Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **00**, 1–33
https://doi.org/10.1093/jrsssb/qkac001

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Original Article

Estimating heterogeneous treatment effects with right-censored data via causal survival forests

Yifan Cui[1], Michael R. Kosorok[2], Erik Sverdrup[3], Stefan Wager[3] and Ruoqing Zhu[4]

[1]Center for Data Science, Zhejiang University, Hangzhou, China
[2]Departments of Biostatistics and Statistics & Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, USA
[3]Stanford Graduate School of Business, Stanford University, Stanford, USA
[4]Department of Statistics, University of Illinois Urbana-Champaign, Champaign, USA

Address for correspondence: Yifan Cui, Center for Data Science, Zhejiang University, Hangzhou, China. Email: cuiyf@zju.edu.cn

Abstract

Forest-based methods have recently gained in popularity for non-parametric treatment effect estimation. Building on this line of work, we introduce causal survival forests, which can be used to estimate heterogeneous treatment effects in survival and observational setting where outcomes may be right-censored. Our approach relies on orthogonal estimating equations to robustly adjust for both censoring and selection effects under unconfoundedness. In our experiments, we find our approach to perform well relative to a number of baselines.

Keywords: causal inference, censored data, heterogeneous treatment effects, machine learning, random forest, survival analysis

1 Introduction

The problem of heterogeneous treatment effect estimation plays a central role in data-driven personalization and, as such, has received considerable attention in the recent literature, including Athey and Imbens (2016), D. J. Foster and Syrgkanis (2019), J. C. Foster et al. (2011), Hahn et al. (2017), Kennedy (2020), Künzel et al. (2019), Luedtke and van der Laan (2016b), Nie and Wager (2021), Tian et al. (2014) and Wager and Athey (2018). One difficulty in applying this line of work to medical applications, however, is that such applications often involve potentially censored survival outcomes—and existing methods for treatment heterogeneity often cannot be used in this setting.

To address this challenge, we propose *causal survival forests* (CSF), an adaptation of the causal forest algorithm of Athey et al. (2019) that adjusts for censoring using doubly robust estimating equations developed in the survival analysis literature (Tsiatis, 2007; van der Laan & Robins, 2003). The method is robust, computationally tractable, and outperforms available baselines in our experiments. We also study the asymptotic behaviour of CSF, and establish conditions under which its predictions are consistent and asymptotically normal.

We focus on a statistical setting where, for $i = 1, \ldots, n$ training examples, we assume independent and identically distributed (i.i.d.) tuples $\{X_i, T_i, C_i, W_i\}$, where $X_i \in \mathcal{X}$ denote covariates, $T_i \in \mathbb{R}_+$ is the survival time for the ith unit, $C_i \in \mathbb{R}_+$ is the time at which the ith unit gets censored, and $W_i \in \{0, 1\}$ denotes treatment assignment. The goal of a statistician is to measure the average

Received: April 26, 2021. **Revised:** September 5, 2022. **Accepted:** September 6, 2022

effect of the treatment W_i on survival time T_i conditional on $X_i = x$. We define causal effects in terms of the standard potential outcomes framework (Imbens & Rubin, 2015), i.e., we posit potential outcomes $\{T_i(0), T_i(1)\}$ such that $T_i = T_i(W_i)$, and seek to estimate the conditional average treatment effect (CATE) on the survival time,

$$\tau(x) = \mathbb{E}[T_i(1) - T_i(0) \mid X_i = x], \tag{1}$$

or some relevant transformation of the survival time. However, unlike in the usual setting for heterogeneous treatment effect estimation discussed in, e.g., Künzel et al. (2019) and Nie and Wager (2021), we do not always get to see the realized survival time T_i; rather, we only observe $U_i = T_i \wedge C_i$ along with a non-censoring indicator $\Delta_i = 1\{T_i \leq C_i\}$. The main challenge for CSF is thus in estimating $\tau(x)$ in (1) while only observing U_i and Δ_i as opposed to the target outcome T_i.

To illustrate the promise of CSF, consider the following simple simulation experiment. We compare our proposed method (CSF) to both an adaptation of the virtual twins (VT) method proposed by J. C. Foster et al. (2011), and a variant of what Künzel et al. (2019) call the *S*-learner approach implemented using the survival forests from Ishwaran et al. (2008) (SRC). We defer the details of these implementations to Section 4; however, we emphasize that—unlike our method—these baselines do not leverage insights from the literature on doubly robust survival analysis to target treatment effects. Moreover, following Athey and Imbens (2016), our method directly targets heterogeneity in the CATE when choosing where to place splits. In Figure 1, we plot estimated effects $\hat{\tau}(X_i)$ against the true $\tau(X_i)$ on randomly drawn test set points across two simulation designs discussed further in Section 4. In both cases, we find our approach to be considerably more accurate than the non-doubly robust baselines here.

Our paper is structured as follows. First, in Section 2, we motivate and present our proposed method, CSF. In Section 3, we study asymptotics of CSF in a generalized random forest setting introduced by Athey et al. (2019). In Section 4, we conduct simulation experiments and find the proposed method to perform well in a wide variety of settings relative to other proposals in the literature. In Section 5, our approach is illustrated via a data application to the data from AIDS Clinical Trials Group Protocol 175 (ACTG175). A software implementation is provided as part of the package `grf` for R and C++, available from CRAN and at github.com/grf-labs/grf (R Core Team, 2019; Tibshirani et al., 2022).

1.1 Related work

As discussed above, there is extensive recent literature on non-parametric methods for heterogeneous treatment effect estimation (Athey & Imbens, 2016; Athey et al., 2019; J. C. Foster et al., 2011; Friedberg et al., 2020; Hahn et al., 2017; Hill, 2011; Kennedy, 2020; Künzel et al., 2019; M. Lu et al., 2018; Luedtke & van der Laan, 2016b; Nie & Wager, 2021; Oprescu et al., 2019; Tian et al., 2014; Wager & Athey, 2018). However, right-censored survival data are frequently encountered in clinical trials and other biomedical research studies, and challenges related to this setting have mostly not been addressed in the existing literature. Here, we take a step towards enabling flexible heterogeneous treatment effect estimation with survival data by adapting the forest-based method of Athey et al. (2019) to this setting.

Random forests were originally proposed by Breiman (2001), and trace back to the work of Breiman et al. (1984) on CART trees. By now, the literature on general random forest methods for survival data has already received considerable attention; and, in particular, Leblanc and Crowley (1993) and Hothorn, Hornik, et al. (2006) studied survival tree models in the context of conditional inference trees; Hothorn, Bühlmann, et al. (2006) proposed using inverse probability of censoring weighting (IPCW) to compensate censoring in forest models; Ishwaran et al. (2008) proposed random survival forests, which extend random forests to handle survival data via using log-rank tests at each split on an individual survival tree (Ciampi et al., 1986; Segal, 1988); Zhu and Kosorok (2012) studied the impact of recursive imputation of survival forests on model fitting; Steingrimsson et al. (2016) proposed doubly robust survival trees by constructing doubly robust loss functions that use more information to improve efficiency; and Steingrimsson et al. (2019) constructed censoring unbiased regression trees and forests by considering a class of

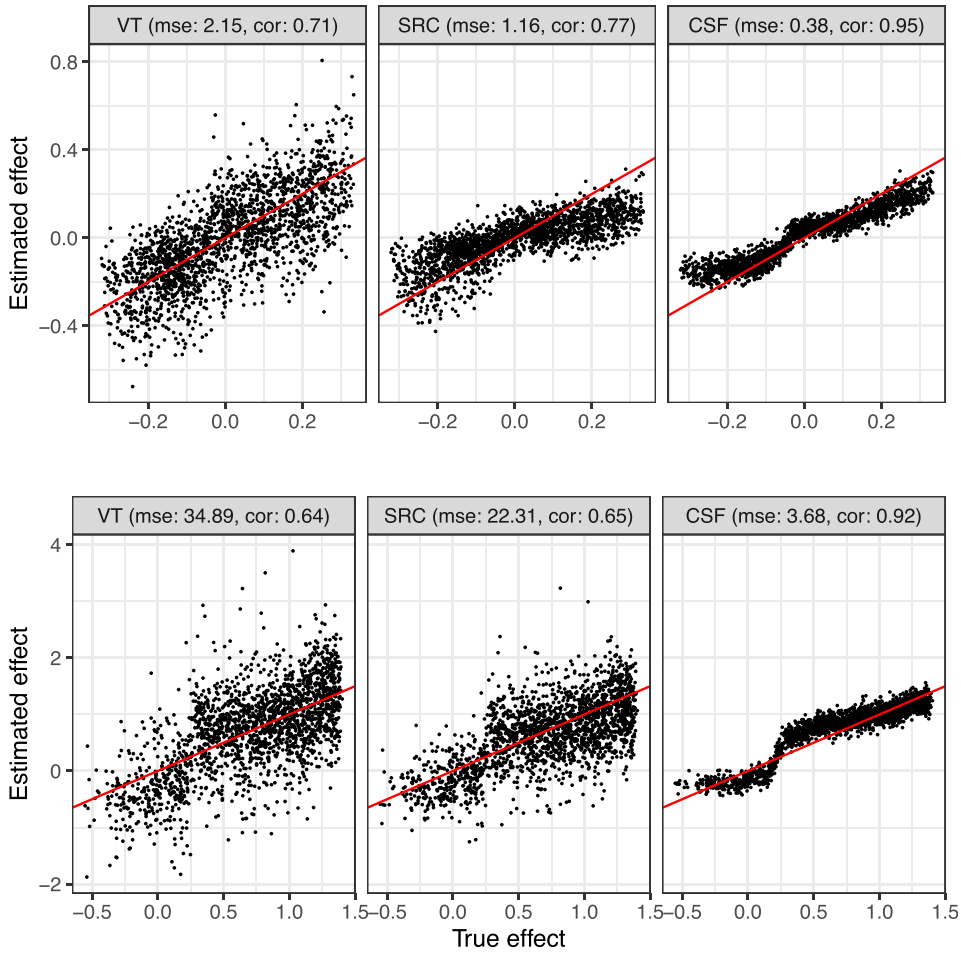

Figure 1. Conditional average treatment effect (CATE) predictions versus ground truth for the proposed method ('causal survival forests, CSF') with two alternatives, 'VT': virtual twins, 'SRC': the *S*-learner. The top row is for a data-generating process with continuous survival time (the second kind described in Section 4) and the bottom row is for a data-generating process with the discrete response (the third kind described in Section 4). Each estimator is fit on a sample size of 5,000, with default package tuning parameters (detailed in Section 4), predictions are for a test set of size 2,000. Mean squared error (MSE) and correlation with the true $\tau(x)$ is in parentheses. MSE is multiplied by 100 for readability.

censoring unbiased loss functions. However, none of these methods was directly targeted at heterogeneous treatment effects in observational studies.

Finally, we note that the problem of heterogeneous treatment effect estimation is closely related to that of optimal treatment regimes or policy learning (Athey & Wager, 2021; Cui, 2021; Luedtke & van der Laan, 2016a; Manski, 2004; Murphy, 2003; Qian & Murphy, 2011; B. Zhang et al., 2012; Y. Zhao et al., 2012). And, unlike in the case of heterogeneous treatment effects, there has been more work on developing methods for optimal treatment regimes under censoring. Adapting the outcome-weighted learning framework (Y. Zhao et al., 2012), Y.-Q. Zhao et al. (2015) proposed two new approaches, inverse censoring weighted outcome-weighted learning, and doubly robust outcome-weighted learning, both of which require semiparametric estimation of the conditional censoring probability given the patient characteristics and treatment choice. Zhu et al. (2017) adopted the accelerated failure time model to estimate an interpretable single-tree treatment decision rule. Cui et al. (2017) proposed a random forest approach for right-censored outcome-weighted learning, which avoids both the IPCW and restrictive modelling assumptions. However, for observational studies with censored survival outcomes, these methods may suffer

from confounding and selection bias. In addition, none of the above approaches focuses on estimating the heterogeneous treatment effect or on associated uncertainty quantification.

