., Lippman, M. E., and Swain, S. M. (2004). "Long-term follow-up for locally advanced and inflammatory breast cancer patients treated with multimodality therapy." *Journal of Clinical Oncology*, 22(20): 4067–4074. 18

Mansournia, M. A., Nazemipour, M., and Etminan, M. (2021). "Causal diagrams for immortal time bias." *International Journal of Epidemiology*, 50(5): 1405–1409. 3

Mi, X., Hammill, B. G., Curtis, L. H., Greiner, M. A., and Setoguchi, S. (2013). "Impact of immortal person-time and time scale in comparative effectiveness research for medical devices: a case for implantable cardioverter-defibrillators." *Journal of Clinical Epidemiology*, 66(8): S138–S144. MR3554996. doi: https://doi.org/10.1002/sim.7019. 3

Miller, K. D., Ortiz, A. P., Pinheiro, P. S., Bandi, P., Minihan, A., Fuchs, H. E., Martinez Tyson, D., Tortolero-Luna, G., Fedewa, S. A., Jemal, A. M., et al. (2021). "Cancer statistics for the US Hispanic/Latino population, 2021." *CA: A Cancer Journal for Clinicians*, 71(6): 466–487. 22

Othus, M., Bansal, A., Koepl, L., Wagner, S., and Ramsey, S. (2017). "Accounting for cured patients in cost-effectiveness analysis." *Value in Health*, 20(4): 705–709. 3

Othus, M., Barlogie, B., LeBlanc, M. L., and Crowley, J. J. (2012). "Cure models as a useful statistical tool for analyzing survival." *Clinical Cancer Research*, 18(14): 3731–3736. 3

Pearl, J. (1995). "Causal diagrams for empirical research." *Biometrika*, 82(4): 669–688. MR1380809. doi: https://doi.org/10.1093/biomet/82.4.669. 9

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and Van Der Laan, M. J. (2012). "Diagnosing and responding to violations in the positivity assumption." *Statistical Methods in Medical Research*, 21(1): 31–54. MR2867537. doi: https://doi.org/10.1177/0962280210386207. 18

Rosenbaum, P. R. (1984). "From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment." *Journal of the American Statistical Association*, 79(385): 41–48. MR0763575. 9

Ročková, V. and van der Pas, S. (2020). "Posterior concentration for Bayesian regression trees and forests." *The Annals of Statistics*, 48(4): 2108–2131. MR4134788. doi: https://doi.org/10.1214/19-AOS1879. 24

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66(5): 688–701. 8

Rubin, D. B. (1978). "Bayesian inference for causal effects: the role of randomization." *The Annals of Statistics*, 6(1): 34–58. MR0472152. 8

Rubin, D. B. (2005). "Causal inference using potential outcomes: design, modeling, decisions." *Journal of the American Statistical Association*, 100(469): 322–331. MR2166071. doi: https://doi.org/10.1198/016214504000001880. 8

Rutqvist, L. E., Wallgren, A., and Nilsson, B. (1984). "Is breast cancer a curable disease? A study of 14,731 women with breast cancer from the Cancer Registry of Norway." *Cancer*, 53(8): 1793–1800. 3

Shiba, K. and Kawahara, T. (2021). "Using propensity scores for causal inference: pitfalls and tips." *Journal of Epidemiology*, 31(8): 457–463. 18

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). "Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package." *Journal of Statistical Software*, 97(1): 1–66. 18

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). "Nonparametric survival analysis using Bayesian Additive Regression Trees (BART)." *Statistics in Medicine*, 35(16): 2741–2753. MR3513715. doi: https://doi.org/10.1002/sim.6893. 2

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science*, 5(4): 465–472. MR1092986. 8

Steingrimsson, J. A. and Morrison, S. (2020). "Deep learning for survival outcomes." *Statistics in Medicine*, 39(17): 2339–2349. MR4119735. doi: https://doi.org/10.1002/sim.8542. 2

Stone, R. (1993). "The assumptions on which causal inferences rest." *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2): 455–466. MR1224409. 9

Suissa, S. (2008). "Immortal time bias in pharmacoepidemiology." *American Journal of Epidemiology*, 167(4): 492–499. 3

Sun, R. and Song, X. (2023). "Supplementary Material for "A tree-based Bayesian accelerated failure time cure model for estimating heterogeneous treatment effect"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1402SUPP. 4

Tan, Y. V. and Roy, J. (2019). "Bayesian additive regression trees and the General BART model." *Statistics in Medicine*, 38(25): 5048–5069. MR4022845. doi: https://doi.org/10.1002/sim.8347. 10

Wang, J., Peduzzi, P., Wininger, M., and Ma, S. (2022). "Statistical methods for accommodating immortal time: a selective review and comparison." *arXiv preprint arXiv:2202.02369*. 4, 7, 23

Wei, L.-J. (1992). "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis." *Statistics in Medicine*, 11(14-15): 1871–1879. 2

Yang, M., Dunson, D. B., and Baird, D. (2010). "Semiparametric Bayes hierarchical models with mean and variance constraints." *Computational Statistics & Data Analysis*, 54(9): 2172–2186. MR2719750. doi: https://doi.org/10.1016/j.csda.2010.03.025. 10

Yu, W., Chen, K., Sobel, M. E., and Ying, Z. (2015). "Semiparametric transformation models for causal inference in time to event studies with all-or-nothing compliance." *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 77(2): 397–415. MR3310532. doi: https://doi.org/10.1111/rssb.12072. 2, 8, 9

Zhao, Q. and Hastie, T. (2021). "Causal interpretations of black-box models." *Journal of Business & Economic Statistics*, 39(1): 272–281. MR4187189. doi: https://doi.org/10.1080/07350015.2019.1624293. 4

Zhou, X. and Song, X. (2021). "Mediation analysis for mixture Cox proportional hazards cure models." *Statistical Methods in Medical Research*, 30(6): 1554–1572. MR4269965. doi: https://doi.org/10.1177/09622802211003113. 3

Zhou, Z., Rahme, E., Abrahamowicz, M., and Pilote, L. (2005). "Survival bias associated with time-to-treatment initiation in drug effectiveness evaluation: a comparison of methods." *American Journal of Epidemiology*, 162(10): 1016–1023. 3, 7

Acknowledgments

The authors are thankful to the editor, the associate editor, and two anonymous reviewers for their valuable comments and suggestions, which have helped improve the article substantially.