2 Causal survival forests

As discussed above, our statistical setting is specified in terms of the distribution of i.i.d. tuples $\{X_i, T_i, C_i, W_i\} \in \mathcal{X} \times \mathbb{R}_+ \times \mathbb{R}_+ \times \{0, 1\}$, and we are interested in measuring the effect of a binary treatment W on some deterministic transformation of the survival time T, i.e.,

$$\tau(x) = \mathbb{E}[y(T_i(1)) - y(T_i(0)) \mid X_i = x]. \tag{2}$$

To streamline our presentation, with a slight abuse of notation, $\tau(x)$ is redefined as (2) hereafter. We observe censored survival times $U_i = T_i \wedge C_i$, and write the non-censoring indicator $\Delta_i = 1\{T_i \leq C_i\}$.

Remark 1 Note that (1) is a special case of (2) with $y(T) = T$. Other transformations include $y(T) = T \wedge h$ for the restricted mean survival time (RMST) and $y(T) = 1\{T \geq h\}$ for the survival probability.

Throughout, we will assume that the transformation $y(\cdot)$ is indifferent to survival beyond some maximal time horizon h. The reason we make this assumption is that, in most studies, all units are censored after some finite amount of time (e.g., in a 10-year study, no units will be observed past the 10-year mark), and functionals of T_i that depend on behaviour past this point will not be identified. The upshot of Assumption 1 is that we can treat any sample tracked past the horizon h as 'observed', even if they are eventually censored.

Assumption 1 (Finite horizon). The outcome transformation $y(\cdot)$ admits a maximal horizon $0 < h < \infty$, such that $y(t) = y(h)$ for all $t \geq h$.

Definition 1 (Effective non-censoring indicator). Under Assumption 1, we define the effective non-censoring indicator as $\Delta_i^h = 1\{(T_i \wedge h) \leq C_i\}$; or, equivalently in terms of the observation U_i, we have $\Delta_i^h = \Delta_i \vee 1\{U_i \geq h\}$.

In order to identify treatment effects, we need to rely on two sets of assumptions. First, we need to make basic causal assumptions that would enable us to estimate the causal effect of W_i on T_i without censoring. The three assumptions below do so following Rosenbaum and Rubin (1983), and correspond exactly to standard assumptions used to identify the CATE (1) in the literature on heterogeneous treatment effect estimation (Künzel et al., 2019; Nie & Wager, 2021).

Assumption 2 (Potential outcomes). There are potential outcomes $\{T_i(0), T_i(1)\}$ such that $T_i = T_i(W_i)$ almost surely.

Assumption 3 (Ignorability). Treatment assignment is as good as random conditionally on covariates, $\{T_i(0), T_i(1)\} \perp W_i \mid X_i$.

Assumption 4 (Overlap). The propensity score $e(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$ is uniformly bounded away from 0 and 1, i.e., $\eta_e \leq e(x) \leq 1 - \eta_e$ for some $0 < \eta_e \leq \frac{1}{2}$.

Second, we need assumptions to guarantee that censoring due to C_i does not break identification results for treatment effects obtained via Assumptions 2–4. To this end, we rely on standard assumptions from survival analysis (e.g., Fleming & Harrington, 2011).

Assumption 5 (Ignorable censoring). Censoring is independent of survival time conditionally on treatment and covariates, $T_i \perp C_i \mid X_i, W_i$.

Assumption 6 (Positivity). $\mathbb{P}[C_i < h \mid X_i, W_i] \leq 1 - \eta_C$ for some $0 < \eta_C \leq 1$.

Assumptions 5 and 6 play a fundamental role for identification. Qualitatively, we note that these two assumptions can be seen as direct analogues to Assumptions 3 and 4, where the former control how censoring due to C_i can mask information about T_i, whereas the latter control how treatment assignment W_i can mask information about the potential outcomes $\{T_i(0), T_i(1)\}$.

Below, we start by reviewing the causal forest algorithm of Athey et al. (2019) that can be used to estimate the CATE without censoring under Assumptions 2–4. Next, we discuss two adaptations of causal forests for censored survival data: First, in Section 2.2, we discuss a simple proposal based on IPCW and then, in Section 2.3, we give our main proposal based on a doubly robust censoring adjustment.

2.1 Causal forests without censoring

The basic causal forest algorithm is motivated by the following fact. If we knew that the treatment effects were constant, i.e., $\tau(x) = \tau$ for all x, then the following estimator $\hat{\tau}$ due to P. M. Robinson (1988) attains $1/\sqrt{n}$ rates of convergence, provided the three assumptions detailed above hold and that we estimate nuisance components sufficiently fast (Chernozhukov et al., 2018):

$$\begin{aligned} &\sum_{i=1}^{n} \psi_{\hat{\tau}}^{(c)}(X_i, y(T_i), W_i;\ \hat{e}, \hat{m}) = 0, \\ &\psi_{\tau}^{(c)}(X_i, y(T_i), W_i;\ \hat{e}, \hat{m}) = [W_i - \hat{e}(X_i)][y(T_i) - \hat{m}(X_i) - \tau(W_i - \hat{e}(X_i))], \end{aligned} \tag{3}$$

where $e(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$, $m(x) = \mathbb{E}[y(T_i) \mid X_i = x]$, and $\hat{e}(X_i)$ and $\hat{m}(X_i)$ are estimates of these quantities derived via cross-fitting (Schick, 1986). We use the superscript (c) to remind ourselves that this estimator requires access to the complete (uncensored) data.

Here, however, our goal is not to estimate a constant treatment effect τ, but rather to fit covariate-dependent treatment heterogeneity $\tau(x)$; and we do so using a random forest method. As background, recall that given a target point x, tree-based methods seek to find training examples which are close to x and use the local kernel weights to obtain a weighted averaging estimator. An essential ingredient of tree-based methods is recursive partitioning of the covariate space $\mathcal{X}$, which induces the local weighting. When the splitting variables are adaptively chosen, the width of a leaf can be narrower along the directions where the causal effect is changing faster. After the tree fitting is completed, the closest points to x are those that fall into the same terminal node as x. The observations that fall into the same node as the target point x can be treated asymptotically as coming from a homogeneous group.

Athey et al. (2019) generalizes the use of random forest-based weights for generic kernel estimation. The most closely related precedent from the perspective of adaptive nearest neighbour estimation are quantile regression forests (Meinshausen, 2006) and bagging survival trees (Hothorn et al., 2004), which can be viewed as special cases of generalized random forests. The idea of adaptive nearest neighbours also underlies theoretical analyses of random forests such as Arlot and Genuer (2014), Biau et al. (2008), and Lin and Jeon (2006). The random forest-based weights α_i are derived from the fraction of trees in which an observation appears in the same terminal node as the target point. Specifically, given a test point x, the weights $\alpha_i(x)$ are the frequency with which the ith training example falls in the same leaf as x, i.e.,

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1\{X_i \in \mathcal{N}_b(x)\}}{|\mathcal{N}_b(x)|}, \tag{4}$$

where $\mathcal{N}_b(x)$ is the terminal node that contains x in the bth tree, B is the number of trees, and $|\cdot|$ denotes the cardinality.

The crux of the causal forest algorithm presented in Athey et al. (2019) is to pair the kernel-based view of forests with Robinson's estimating equation (3). Causal forests seek to grow a forest such that the resulting weighting function $\alpha_i(x)$ can be used to express heterogeneity in $\tau(\cdot)$,

meaning that $\tau(\cdot)$ is roughly constant over observations given positive weight $\alpha_i(x)$ for predicting at x. Then, we estimate $\tau(x)$ by solving a localized version of (3):

$$\sum_{i=1}^{n} \alpha_i(x) \psi^{(c)}_{\hat{\tau}(x)}(X_i, y(T_i), W_i;\ \hat{e}, \hat{m}) = 0. \tag{5}$$

Athey et al. (2019) discuss additional details, including the choice of a splitting rule targeted for treatment heterogeneity, while Nie and Wager (2021) and Kennedy (2020) further explore the idea of using Robinson's transformation to fit treatment heterogeneity. For a general review and discussion of causal forests, see Athey and Wager (2019).

2.2 Adjusting for censoring via weighting

In the presence of censoring, the estimator (5) is no longer applicable because T_i is not always observed. Simply ignoring censoring and running causal forests with complete observations (i.e., those with $\Delta_i^h = 1$) would lead to bias. However, under Assumptions 5 and 6, several general approaches for making statistical estimators robust to censoring are available.

One particularly simple censoring adjustment is via IPCW. Let

$$S^C_w(s \mid x) = \mathbb{P}[C_i \geq s \mid W_i = w, X_i = x] \tag{6}$$

denote the conditional survival function for the censoring process,[1] and note that any given survival time is observed in the sense of Definition 1 with probability

$$\mathbb{P}[\Delta_i^h = 1 \mid X_i, W_i, T_i] = S^C_{W_i}\left(T_i \wedge h \mid X_i\right), \tag{7}$$

The main idea of IPCW estimation is to only consider complete cases, but up-weight all complete observations by $1/S^C_{W_i}\left(T_i \wedge h \mid X_i\right)$ to compensate for censoring. As discussed in van der Laan and Robins (2003, Chapter 3.3), such IPCW estimators succeed in eliminating censoring bias under considerable generality.

In the context of causal forests, IPCW estimation for an outcome transformation satisfying Assumption 1 amounts to estimating $\tau(x)$ as

$$\sum_{\{i:\, \Delta_i^h = 1\}} \frac{\alpha_i(x)}{\hat{S}^C_{W_i}\left(T_i \wedge h \mid X_i\right)} \psi^{(c)}_{\hat{\tau}(x)}(X_i, y(T_i), W_i;\ \hat{e}, \hat{m}) = 0, \tag{8}$$

where $\hat{S}^C_w(s|x)$ is an estimate of (6) that is pre-computed using cross-fitting just like $\hat{e}$ and $\hat{m}$, and $\alpha_i(x)$ is obtained as in (4) trained on complete observations only. This algorithm forms the first adaptation of causal forests for censored data. We note that, when using (8), the nuisance component $\hat{m}(x)$ also needs to be estimated in a way that accounts for censoring; here, the IPCW approach can again be used. Concretely, we implement IPCW causal forests by training a model on samples with observed failure times only ($\Delta_i^h = 1$), and pass $1/\hat{S}^C_{W_i}\left(T_i \wedge h \mid X_i\right)$ as a 'sample weight' to the function `causal_forest` in `grf` (Tibshirani et al., 2022). We refer to the `grf` package documentation for more details of how sample weights are incorporated in all parts of the algorithm (including in splitting).

2.3 A doubly robust correction

While the IPCW approach discussed above is attractive in terms of its simplicity, it has several statistical drawbacks. First, this estimation strategy requires us to effectively throw away all observations with $\Delta_i^h = 0$, and this may hurt us in terms of efficiency: After all, if we see that a sample got censored at time $U_i > 0$, then at least we know they did not die before U_i, and a powerful

[1] The conditional survival function for the censoring process is defined using weak inequality because if the censoring event occurs at the same time as the failure event, the failure event is observed.

estimation procedure should be able to take this into account. Second, in general, we need to run IPCW with an estimate of $\hat{S}^C_w(s \mid x)$, and IPCW-type methods are generally not robust to estimation errors in this quantity—which will be a problem especially if we want to use flexible methods like random survival forests for estimating $\hat{S}^C_w(s \mid x)$ (Chernozhukov et al., 2018; van der Laan & Rose, 2011).

For this reason, our preferred CSF method does not rely on IPCW, and instead relies on a more robust approach to making estimating equations robust to censoring, as described in Tsiatis (2007, Chapter 10.4). If the true value τ of our parameter of interest is identified by a complete data estimating equation, $\mathbb{E}[\psi^{(c)}_\tau(X_i, y(T_i), W_i)] = 0$, then, under Assumption 1, τ is also identified via the following estimator that generalizes the celebrated augmented inverse propensity weighting estimator of J. M. Robins et al. (1994). We have $\mathbb{E}[\psi_\tau(X_i, y(U_i), U_i \wedge h, W_i, \Delta^h_i)] = 0$ with scores

$$\begin{aligned}\psi_\tau(X_i, y(U_i), U_i \wedge h, W_i, \Delta^h_i) = &\frac{\Delta^h_i \psi^{(c)}_\tau(X_i, y(U_i), W_i)}{S^C_{W_i}(U_i \wedge h \mid X_i)} \\ &+ \frac{(1 - \Delta^h_i)\mathbb{E}[\psi^{(c)}_\tau(X_i, y(T_i), W_i) \mid T_i \wedge h > U_i \wedge h, W_i, X_i]}{S^C_{W_i}(U_i \wedge h \mid X_i)} \\ &- \int_0^{U_i \wedge h} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)} \mathbb{E}[\psi^{(c)}_\tau(X_i, y(T_i), W_i) \mid T_i \wedge h > s,\ W_i, X_i]\, ds,\end{aligned} \tag{9}$$

where $S^C_w(s \mid x)$ is the conditional survival function as defined in (6) and

$$\lambda^C_w(s \mid x) = -\frac{d}{ds} \log S^C_w(s \mid x) \tag{10}$$

is the associated conditional hazard function.

When applied to the complete data estimating equation (3) used to motivate causal forests, this approach gives us scores

$$\begin{aligned}&\psi_\tau(X_i, y(U_i), U_i \wedge h, W_i, \Delta^h_i;\ \hat{e}, \hat{m}, \hat{\lambda}^C_w, \hat{S}^C_w, \hat{Q}_w) \\ &= \left(\frac{\hat{Q}_{W_i}(U_i \wedge h \mid X_i) + \Delta^h_i [y(U_i) - \hat{Q}_{W_i}(U_i \wedge h \mid X_i)] - \hat{m}(X_i) - \tau(W_i - \hat{e}(X_i))}{\hat{S}^C_{W_i}(U_i \wedge h \mid X_i)} \right. \\ &\quad \left. - \int_0^{U_i \wedge h} \frac{\hat{\lambda}^C_{W_i}(s \mid X_i)}{\hat{S}^C_{W_i}(s \mid X_i)} [\hat{Q}_{W_i}(s \mid X_i) - \hat{m}(X_i) - \tau(W_i - \hat{e}(X_i))]\, ds \right) \big(W_i - \hat{e}(X_i)\big),\end{aligned} \tag{11}$$

where $Q_w(s|x) = \mathbb{E}[y(T_i) \mid X_i = x,\ W_i = w,\ T_i \wedge h > s]$ is the conditional expectation of the transformed survival time, while $\hat{Q}_w(s \mid x)$, $\hat{S}^C_w(s \mid x)$ and $\hat{\lambda}^C_w(s \mid x)$ are cross-fit nuisance parameter estimates. For example, $\hat{Q}_w(s \mid x)$ can be estimated by an integral of estimated survival functions and $\hat{\lambda}^C_w(s \mid x)$ can be estimated as a forward difference of $-\log(\hat{S}^C_w(s \mid x))$.

The standard way of using scores is to construct a Neyman-orthogonal robust estimator of a relevant global parameter; see Chernozhukov et al. (2018) for a general discussion and references. In our case, the scores (11) induce a robust estimator $\hat{\tau}$ for a global constant treatment effect parameter τ:

$$\sum_{i=1}^n \psi_{\hat{\tau}}(X_i, y(U_i), U_i \wedge h, W_i, \Delta^h_i;\ \hat{e}, \hat{m}, \hat{\lambda}^C_w, \hat{S}^C_w, \hat{Q}_w) = 0. \tag{12}$$

This estimator is Neyman-orthogonal in a sense discussed in Chernozhukov et al. (2018), and attains a $1/\sqrt{n}$ rate of convergence for τ under fourth root rates for the nuisance components, provided we

use cross-fitting and that Assumptions 2–6 hold. Then the following result, given here for reference, illustrates the type of results one can get from this approach. The proof follows from arguments made in Chernozhukov et al. (2018) and Tsiatis (2007); see also van der Laan and Robins (2003) for an extensive discussion of double robustness under right-censoring.

Proposition 1. Assume that the CATEs are constant, $\tau(x) = \tau$ for all $x \in \mathcal{X}$, and let $\tilde{\tau}$ be an oracle estimator for τ, i.e., the solution to equation (12) with estimated nuisance components $\hat{e}$, $\hat{m}$, $\hat{\lambda}_w^C/\hat{S}_w^C$, $\hat{S}_w^C$, $\hat{Q}_w$ replaced by true values for e, m, λ_w^C/S_w^C, S_w^C, Q_w. Suppose $\sup_{x\in\mathcal{X}} |\hat{e}(x) - e(x)| = o_p(1)$, $\sup_{x\in\mathcal{X}} |\hat{m}(x) - m(x)| = o_p(1)$, and $\sup_{x\in\mathcal{X}, s\le h} |\hat{S}_w^C(s \mid x) - S_w^C(s \mid x)| = o_p(1)$, $\sup_{x\in\mathcal{X}, s\le h} |\hat{Q}_w(s \mid x) - Q_w(s \mid x)| = o_p(1)$,

$$\sup_{x\in\mathcal{X}, s\le h} \left| \frac{\hat{\lambda}_w^C(s \mid x)}{\hat{S}_w^C(s \mid x)} - \frac{\lambda_w^C(s \mid x)}{S_w^C(s \mid x)} \right| = o_p(1)$$

for both $w = 0, 1$. In addition, we assume that $\mathbb{E}[|\hat{e}(X) - e(X)|^2] = o(b_n^2)$, $\mathbb{E}[|\hat{m}(X) - m(X)|^2] = o(c_n^2)$, and $\sup_{s\le h} \mathbb{E}[|\hat{S}_w^C(s \mid X) - S_w^C(s \mid X)|^2] = o(c_n^2)$, $\sup_{s\le h} \mathbb{E}[|\hat{Q}_w(s \mid X) - Q_w(s \mid X)|^2] = o(c_n^2)$,

$$\sup_{s\le h} \mathbb{E}\left[\left| \frac{\hat{\lambda}_w^C(s \mid X)}{\hat{S}_w^C(s \mid X)} - \frac{\lambda_w^C(s \mid X)}{S_w^C(s \mid X)} \right|^2 \right] = o(d_n^2)$$

for both $w = 0, 1$. Then, provided Assumptions 2–6 hold, we have $\hat{\tau} - \tilde{\tau} = o_p(\max((c_n + d_n)c_n, b_n c_n, b_n^2))$. Furthermore, the solution $\hat{\tau}$ to estimating equation (12) has an asymptotically normal sampling distribution and attains a $1/\sqrt{n}$ rate of convergence, provided the nuisance components $\hat{e}$, $\hat{m}$, $\hat{\lambda}_w^C/\hat{S}_w^C$, $\hat{S}_w^C$, $\hat{Q}_w$ are learned using cross-fitting, are uniformly consistent and achieve 4th root rates of convergence in root-MSE.

Our CSF also use the doubly robust scores (11), but now in the context of a forest-based estimator for treatment heterogeneity. First, we estimate the nuisance components $\hat{e}$, $\hat{m}$, $\hat{\lambda}_w^C$, $\hat{S}_w^C$, $\hat{Q}_w$ required to form the score (11), and then pair this estimating equation with the forest weighting scheme (5), resulting in estimates $\hat{\tau}(x)$ characterized by

$$\sum_{i=1}^{n} \alpha_i(x) \psi_{\hat{\tau}(x)}(X_i, y(U_i), U_i \wedge h, W_i, \Delta_i^h;\ \hat{e}, \hat{m}, \hat{\lambda}_w^C, \hat{S}_w^C, \hat{Q}_w) = 0. \tag{13}$$

In order to use this estimator, we of course need to specify how to grow the forest, so that the resulting forest weights $\alpha_i(x)$ adequately express heterogeneity in the underlying signal $\tau(x)$. Here, for the splitting rule, we use the $\widetilde{\Delta}$-criterion proposed in Athey et al. (2019). In particular, we generate pseudo-outcomes by the following relabelling strategy at each parent node $\mathcal{N}$,

$$\rho_i = \psi^i_{\hat{\tau}_\mathcal{N}} \times \left[\frac{1}{|\{j : X_j \in \mathcal{N}\}|} \sum_{j:X_j\in\mathcal{N}} (W_j - \hat{e}(X_j))^2 \left(\frac{1}{\hat{S}^C_{W_j}(U_j \wedge h \mid X_j)} - \int_0^{U_j\wedge h} \frac{\hat{\lambda}^C_{W_j}(s \mid X_j)}{\hat{S}^C_{W_j}(s \mid X_j)}\, ds \right) \right]^{-1}, \tag{14}$$

where $\psi^i_{\hat{\tau}_\mathcal{N}}$ is a shorthand of $\psi_{\hat{\tau}_\mathcal{N}}(X_i, y(U_i), U_i \wedge h, W_i, \Delta_i^h;\ \hat{e}, \hat{m}, \hat{\lambda}_w^C, \hat{S}_w^C, \hat{Q}_w)$, and $\hat{\tau}_\mathcal{N}$ is the estimated τ in $\mathcal{N}$ by (12). Next, the splitting criterion proceeds exactly the same as a regression tree (Breiman et al., 1984) problem by treating the pseudo-outcomes ρ_i's as a continuous outcome variable. Specifically, we split the parent node into two child nodes $\mathcal{N}_L$ and $\mathcal{N}_R$ so as to maximize the following quantity:

$$\widetilde{\Delta}(L, R) = \frac{1}{|\{i : X_i \in \mathcal{N}_L\}|} \left(\sum_{i:X_i\in\mathcal{N}_L} \rho_i \right)^2 + \frac{1}{|\{i : X_i \in \mathcal{N}_R\}|} \left(\sum_{i:X_i\in\mathcal{N}_R} \rho_i \right)^2.$$

Remark 2 Throughout this paper, we will present our method and formal results in a general setting that covers any outcome transformation $y(\cdot)$ satisfying Assumption 1. However, the implementation of our doubly robust method, provided as the `causal_survival_forest` function in `grf`, only considers two target outcomes: The RMST,

$$\tau^h(x) = \mathbb{E}[T_i(1) \wedge h \mid X_i = x] - \mathbb{E}[T_i(0) \wedge h \mid X_i = x], \tag{15}$$

which arises from our framework using $y(T) = T \wedge h$, and the survival probability,

$$\pi^h(x) = \mathbb{P}[T_i(1) \geq h \mid X_i = x] - \mathbb{P}[T_i(0) \geq h \mid X_i = x], \tag{16}$$

which arises from using $y(T) = 1\{T \geq h\}$. Methodological extensions to other outcome transformations is straightforward, but requires modifying estimators for the nuisance components $\hat{Q}_w(s \mid x)$ and $\hat{m}(x)$. In our implementation for estimating $\tau^h(x)$ and $\pi^h(x)$, $\hat{Q}_w(s \mid x)$ and $\hat{m}(x)$ are derived from the estimated conditional survival function of the survival time.

Remark 3 Our splitting rule based on pseudo-outcomes as in (14) is a direct application of the abstract GRF-algorithm of Athey et al. (2019). This algorithmic technique is motivated by classical influence function-based approximations for semiparametric inference (Andrews, 1993; Zeileis, 2005). An alternative approach to splitting that could also be used with our doubly robust estimating equation is to maximize a test statistic for treatment heterogeneity as in, e.g., Zeileis et al. (2008) or Yang et al. (2021). Athey et al. (2019) have a further discussion of both statistical and computational properties of pseudo-outcome-based splitting.

3 Asymptotics and inference

The main draw of forest-based methods is that they have repeatedly proven themselves to be both robust and flexible in practice. However, to further our statistical understanding of the method, it is helpful to study the behaviour of the method under a more restricted asymptotic setting where sharp formal characterizations are available. To this end, we now study the asymptotics of CSF in a setting that builds on the one used by Wager and Athey (2018) to study inference with forests in problems without censoring. Throughout this section, we assume that the covariates $X \in \mathcal{X} = [0, 1]^p$ are distributed according to a density that is bounded away from zero and infinity. The following assumption guarantees the smoothness of $\mathbb{E}[\psi_{\tau(x)} \mid X = x]$.

Assumption 7 (Lipschitz continuity). The treatment effect function $\tau(x)$ is Lipschitz continuous in terms of x. In addition, the nuisance components $e(x)$, $m(x)$ are Lipschitz continuous in terms of x, and $Q_w(s \mid x)$, $\lambda_w^C(s \mid x)$, $S_w^C(s \mid x)$ are Lipschitz continuous in terms of x for both $w \in \{0, 1\}$ and all $s \leq h$.

In addition, our trees are symmetric, i.e., their output is invariant to permuting the indices of training samples. Our algorithm also guarantees honesty (Wager & Athey, 2018), and the following two conditions.

Random split tree: At each internal node, the probability of splitting at the jth dimension is greater than ς, where $0 < \varsigma < 1$ for $j = 1, \ldots, p$.

Subsampling: Each child node contains at least a fraction v of the data points in its parent node for some $0 < v < 0.5$, and trees are grown on subsamples of size ℓ scaling as

$$\ell = n^{\gamma}, \quad \kappa < \gamma < 1, \quad \kappa \equiv 1 - \left[1 + \varsigma^{-1}(\log(v^{-1}))/(\log(1-v)^{-1})\right]^{-1}. \tag{17}$$

We note that, in Theorem 3, we will obtain a rate of convergence of $\widetilde{O}(\sqrt{\ell/n})$ for $\hat{\tau}(x)$, where $\widetilde{O}(\cdot)$ denotes the term neglecting the logarithmic factors; thus (17) captures the asymptotic behaviour we can get; see Wager and Athey (2018) for an in-depth discussion of this assumption.

Moreover, we need Assumption 8 to couple $\hat{\tau}(x)$ and $\tilde{\tau}(x)$, where $\tilde{\tau}(x)$ is an oracle estimator, with nuisance components e, m, λ_w^C, S_w^C, Q_w being the underlying truth instead of their empirical analogues in equation (11).

Assumption 8 Consistency of the non-parametric plug-in estimators: we have the following convergences in probability,

$$\sup_{x\in\mathcal{X}} |\hat{e}(x) - e(x)| \to 0, \quad \sup_{x\in\mathcal{X}} |\hat{m}(x) - m(x)| \to 0,$$

and

$$\sup_{x\in\mathcal{X}, s\le h} |\hat{S}_w^C(s\,|\,x) - S_w^C(s\,|\,x)| \to 0, \quad \sup_{x\in\mathcal{X}, s\le h} |\hat{Q}_w(s\,|\,x) - Q_w(s\,|\,x)| \to 0,$$

$$\sup_{x\in\mathcal{X}, s\le h} \left| \frac{\hat{\lambda}_w^C(s\,|\,x)}{\hat{S}_w^C(s\,|\,x)} - \frac{\lambda_w^C(s\,|\,x)}{S_w^C(s\,|\,x)} \right| \to 0,$$

for both $w = 0, 1$. Furthermore, suppose that

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{e}(x) - e(x)|^2\right] = o(b_n^2), \quad \mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{m}(x) - m(x)|^2\right] = o(c_n^2),$$

and

$$\sup_{s\le h} \mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{S}_w^C(s\,|\,x) - S_w^C(s\,|\,x)|^2\right] = o(c_n^2), \tag{18}$$

$$\sup_{s\le h} \mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{Q}_w(s\,|\,x) - Q_w(s\,|\,x)|^2\right] = o(c_n^2), \tag{19}$$

$$\sup_{s\le h} \mathbb{E}\left[\sup_{x\in\mathcal{X}} \left| \frac{\hat{\lambda}_w^C(s\,|\,x)}{\hat{S}_w^C(s\,|\,x)} - \frac{\lambda_w^C(s\,|\,x)}{S_w^C(s\,|\,x)} \right|^2\right] = o(d_n^2) \tag{20}$$

for both $w = 0, 1$.

Equations (18)–(20) imply the corresponding rate assumptions of Proposition 1. Assumption 8 is quite general and can be achieved by many existing models. Biau (2012) and Wager and Walther (2015) show that for the random forest models, b_n^2 can be faster than $n^{-2/(p+2)}$ as long as the intrinsic signal dimension is less than $0.54p$. As shown in Cui et al. (2022), $c_n^2 = n^{-1/(p+2)}$ is achievable for survival forest models. Non-parametric kernel smoothing methods such as Sun et al. (2019) provide estimation with $d_n^2 = n^{-1/2+\kappa}$, where $\kappa < 1/2$ depending on the dimension p.

The following lemma provides an intermediate result for our main theorem, which bounds the difference between $\hat{\tau}(x)$ and $\tilde{\tau}(x)$.

Lemma 2. We assume Assumptions 6 and 8 hold. Then, for any $x \in \mathcal{X}$, we have that $\hat{\tau}(x) - \tilde{\tau}(x) = o_p(\max((c_n + d_n)c_n, b_n c_n, b_n^2))$.

The proof of Lemma 2 is deferred to the Appendix. The technical results in Wager and Athey (2018), Athey et al. (2019) paired with Lemma 2 lead to the following asymptotic Gaussianity result.

Theorem 3. Assume Assumptions 2–8 hold, and that ℓ scales as in (17). If $o_p(\max((c_n + d_n)c_n, b_n c_n, b_n^2))$ converges to zero with a faster rate than $\text{polylog}(n/\ell)^{-1/2}(\ell/n)^{1/2}$, where $\text{polylog}(n/\ell)$ is a function that is bounded away from 0 and increases at most polynomially with the log-inverse sampling ratio $\log(n/\ell)$. Then there exists a sequence $\sigma_n(x)$ such that for any $x \in \mathcal{X}$,

$$[\hat{\tau}(x) - \tau(x)]/\sigma_n(x) \to N(0, 1), \tag{21}$$

where $\sigma_n^2(x) = \text{polylog}(n/\ell)^{-1}\ell/n$.

The proof of Theorem 3 is deferred to the Appendix. This asymptotic Gaussianity result yields valid asymptotic confidence intervals for the true treatment effect $\tau(x)$.

3.1 Pointwise confidence intervals

One important consequence of Theorem 3 is that, in settings where its assumptions hold, we can use (21) to build pointwise confidence intervals for $\tau(x)$: All we need for confidence intervals is a consistent estimator for the asymptotic variance of $\hat{\tau}(x)$. We proceed using the 'bootstrap of little bags' algorithm (Athey et al., 2019; Sexton & Laake, 2009). The main idea of the bootstrap of little bags is to introduce some cluster structure into the random samples used by the random forest to grow each tree, and then use within versus between-cluster correlations in the behaviours of individual trees to reason about what we would have gotten by running (computationally infeasible) bootstrap uncertainty quantification on the whole forest.

Now, following the line of argument in Section 4 from Athey et al. (2019), under the conditions of Theorem 3 our estimator $\hat{\tau}(x)$ is asymptotically equivalent to

$$\tilde{\tau}^*(x) = \tau(x) + \sum_{i=1}^{n} \alpha_i(x)\rho_i^*(x), \tag{22}$$

where $\rho_i^*(x)$ is the influence function of the ith observation with respect to the true parameter value $\tau(x)$, i.e., $\rho_i^*(x) = \psi^i_{\tau(x)} V(x)^{-1}$ with

$$V(x) = \mathbb{E}\left[\left(W - e(X)\right)^2 \left(\frac{1}{S^C_W(U \wedge h \mid X)} - \int_0^{U\wedge h} \frac{\lambda^C_W(s \mid X)}{S^C_W(s \mid X)}\, ds\right) \middle| X = x\right]. \tag{23}$$

It thus follows that, to build confidence intervals, it suffices to estimate

$$\sigma_n^2(x) = \text{Var}[\tilde{\tau}^*(x)] = \text{Var}\left[\sum_{i=1}^{n} \alpha_i(x)\psi^i_{\tau(x)}\right] V(x)^{-2}, \tag{24}$$

where we note $\sum_{i=1}^n \alpha_i(x)\psi^i_{\tau(x)}$ is formally equivalent to the prediction made by a regression forest with 'outcome' $\psi^i_{\tau(x)}$. To this end, we set

$$\hat{\sigma}_n^2 = \hat{H}_n(x)\hat{V}_n(x)^{-2}, \tag{25}$$

where $\hat{H}_n(x)$ is estimated via the bootstrap of little bags device described above, and $\hat{V}_n(x)$ is a sample-version of (23) with forest weights $\alpha_i(x)$. We refer to Section 4 of Athey et al. (2019) for further details and discussion of this approach.

3.2 Inference on linear projections of the CATE

The pointwise confidence intervals discussed above provide a valuable method for assessing the stability of CSF, and also can be helpful in getting a handle on the behaviour of $\tau(x)$ in large samples. However, using these estimates in practice can sometimes lead to difficulties. First, the problem of pointwise non-parametric inference of $\tau(x)$ is fundamentally a very difficult problem as the function $\tau(\cdot)$ may take on complex shapes (Chernozhukov et al., 2018; Imai & Li, 2019), which means that intervals for $\tau(x)$ will, in general, be quite wide. Second, the result (21) relies on the forest being tuned for 'undersmoothing', i.e., that errors of the forest are dominated by variance and bias is negligible. In practice, it can be difficult to detect cases where undersmoothing does not hold (and confidence intervals are centred on potentially biased point estimates); see Appendix C3 of Athey et al. (2019) for further discussion.

In order to get around these difficulties, Semenova and Chernozhukov (2021) advocate focusing inference on low-dimensional summaries of $\tau(\cdot)$, including projections (Beran, 1977; Buja et al., 2019; White, 1980, 1982); see also van der Laan (2006) and Neugebauer and van der Laan (2007). Given a set of covariates A_i, the best linear projection (BLP) of the CATE function $\tau(\cdot)$ is

$$\{\beta_0^*, \beta^*\} = \underset{\beta_0, \beta}{\operatorname{argmin}}\ \mathbb{E}[\left(\tau(X_i) - \beta_0 - A_i\beta\right)^2]. \tag{26}$$

Typically, the A_i will be chosen as a subset (or transformations) of the X_i. As argued in Semenova and Chernozhukov (2021), such linear projections can be used to meaningfully interpret and summarize treatment heterogeneity, but remain simple enough that we can still provide robust inference for them using well-understood techniques from the literature on semiparametrics. This means a researcher is able to pre-specify a hypothesis of how they believe the conditional mean of $\tau(X)$ varies with, for example, age and gender, and achieve valid inference for this association measure, even though point estimates of $\tau(X)$ may be quite uncertain and obtained non-parametrically.

To implement this approach, we first construct relevant doubly robust scores, based on the augmented inverse propensity weighting (J. M. Robins et al., 1994)

$$\widehat{\Gamma}_i = \hat{\tau}(X_i) + \frac{\psi_{\hat{\tau}(X_i)}(X_i, y(U_i), U_i \wedge h, W_i, \Delta_i^h)}{\hat{e}(X_i)(1 - \hat{e}(X_i))}, \tag{27}$$

where $\psi_\tau(\cdot)$ is as in (11) and $\hat{\tau}(X_i)$ are CATE estimates provided by the CSF. We then estimate the BLP parameters (26) by running a linear regression of $\widehat{\Gamma}_i$ on the features of interest A_i; i.e., the regression $\widehat{\Gamma}_i \sim A_i$. Confidence intervals can be derived via any misspecification- and heteroskedasticity-robust approach to inference with ordinary least squares; in our implementation, we use the sandwich variance estimator with HC_3 covariance weights following MacKinnon and White (1985).

Semenova and Chernozhukov (2021) show that this approach yields a $\sqrt{n}$-rate central limit theorem for β and β_0 under conditions analogous to those of Proposition 1 on the nuisance components. In the case of our forest-based approach, the rates of convergence for nuisance components may be slower than those required by the result of Semenova and Chernozhukov (2021) so we cannot formally verify that a $\sqrt{n}$-rate central limit theorem holds for us; however, as shown in our simulation experiments (see Figure 2), the estimator of Semenova and Chernozhukov (2021) has a very nearly normal sampling distribution.

Remark 4 It would also be possible to consider the pseudo-outcomes from (27) as left-hand side variables for non-parametric regression, thus providing an alternative starting point for non-parametric estimation of CATE with survival outcomes. This class of approaches has received considerable attention in recent years; see, e.g., Q. Fan et al. (2022), Kennedy (2020), and Zimmert and Lechner (2019). Meanwhile, another alternative to BLPs is to apply assumption-lean inference recently proposed by Vansteelandt and Dukes

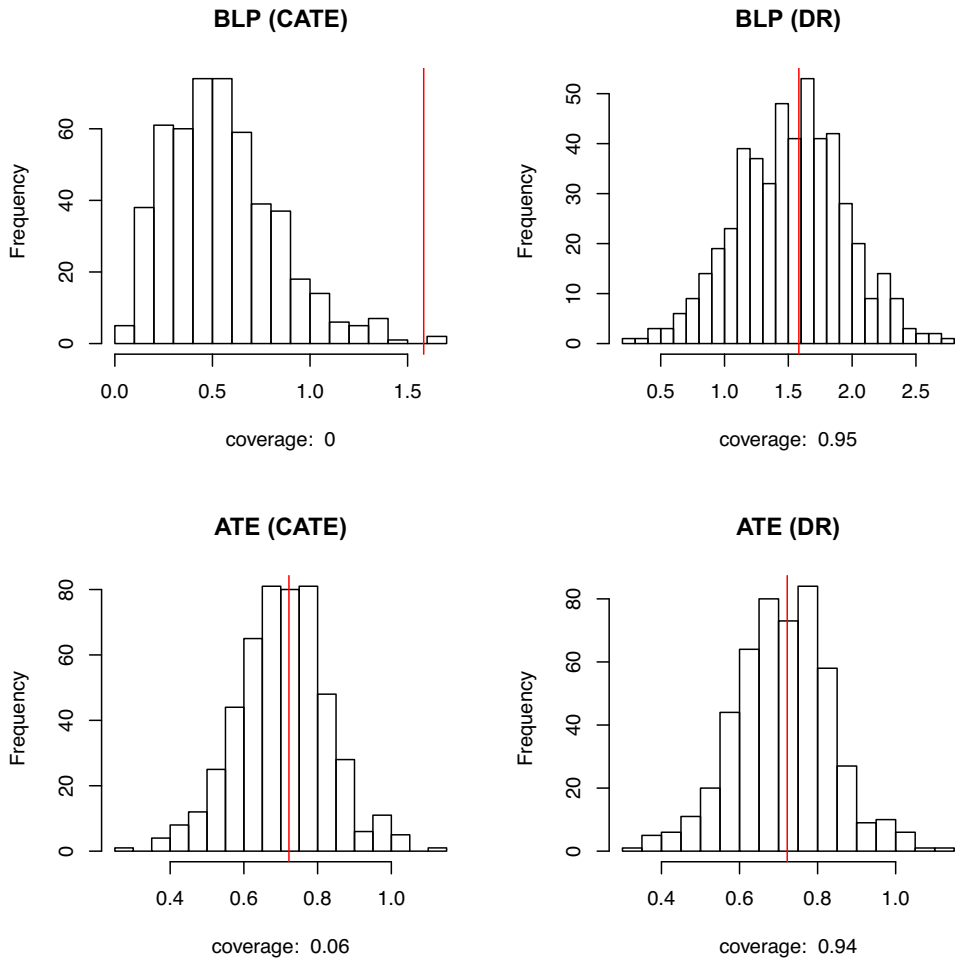

Figure 2. First row: histogram of the estimated best linear projection (BLP) coefficient $\hat{\beta}_1$ in the regression $\hat{\tau}(x) = \hat{\beta}_0 + \hat{\beta}_1 X_{(1)} + \hat{\beta}_2 X_{(2)}$, without (left) and with (right) a robustness correction to $\hat{\tau}(x)$, the estimated conditional average treatment effects (CATEs) with causal survival forests (CSF). Second row: histogram of the estimated coefficient β_0 (ATE) in the regression $\hat{\tau}(x) = \hat{\beta}_0$. The solid red line indicates the true population coefficient. Coverage is based on 95% confidence intervals computed as $\hat{\beta} \pm z_{0.975} se(\hat{\beta})$, where z is standard normal quantiles and $se(\hat{\beta})$ is derived via the HC_3 variance estimate (MacKinnon & White, 1985). Data is generated according to *Setting 3* with $n =$ 2, 000 training samples and $p = 15$ covariates. The number of repetitions is 500.

(2022) who use non-parametric projection to assess associations of the CATE with one variable at a time. Vansteelandt and Dukes (2022) argue that this approach provides robust summaries of the CATE when $\tau(\cdot)$ is non-linear in terms of the summarizing variables A_i.

4 Simulation study

We now turn to a simulation study to assess the empirical performance of CSF. As discussed in Remark 2, given a target horizon h, we focus on estimating the effect of the treatment on both the RMST at h and on survival probability at h. The choice of horizon h is given in each simulation specification.

In our experiments, we compare the following methods. (1) An adaptation of the VT method of J. C. Foster et al. (2011) to survival data, using random survival forests as implemented in the package `randomForestSRC` (Ishwaran & Kogalur, 2019): We train two random survival

forests, the first using the observations in the control group $\{(X, U, \Delta)\}_{W=0}$ to estimate $\mu_0(x) = \mathbb{E}[y(T_i) \mid X_i = x, W_i = 0]$, and the second using the observations in the treated group to estimate $\mu_1(x) = \mathbb{E}[y(T_i) \mid X_i = x, W_i = 1]$, where the conditional expectation estimation is constructed from the estimated conditional survival function, and then report $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$. (2) An instantiation of the *S*-learner strategy discussed in Künzel et al. (2019) using random survival forests (SRC1) as implemented in the package `randomForestSRC`: we train a single random survival forest with features (X, W) to estimate $\mu(x, w) = \mathbb{E}[y(T_i) \mid X_i = x, W_i = w]$, where the conditional expectation estimation is constructed from the estimated conditional survival function, and then report $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$. (3) Enriched random survival forests (SRC2) derived as above, except we train the forest with additional interaction features, i.e., (X, W, XW) as considered in M. Lu et al. (2018). (4) IPCW causal forests, as described in Section 2.2. (5) Our proposed CSF, as described in Section 2.3. Both of our methods, IPCW and CSF, are run using publicly available implementations in the `R` package `grf` version 2.1, available from CRAN and at github.com/grf-labs/grf (R Core Team, 2019; Tibshirani et al., 2022).

We run the above methods on the following simulation designs. In each, we generated independent covariates from a uniform distribution on $[0, 1]^p$ with $p = 15$. We consider two estimands, the RMST and survival probability. For the first estimand, the truncation time is listed in the simulation settings; for the second estimand, we consider surviving past the 90th percentile of U. The Supplementary material shows results when the covariates have a non-diagonal covariance matrix V with entries $V_{ij} = 0.5^{|i-j|}$. *Setting 1*. We generate T from an accelerated failure time model, and C from a Cox model,

$$\log(T) = -1.85 - 0.8I(X_{(1)} < 0.5) + 0.7X_{(2)}^{1/2} + 0.2X_{(3)} + (0.7 - 0.4I(X_{(1)} < 0.5) - 0.4X_{(2)}^{1/2})W + \epsilon,$$

$$\lambda_C(t \mid W, X) = \lambda_0(t)\exp[-1.75 - 0.5X_{(2)}^{1/2} + 0.2X_{(3)} + (1.15 + 0.5I(X_{(1)} < 0.5) - 0.3X_{(2)}^{1/2})W],$$

where the baseline hazard function $\lambda_0(t) = 2t$, and ϵ follows a standard normal distribution. The follow-up time is $h = 1.5$, and propensity score is $e(x) = (1 + \beta(x_{(1)}; 2, 4))/4$, where $\beta(\cdot; a, b)$ is the density function of a Beta distribution with shape parameters a and b. *Setting 2*. We generate T from a Cox model with a non-linear structure, and C from a uniform distribution on (0,3),

$$\lambda_T(t \mid W, X) = \lambda_0(t)\exp[X_{(1)} + (-0.5 + X_{(2)})W],$$

where the baseline hazard function is $\lambda_0(t) = 1/2t^{-1/2}$, and ϵ follow a standard normal distribution. The maximum follow-up time is $h = 2$, and the propensity score is $e(x) = (1 + \beta(x_{(2)}; 2, 4))/4$. *Setting 3*. We generate T from a Poisson distribution with mean $X_{(2)}^2 + X_{(3)} + 6 + 2(X_{(1)}^{1/2} - 0.3)W$, and C from a Poisson distribution with mean $12 + \log(1 + \exp(X_{(3)}))$. The maximum follow-up time is $h = 15$, and the propensity score is $e(x) = (1 + \beta(x_{(1)}; 2, 4))/4$. *Setting 4*. We generate T from a Poisson distribution with mean $X_{(2)} + X_{(3)} + \max(0, X_{(1)} - 0.3)W$, and C from a Poisson distribution with mean $1 + \log(1 + \exp(X_{(3)}))$. The maximum follow-up time is $h = 3$, and propensity score is $e(x) = [(1 + \exp(-x_{(1)}))(1 + \exp(-x_{(2)}))]^{-1}$. Note that for subjects with $X_{(1)} < 0.3$, treatment does not affect survival time, and thus the classification error rate is evaluated on the subgroup of $X_{(1)} \geq 0.3$.

Default tuning parameters were used for different forest-based methods. For the proposed CSF, and for causal forests with inverse probability of censoring weights, the default values are given in the `grf` package documentation: the number of trees is 2,000, the minimum node size is 5, the subsampling fraction is 0.5, and the number of split variables $\min(p, \lceil\sqrt{p}\rceil + 20)$, where $\lceil x\rceil$ denotes the least integer greater than or equal to x. For estimating the nuisance components, we also use default values: 500 trees, and a minimum node size of 15 for the survival forests, and the remaining parameters, the subsampling fraction and the number of split variables are the same as that of CSF. For random survival forests (Ishwaran & Kogalur, 2019), the minimal number of observations in each terminal node was chosen as the default, i.e., 15; The number of variables available for splitting at each tree node was chosen as the default, i.e., $\lceil p^{1/2}\rceil$. The total number of trees was set to 500.

Remark 5 The default tuning parameters in `grf` were chosen based on empirical performance in a series of experiments (performed independently from the research reported in this paper). This strategy reflects standard practice in applications. Another approach (not taken here) would have been to choose tuning parameters based on n and p using an algorithm guaranteed to satisfy asymptotic scaling conditions assumed in Theorem 3.

4.1 Results

Table 1 reports simulation results in terms of test set MSE for $\tau(X_i)$. Table 2 considers classification accuracy, i.e., $1 - (1/n_{\text{test}}) \sum_{i=1}^{n_{\text{test}}} 1\{\text{sign}(\hat{\tau}(X_i)) = \text{sign}(\tau(X_i))\}$. These two tables report results with training set size $n = 2,000$, while Figures A1–A4 in Section A.4 in the Appendix show results across several values of n.

In almost all scenarios, the proposed CSF achieves the best performance among all competing methods and, overall, the proposed CSF is superior to ordinary random survival forests which do not target the causal parameter directly. The exception is for setting 4 where, when targeting RMST, IPCW performs best, and when targeting the survival probability, SRC1 performs best. IPCW generally outperforms the other baselines (except for setting 4 where SRC1 is best). Two aspects of our algorithm that may help explain this finding is that, first, both IPCW and CSF are designed to target treatment effects specifically, and so are able to focus on covariates interacting with W rather than on the covariates which only appear in the main effect. Second, all of these simulation processes involve non-trivial right-censoring or confounding mechanism and, as emphasized throughout the paper, CSF was designed to robustly adjust for such censoring and confounding.

Table 1. Mean squared error (MSE) multiplied by 100 and excess MSE for four scenarios, defined as $(1/B) \sum_{b=1}^{B} \text{MSE}_b(\text{method}) / \min\{\text{MSE}_b(\text{m}) : \text{m} \in \{\text{all methods}\}\}$, respectively, for two estimands, restricted mean survival time (RMST) and survival probability

Setting	Metric	VT	SRC1	SRC2	IPCW	CSF
Panel A: RMST						
1	MSE	0.82	0.46	0.71	0.60	0.25
	Excess MSE	3.97	2.07	3.50	2.63	1.02
2	MSE	2.96	1.89	2.16	1.75	1.16
	Excess MSE	3.11	1.93	2.23	1.67	1.02
3	MSE	49.19	33.31	42.21	16.88	13.69
	Excess MSE	4.28	2.79	3.66	1.32	1.02
4	MSE	5.86	3.79	4.58	2.79	3.04
	Excess MSE	2.52	1.60	1.94	1.10	1.23
Panel B: Survival probability						
1	MSE	0.45	0.26	0.41	0.19	0.14
	Excess MSE	4.16	2.23	3.92	1.42	1.07
2	MSE	0.74	0.46	0.54	0.41	0.37
	Excess MSE	2.41	1.46	1.72	1.21	1.06
3	MSE	0.38	0.25	0.32	0.16	0.15
	Excess MSE	2.86	1.78	2.42	1.11	1.05
4	MSE	0.46	0.25	0.38	0.28	0.27
	Excess MSE	2.06	1.11	1.70	1.21	1.21

Note. The simulation settings and methods under consideration are as described in Section 4. All forests were trained on $n = 2,000$ samples, and an independent test set with size $n_{\text{test}} = 2,000$ was used to evaluate error. Each simulation was repeated 250 times.

Table 2. Classification error for two estimands, restricted mean survival time (RMST) and survival probability, for four scenarios, defined as $1 - (1/n_{\text{test}}) \sum_{i=1}^{n_{\text{test}}} 1\{\text{sign}(\hat{\tau}(X_i)) = \text{sign}(\tau(X_i))\}$ (for the last setting, the classification error rate is evaluated on the subgroup of $X_{(1)} \geq 0.3$)

Setting	VT	SRC1	SRC2	IPCW	CSF
Panel A: RMST					
1	0.27	0.23	0.27	0.23	0.22
2	0.25	0.21	0.23	0.26	0.15
3	0.18	0.15	0.19	0.08	0.08
4	0.14	0.10	0.11	0.01	0.00
Panel B: Survival probability					
1	0.28	0.25	0.28	0.22	0.22
2	0.25	0.22	0.24	0.25	0.22
3	0.19	0.16	0.17	0.09	0.09
4	0.16	0.11	0.11	0.05	0.03

Note. The simulation settings and methods under consideration are as described in Section 4. All forests were trained on $n = 2,000$ samples, and an independent test set with size $n_{\text{test}} = 2,000$ was used to evaluate classification error. Each simulation was repeated 250 times.

Table 3. Coverage (%) of the proposed 95% confidence intervals and average length at four deterministic points described in Section 4, with a training set size $n = 2,000$

	Coverage				Length			
Setting	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
Panel A: RMST								
1	0.84	0.69	0.88	0.95	0.09	0.08	0.15	0.18
2	0.75	0.93	0.92	0.82	0.32	0.25	0.21	0.23
3	0.82	0.96	0.89	0.72	0.96	0.81	0.82	0.95
4	0.37	0.68	0.96	0.88	0.36	0.31	0.32	0.39
Panel B: Survival probability								
1	0.83	0.76	0.91	0.93	0.07	0.06	0.14	0.16
2	0.71	0.92	0.93	0.86	0.18	0.14	0.12	0.13
3	0.84	0.95	0.87	0.62	0.10	0.09	0.10	0.12
4	0.67	0.84	0.92	0.43	0.10	0.10	0.11	0.15

Note. $B = 10,000$ trees are used to fit confidence intervals. The numbers are aggregated over 1,000 simulation replications.

Next, we consider the accuracy of the pointwise confidence intervals for CSF proposed in Section 3.1. To do so, we evaluated the coverage of the proposed 95% confidence intervals at four deterministic points, namely $x_1 = (0.2, \ldots, 0.2)^T$, $x_2 = (0.4, \ldots, 0.4)^T$, $x_3 = (0.6, \ldots, 0.6)^T$, and $x_4 = (0.8, \ldots, 0.8)^T$. The true treatment effect was estimated by the Monte Carlo method with a sample size 100,000. The results, summarized in Table 3, are mostly promising, and suggests that in most of these examples the bias-variance trade-off of CSF puts us in a regime where discussions from Section 3.1 apply. However, at other points we observe poor coverage, especially at x_1 and x_4, which are near the corners of the feature space.

Finally, we investigate the performance of the BLP estimator discussed in Section 3.2 which provides summaries of the CATE that are amenable to more robust inference. In Figure 2, given various choices of projection variables A_i in (26), we consider both the doubly robust method of

Semenova and Chernozhukov (2021) (DR) and a naive baseline that simply regresses the fitted $\hat{\tau}(X_i)$ estimates against A_i without a doubly robust correction (CATE). We find that the proposed BLP method performs well; in contrast, the naive direct regression can be biased (top row), and badly underestimates the sampling variability of these estimators, thus resulting in poor coverage (both rows).

5 HIV data analysis

We demonstrate the proposed method by an application to the data from AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al., 1996). The original dataset consists of 2,139 HIV-infected subjects. The enrolled subjects were randomized to four treatment groups: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine, and ddI monotherapy. We focus on the subset of patients receiving the treatment ZDV+ddI or ddI monotherapy as considered in W. Lu et al. (2013). Treatment indicator $W = 0$ denotes the treatment ddI with 561 subjects, and $W = 1$ denotes the treatment ZDV+ddI with 522 subjects. Though ACTG175 is a randomized study, there seem to be some selection effects in the subsets used here. For example, for covariate race equals to 1, there are 138 receiving ZDV+ddI and 173 receiving ddI. A binomial test with null probability 0.5 gives p-value 0.05. For this reason, we analyse the study as an observational rather than randomized study.

Here, we are interested in the causal effect between ZDV+ddI and ddI on survival time of HIV-infected patients. 12 selected baseline covariates were studied in Tsiatis et al. (2008), M. Zhang et al. (2008), W. Lu et al. (2013), and C. Fan et al. (2017). There are five continuous covariates: age (year), weight (kg), Karnofsky score (scale of 0–100), CD4 count (cells/mm^3) at baseline, CD8 count (cells/mm^3) at baseline. There are seven binary variables: gender (male = 1, female = 0), homosexual activity (yes = 1, no = 0), race (non-white = 1, white = 0), symptomatic status (symptomatic = 1, asymptomatic = 0), history of intravenous drug use (yes = 1, no = 0), haemophilia (yes = 1, no = 0), and antiretroviral history (experienced = 1, naive = 0). As the

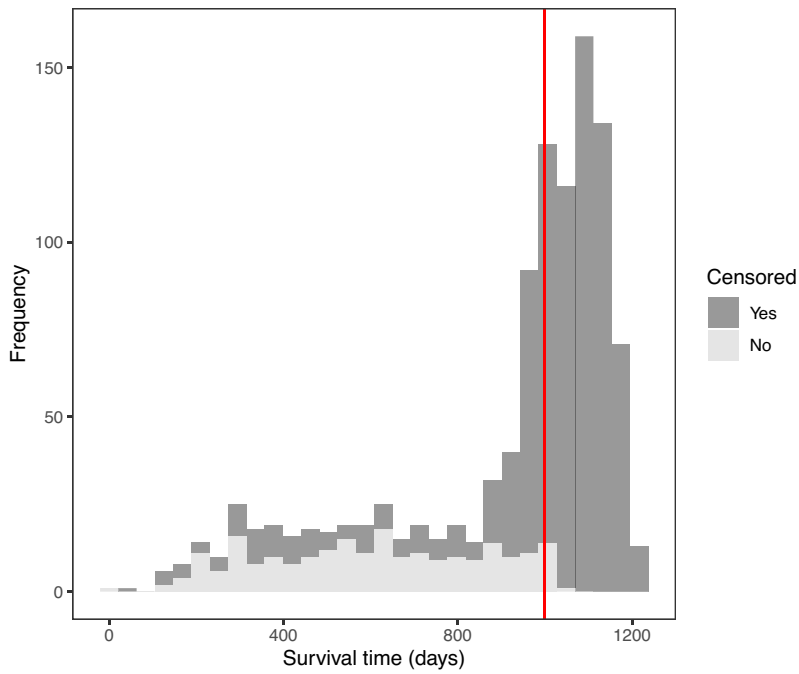

Figure 3. Histogram of survival time for censored/non-censored subjects. The solid red line is the suggested truncation time $h = 1,000$. The data is from the AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al., 1996) with control set to the *didanosine* treatment with 561 subjects and the treatment group to *zidovudine +didanosine* with 522 subjects.

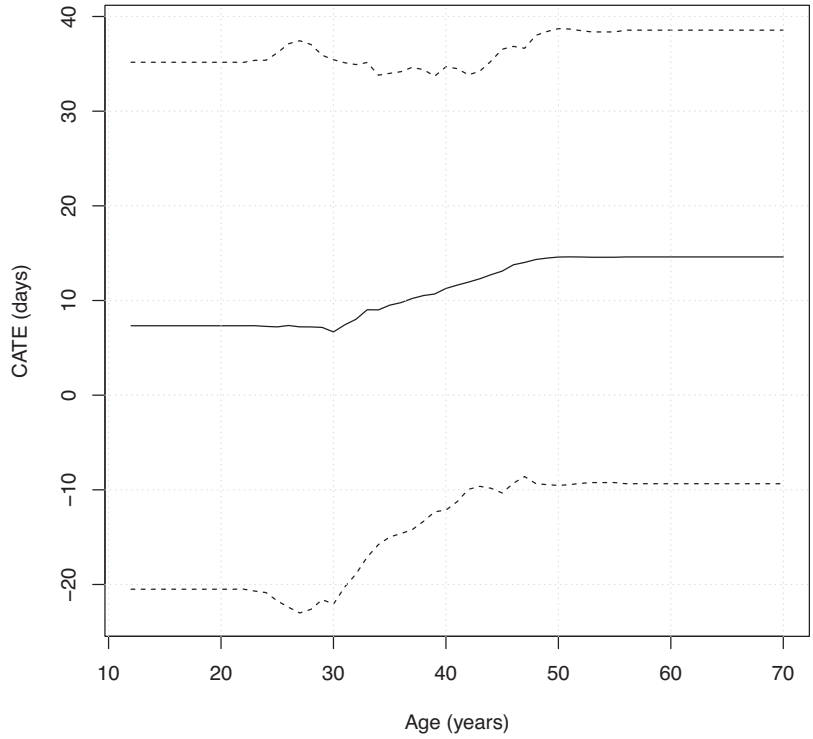

Figure 4. Estimated conditional average treatment effects (CATEs) (survival time in days) versus age (in years) and 95% confidence bars (dash lines), with all other covariates set to their median value. CATEs are estimated using a causal survival forest with default options (the number of trees is set to 10,000 for confidence intervals) and the survival time threshold at 1, 000 days. The data is from AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al., 1996) with control set to the *didanosine* treatment with 561 subjects and the treatment group to *zidovudine +didanosine* with 522 subjects. 14 covariates are included in the analysis, five continuous: age (year), weight (kg), Karnofsky score (scale of 0–100), CD4 count (cells/mm^3) at baseline, and at 20 ± 5 weeks, CD8 count (cells/mm^3) at baseline, and at 20 ± 5 weeks, and seven binary: gender (male = 1, female = 0), homosexual activity (yes = 1, no = 0), race (non-white = 1, white = 0), symptomatic status (symptomatic = 1, asymptomatic = 0), history of intravenous drug use (yes = 1, no = 0), haemophilia (yes = 1, no = 0), and antiretroviral history (experienced = 1, naive = 0).

outcome considered here is the survival time, we also include CD4 count (cells/mm^3) at 20 ± 5 weeks and CD8 count (cells/mm^3) at 20 ± 5 weeks as covariates, in addition to the 12 covariates described above.

We applied the proposed CSF to this dataset. We used the default tuning parameters, described in Section 4, with the exception of the number of trees which is set to $B = 10,000$ for computing confidence intervals. As mentioned in Section 2, Assumption 6 warrants some extra consideration. The high follow-up time in this study requires care in focusing on an appropriate estimand, namely a suitable h for $T_i \wedge h$. This study includes a large amount of near end-time censored subjects observed up to almost 6 months after the last failure, which occurs at around 3 years. For this reason, we truncate the survival time right before 3 years, setting $h = 1,000$. This assures us that the estimated censoring probabilities $\hat{S}^{C}_{W_i}(U_i \wedge h \mid X_i)$ all lie in a reasonable range and suggest that we are in a regime where this identifying assumption holds. Figure 3 shows a histogram depicting the issue. In the raw data, the observations with the largest values of T_i are all censored, and so moments of T_i are not identified. However, given a focus on restricted survival time with a judicious choice of h, we get to re-code all (censored or uncensored) observations with $U_i > h$ as uncensored observations with $T_i = h$, thus eliminating the positivity problem.

Following W. Lu et al. (2013) and C. Fan et al. (2017), we consider what role age serves in treatment efficacy. Figure 4 shows the estimated CATEs against age with all other covariates

set to their median value, and as the plot suggests, age appears to have a positive effect, with older patients benefiting more from the treatment ZDV+ddI. Table 4 shows point estimates and standard errors from a random sample of patients. We also consider the BLP proposed in Section 3.2, in particular, we regress the obtained doubly robust scores on all covariates, and

Table 4. Conditional average treatment effect (CATE) estimates and standard errors from a random sample of 10 individuals as well as patient characteristics corresponding to the four covariates with the highest split frequency obtained from fitting a causal survival forest with default options (survival time threshold at 1, 000 days) on data from AIDS Clinical Trials Group Protocol 175

	All covariates	Age only
Constant	−151.65 (240.10)	−25.85 (48.49)
Age	1.26 (1.46)	1.01 (1.35)
Weight	−1.99 (1.02)	
Karnofsky score	2.75 (2.13)	
CD4 count	−0.06 (0.13)	
CD8 count	0.05 (0.05)	
Gender	44.15 (41.90)	
Homosexual activity	−33.08 (34.47)	
Race	31.61 (25.61)	
Symptomatic status	−10.08 (35.40)	
Intravenous drug use	48.36 (31.94)	
Haemophilia	−63.46 (44.97)	
Antiretroviral history	−7.57 (24.11)	
CD4 count 20 ± 5 weeks	−0.03 (0.11)	
CD8 count 20 ± 5 weeks	−0.04 (0.04)	

Note. The data is from Hammer et al. (1996) with control set to the *didanosine* treatment with 561 subjects and the treatment group to *zidovudine+didanosine* with 522 subjects. 14 covariates are included in the analysis, five continuous: age (year), weight (kg), Karnofsky score (scale of 0–100), CD4 count (cells/mm^3) at baseline, and at 20 ± 5 weeks, CD8 count (cells/mm^3) at baseline, and at 20 ± 5 weeks, and seven binary: gender (male = 1, female = 0), homosexual activity (yes = 1, no = 0), race (non-white = 1, white = 0), symptomatic status (symptomatic = 1, asymptomatic = 0), history of intravenous drug use (yes = 1, no = 0), haemophilia (yes = 1, no = 0), and antiretroviral history (experienced = 1, naive = 0). HC_3 (MacKinnon & White, 1985) standard errors in parentheses.

Table 5. Best linear projection $\hat{\tau}_{DR}(x) = \hat{\beta}_0 + A\hat{\beta}$ on doubly robust estimates obtained from fitting a causal survival forest with default options (survival time threshold at 1, 000 days) on data from AIDS Clinical Trials Group Protocol 175

CATE	se(CATE)	Haemophilia	Gender	Homosexual activity	Antiretroviral history
−7.90	10.45	No	Male	Yes	Experienced
−6.90	13.89	No	Male	Yes	Naive
−4.38	17.06	No	Male	Yes	Naive
0.43	17.28	No	Male	No	Naive
7.26	19.17	No	Male	No	Naive
8.20	11.45	No	Male	Yes	Naive
12.80	13.82	No	Female	Yes	Naive
13.42	14.16	No	Male	Yes	Naive
14.95	14.13	No	Male	Yes	Naive
19.99	43.53	Yes	Male	No	Experienced

Note. The data is from Hammer et al. (1996) with control set to the *didanosine* treatment with 561 subjects and the treatment group to *zidovudine+didanosine* with 522 subjects.

age only. The results in Table 5 suggest that we should exercise caution in interpreting heterogeneity in $\hat{\tau}(\cdot)$, as none of the coefficients in the considered projections is significantly from zero.

Acknowledgments

We are thankful to the three referees, associate editor, and editor for helpful comments which led to an improved manuscript. We are grateful to Susan Athey, Scott Fleming, Vitor Hadad, David Hirshberg, Ayush Kanodia, Julie Tibshirani, Yizhe Xu, and Steve Yadlowsky for helpful conversations and suggestions. We also particularly thank Julie Tibshirani for performing a code review of our implementation and helping us merge it into `grf`, and David Hirshberg for contributing to `grf` the implementation of sample weighting used in our IPCW approach.

Supplementary material

Supplementary data is available online at *Journal of the Royal Statistical Society* online.

Data availability

The HIV dataset used in this paper is publicly available.

Conflict of interest: None declared.

Funding

Yifan Cui is supported in part by the National Natural Science Foundation of China. Ruoqing Zhu is supported in part by the National Science Foundation grant DMS-2210657.

Appendix

A.1 Proof of Proposition 1

To simplify the notation, we write $T \wedge h$ as $\tilde{T}$ and $U \wedge h$ as $\tilde{U}$ in this section. We first start with a proof in the context of estimating $\mu = \mu(x) = \mathbb{E}[y(T)]$ by the estimating equation (9), which might also be of independent interest. Namely, we show that

$$\hat{\mu} - \tilde{\mu} = o_p(\max(c_n^2, c_n d_n)), \tag{A1}$$

where $\tilde{\mu}$ is an oracle estimator for μ.

Proof of (A1). At a high level, cross-fitting uses cross-fold estimation to avoid bias due to overfitting. Recall that the cross-fitting first splits the data (at random) into two halves I_1 and I_2, and then uses an estimator

$$\hat{\mu} = \frac{n_1}{n}\hat{\mu}^{I_1} + \frac{n_2}{n}\hat{\mu}^{I_2},$$

where $n_1 = |I_1|$, $n_2 = |I_2|$, and $\hat{\mu}^{I_1}$, $\hat{\mu}^{I_2}$ are estimated using nuisances estimated from samples I_2, I_1, respectively. We essentially need to show that

$$\hat{\mu}^{I_1} - \tilde{\mu}^{I_1} = o_p(\max(c_n^2, c_n d_n)),$$

where $\tilde{\mu}^{I_1}$ is the oracle estimator obtained by solving

$$\frac{1}{n_1}\sum_{i:i\in I_1}\frac{\Delta_i^h(y(U_i)-\mu)}{S^C(\tilde{U}_i \mid X_i)} + \frac{(1-\Delta_i^h)}{S^C(\tilde{U}_i \mid X_i)}\mathbb{E}[y(T_i)-\mu \mid X_i, \tilde{T}_i > \tilde{U}_i]$$

$$-\int_0^{\tilde{U}_i}\frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}\mathbb{E}[y(T_i)-\mu \mid X_i, \tilde{T}_i > s]\,ds = 0.$$

Note that we have the following decomposition of $\hat{\mu}^{I_1} - \tilde{\mu}^{I_1}$:

$$\begin{aligned}
\hat{\mu}^{I_1} - \tilde{\mu}^{I_1} &= \frac{1}{n_1}\left[\sum_{i:i\in I_1} \Delta_i^b \frac{y(U_i)}{\hat{S}^C(\tilde{U}_i \mid X_i)} + (1-\Delta_i^b)\frac{\hat{Q}(\tilde{U}_i \mid X_i)}{\hat{S}^C(\tilde{U}_i \mid X_i)} - \int_0^{\tilde{U}_i} \frac{\hat{\lambda}^C(s \mid X_i)}{\hat{S}^C(s \mid X_i)} \hat{Q}(s \mid X_i)\, ds\right] \\
&\quad - \frac{1}{n_1}\left[\sum_{i:i\in I_1} \Delta_i^b \frac{y(U_i)}{S^C(\tilde{U}_i \mid X_i)} + (1-\Delta_i^b)\frac{Q(\tilde{U}_i \mid X_i)}{S^C(\tilde{U}_i \mid X_i)} - \int_0^{\tilde{U}_i} \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)} Q(s \mid X_i)\, ds\right] \\
&= \frac{1}{n_1}\left[\sum_{i:i\in I_1} \frac{(1-\Delta_i^b)}{S^C(\tilde{U}_i \mid X_i)}(\hat{Q}(\tilde{U}_i \mid X_i) - Q(\tilde{U}_i \mid X_i)) - \int_0^{\tilde{U}_i} \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}(\hat{Q}(s \mid X_i) - Q(s \mid X_i))\, ds\right. \\
&\quad + (1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i \mid X_i)} - \frac{1}{S^C(\tilde{U}_i \mid X_i)}\right)(\hat{Q}(\tilde{U}_i \mid X_i) - Q(\tilde{U}_i \mid X_i)) \\
&\quad - \int_0^{\tilde{U}_i}\left(\frac{\hat{\lambda}^C(s \mid X_i)}{\hat{S}^C(s \mid X_i)} - \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}\right)(\hat{Q}(s \mid X_i) - Q(s \mid X_i))\, ds \\
&\quad + (1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i \mid X_i)} - \frac{1}{S^C(\tilde{U}_i \mid X_i)}\right) Q(\tilde{U}_i \mid X_i) \\
&\quad + \Delta_i^b\left(\frac{1}{\hat{S}^C(\tilde{U}_i \mid X_i)} - \frac{1}{S^C(\tilde{U}_i \mid X_i)}\right) y(U_i) \\
&\quad \left. - \int_0^{\tilde{U}_i}\left(\frac{\hat{\lambda}^C(s \mid X_i)}{\hat{S}^C(s \mid X_i)} - \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}\right) Q(s \mid X_i)\, ds\right],
\end{aligned} \tag{A2}$$

where we denote $\widehat{\mathbb{E}}[y(T_i)|\tilde{T}_i > \tilde{U}_i, X_i]$ by $\hat{Q}(\tilde{U}_i \mid X_i)$, and $\mathbb{E}[y(T_i)|\tilde{T}_i > \tilde{U}_i, X_i]$ by $Q(\tilde{U}_i \mid X_i)$. At a high level, this decomposition separates $\hat{\mu}^{I_1} - \tilde{\mu}^{I_1}$ to four terms: two mean zero terms and two product terms.

Note that by the double robustness of equation (9) shown in Tsiatis (2007, Chapter 10.4),

$$\mathbb{E}\left[\frac{(1-\Delta_i^b)}{S^C(\tilde{U}_i \mid X_i)}(\hat{Q}(\tilde{U}_i \mid X_i) - Q(\tilde{U}_i \mid X_i)) - \int_0^{\tilde{U}_i} \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}(\hat{Q}(s \mid X_i) - Q(s \mid X_i))\, ds\right] = 0.$$

Thanks to our cross-fitting construction, the nuisance components can effectively be treated as deterministic. Thus, after conditioning on I_2, the summands used to build the following term become mean zero and independent:

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n_1}\sum_{i:i\in I_1} \frac{(1-\Delta_i^b)}{S^C(\tilde{U}_i \mid X_i)}(\hat{Q}(\tilde{U}_i \mid X_i) - Q(\tilde{U}_i \mid X_i)) - \int_0^{\tilde{U}_i} \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}(\hat{Q}(s \mid X_i) - Q(s \mid X_i))\, ds\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{n_1}\sum_{i:i\in I_1} \frac{(1-\Delta_i^b)}{S^C(\tilde{U}_i \mid X_i)}(\hat{Q}(\tilde{U}_i \mid X_i) - Q(\tilde{U}_i \mid X_i)) - \int_0^{\tilde{U}_i} \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}(\hat{Q}(s \mid X_i) - Q(s \mid X_i))\, ds\right)^2 \middle| I_2\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left[\left(\frac{1}{n_1}\sum_{i:i\in I_1} \frac{(1-\Delta_i^b)}{S^C(\tilde{U}_i \mid X_i)}(\hat{Q}(\tilde{U}_i \mid X_i) - Q(\tilde{U}_i \mid X_i)) - \int_0^{\tilde{U}_i} \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}(\hat{Q}(s \mid X_i) - Q(s \mid X_i))\, ds\right) \middle| I_2\right]\right] \\
&= \frac{1}{n_1}\mathbb{E}\left[\mathrm{Var}\left[\left(\frac{(1-\Delta_i^b)}{S^C(\tilde{U}_i \mid X_i)}(\hat{Q}(\tilde{U}_i \mid X_i) - Q(\tilde{U}_i \mid X_i)) - \int_0^{\tilde{U}_i} \frac{\lambda^C(s \mid X_i)}{S^C(s \mid X_i)}(\hat{Q}(s \mid X_i) - Q(s \mid X_i))\, ds\right) \middle| I_2\right]\right] \\
&\le \frac{O_p(1)}{n_1} \sup_{x\in\mathcal{X}, s\le b} |\hat{Q}(s \mid x) - Q(s \mid x)|^2 = \frac{o_p(1)}{n}.
\end{aligned} \tag{A3}$$

The same logic applies to the term

$$(1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)Q(\tilde{U}_i\,|\,X_i)+\Delta_i^b\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)y(U_i)$$
$$-\int_0^{\tilde{U}_i}\left(\frac{\hat{\lambda}^C(s\,|\,X_i)}{\hat{S}^C(s\,|\,X_i)}-\frac{\lambda^C(s\,|\,X_i)}{S^C(s\,|\,X_i)}\right)Q(s\,|\,X_i)\,ds,$$

as

$$\mathbb{E}\left((1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)Q(\tilde{U}_i\,|\,X_i)+\Delta_i^b\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)y(U_i)\right.$$
$$\left.-\int_0^{\tilde{U}_i}\left(\frac{\hat{\lambda}^C(s\,|\,X_i)}{\hat{S}^C(s\,|\,X_i)}-\frac{\lambda^C(s\,|\,X_i)}{S^C(s\,|\,X_i)}\right)Q(s\,|\,X_i)\,ds\right)=0,$$

so we have that

$$\mathbb{E}\left(\frac{1}{n_1}\sum_{i:i\in I_1}(1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)Q(\tilde{U}_i\,|\,X_i)+\Delta_i^b\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)y(U_i)\right.$$
$$\left.-\int_0^{\tilde{U}_i}\left(\frac{\hat{\lambda}^C(s\,|\,X_i)}{\hat{S}^C(s\,|\,X_i)}-\frac{\lambda^C(s\,|\,X_i)}{S^C(s\,|\,X_i)}\right)Q(s\,|\,X_i)\,ds\right)\le\frac{o_p(1)}{n}. \tag{A4}$$

In addition, by Cauchy–Schwarz inequality,

$$\frac{1}{n_1}\sum_{i:i\in I_1}\left((1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)(\hat{Q}(\tilde{U}_i\,|\,X_i)-Q(\tilde{U}_i\,|\,X_i))\right.$$
$$\left.-\int_0^{\tilde{U}_i}\left(\frac{\hat{\lambda}^C(s\,|\,X_i)}{\hat{S}^C(s\,|\,X_i)}-\frac{\lambda^C(s\,|\,X_i)}{S^C(s\,|\,X_i)}\right)(\hat{Q}(s\,|\,X_i)-Q(s\,|\,X_i))\,ds\right)$$
$$\le\sqrt{\frac{1}{n_1}\sum_{i:i\in I_1}(1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)^2}\times\sqrt{\frac{1}{n_1}\sum_{i:i\in I_1}(1-\Delta_i^b)(\hat{Q}(\tilde{U}_i\,|\,X_i)-Q(\tilde{U}_i\,|\,X_i))^2} \tag{A5}$$
$$+\int_0^{\tilde{U}_i}\sqrt{\frac{1}{n_1}\sum_{i:i\in I_1}\left(\frac{\hat{\lambda}^C(s\,|\,X_i)}{\hat{S}^C(s\,|\,X_i)}-\frac{\lambda^C(s\,|\,X_i)}{S^C(s\,|\,X_i)}\right)^2}\times\sqrt{\frac{1}{n_1}\sum_{i:i\in I_1}(\hat{Q}(s\,|\,X_i)-Q(s\,|\,X_i))^2}\,ds$$
$$=o_p(\max(c_n^2,c_nd_n)).$$

Therefore, combining equations (A3), (A4), and (A5), we have that $\hat{\mu}-\tilde{\mu}=o_p(\max(c_n^2,c_nd_n))$. □

Now, we turn to estimating $\tau=\tau(x)=\mathbb{E}[y(T(1))-y(T(0))]$.

Proof of Proposition 1. Using the same notation as the proof of equation (A1) and consider an estimator

$$\hat{\tau}=\frac{n_1}{n}\hat{\tau}^{I_1}+\frac{n_2}{n}\hat{\tau}^{I_2},$$

where $n_1=|I_1|$, $n_2=|I_2|$, and $\hat{\tau}^{I_1}$, $\hat{\tau}^{I_2}$ are estimated using nuisances estimated from samples I_2, I_1, respectively. Note that

$$\hat{\tau}^{I_1}-\tilde{\tau}^{I_1}=\left(\frac{1}{n_1}\sum_{i:i\in I_1}(W_i-\hat{e}(X_i))^2\right)^{-1}$$
$$\times\frac{1}{n_1}\sum_{i:i\in I_1}\left(\frac{\hat{\Upsilon}_i}{\hat{S}^C_{W_i}(\tilde{U}_i\,|\,X_i)}-\int_0^{\tilde{U}_i}\frac{\hat{\lambda}^C_{W_i}(s\,|\,X_i)}{\hat{S}^C_{W_i}(s\,|\,X_i)}(W_i-\hat{e}(X_i))[\hat{Q}_{W_i}(s\,|\,X_i)-\hat{m}(X_i)]\,ds\right)$$
$$-\left(\frac{1}{n_1}\sum_{i:i\in I_1}(W_i-e(X_i))^2\right)^{-1}$$
$$\times\frac{1}{n_1}\sum_{i:i\in I_1}\left(\frac{\Upsilon_i}{S^C_{W_i}(\tilde{U}_i\,|\,X_i)}-\int_0^{\tilde{U}_i}\frac{\lambda^C_{W_i}(s\,|\,X_i)}{S^C_{W_i}(s\,|\,X_i)}(W_i-e(X_i))[Q_{W_i}(s\,|\,X_i)-m(X_i)]\,ds\right),$$

where $\tilde{\tau}^{I_1}$ is the oracle estimator,

$$\Upsilon_i = \begin{cases} \{W_i - e(X_i)\}\{y(U_i) - m(X_i)\} & \text{if } \Delta_i^b = 1, \\ \{W_i - e(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - m(X_i)\} & \text{o.w.} \end{cases}$$

and

$$\hat{\Upsilon}_i = \begin{cases} \{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\} & \text{if } \Delta_i^b = 1, \\ \{W_i - \hat{e}(X_i)\}\{\hat{Q}_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\} & \text{o.w.} \end{cases}$$

Denote $\hat{\tau}^{I_1} - \tilde{\tau}^{I_1}$ by

$$\frac{K_1}{J_1} - \frac{K_2}{J_2}.$$

Thus, we have

$$\begin{aligned} \frac{K_1}{J_1} - \frac{K_2}{J_2} &= \frac{K_1 + K_2 - K_2}{J_1} - \frac{K_2}{J_2} \\ &= \frac{K_1 - K_2}{J_1} + K_2\left(\frac{1}{J_1} - \frac{1}{J_2}\right) = \frac{K_1 - K_2}{J_1} + \frac{K_2}{J_1 J_2}(J_2 - J_1). \end{aligned} \tag{A6}$$

We essentially need to bound $K_1 - K_2$, and we have the following decomposition:

$$\begin{aligned} K_1 - K_2 = \frac{1}{n_1}\Bigg[& \sum_{i:i\in I_1} \Delta_i^b \frac{\{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\}}{\hat{S}^C_{W_i}(\tilde{U}_i \mid X_i)} + (1 - \Delta_i^b)\frac{\{W_i - \hat{e}(X_i)\}\{\hat{Q}_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}}{\hat{S}^C_{W_i}(\tilde{U}_i \mid X_i)} \\ & - \int_0^{\tilde{U}_i} \frac{\hat{\lambda}^C_{W_i}(s \mid X_i)}{\hat{S}^C_{W_i}(s \mid X_i)}(W_i - \hat{e}(X_i))[\hat{Q}_{W_i}(s \mid X_i) - \hat{m}(X_i)]\, ds \\ & - \Delta_i^b \frac{\{W_i - e(X_i)\}\{y(U_i) - m(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} - (1 - \Delta_i^b)\frac{\{W_i - e(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - m(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} \\ & + \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i - e(X_i))[Q_{W_i}(s \mid X_i) - m(X_i)]\, ds\Bigg] \\ = \frac{1}{n_1}\Bigg[& \sum_{i:i\in I_1} \Delta_i^b \frac{\{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\}}{\hat{S}^C_{W_i}(\tilde{U}_i \mid X_i)} + (1 - \Delta_i^b)\frac{\{W_i - \hat{e}(X_i)\}\{\hat{Q}_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}}{\hat{S}^C_{W_i}(\tilde{U}_i \mid X_i)} \\ & - \int_0^{\tilde{U}_i} \frac{\hat{\lambda}^C_{W_i}(s \mid X_i)}{\hat{S}^C_{W_i}(s \mid X_i)}(W_i - \hat{e}(X_i))[\hat{Q}_{W_i}(s \mid X_i) - \hat{m}(X_i)]\, ds \\ & - \Delta_i^b \frac{\{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} - (1 - \Delta_i^b)\frac{\{W_i - \hat{e}(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} \\ & + \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i - \hat{e}(X_i))[Q_{W_i}(s \mid X_i) - \hat{m}(X_i)]\, ds \\ & + \Delta_i^b \frac{\{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} + (1 - \Delta_i^b)\frac{\{W_i - \hat{e}(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} \\ & - \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i - \hat{e}(X_i))[Q_{W_i}(s \mid X_i) - \hat{m}(X_i)]\, ds \\ & - \Delta_i^b \frac{\{W_i - e(X_i)\}\{y(U_i) - m(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} - (1 - \Delta_i^b)\frac{\{W_i - e(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - m(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} \\ & + \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i - e(X_i))[Q_{W_i}(s \mid X_i) - m(X_i)]\, ds\Bigg]. \end{aligned}$$

At a high level, this decomposition separates $K_1 - K_2$ into two terms: the first term takes $\hat{m}$ and $\hat{e}$ as given, and follow the construction in the survival-related nuisance components (A2) to bound errors caused by using $\hat{\lambda}^C_w$, $\hat{S}^C_w$, $\hat{Q}_w$ instead of λ^C_w, S^C_w, Q_w; the second term bounds errors caused by using $\hat{m}$ and $\hat{e}$ instead of m and e.

For the term

$$\begin{aligned}
&\Delta_i^b \frac{\{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\}}{\hat{S}^C_{W_i}(\tilde{U}_i \mid X_i)} + (1-\Delta_i^b)\frac{\{W_i - \hat{e}(X_i)\}\{\hat{Q}_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}}{\hat{S}^C_{W_i}(\tilde{U}_i \mid X_i)} \\
&- \int_0^{\tilde{U}_i} \frac{\hat{\lambda}^C_{W_i}(s \mid X_i)}{\hat{S}^C_{W_i}(s \mid X_i)}(W_i - \hat{e}(X_i))[\hat{Q}_{W_i}(s \mid X_i) - \hat{m}(X_i)]\, ds \\
&- \Delta_i^b \frac{\{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\}}{S^C_{W_i}(U_i|X_i)} - (1-\Delta_i^b)\frac{\{W_i - \hat{e}(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} \\
&+ \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i - \hat{e}(X_i))[Q_{W_i}(s \mid X_i) - \hat{m}(X_i)]\, ds \equiv (I),
\end{aligned}$$

the proof follows from the same decomposition of equation (A2) with $\tilde{U}_i$ replaced by $\{W_i - \hat{e}(X_i)\}\{\tilde{U}_i - \hat{m}(X_i)\}$ and $\hat{Q}(\tilde{U}_i \mid X_i)$ replaced by $\{W_i - \hat{e}(X_i)\}\{\hat{Q}_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}$. So we have that

$$(I) = o_p(c_n^2 + c_n d_n). \tag{A7}$$

For the term

$$\begin{aligned}
&\Delta_i^b \frac{\{W_i - \hat{e}(X_i)\}\{y(U_i) - \hat{m}(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} + (1-\Delta_i^b)\frac{\{W_i - \hat{e}(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - \hat{m}(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} \\
&- \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i - \hat{e}(X_i))[Q_{W_i}(s \mid X_i) - \hat{m}(X_i)]\, ds \\
&- \Delta_i^b \frac{\{W_i - e(X_i)\}\{y(U_i) - m(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} - (1-\Delta_i^b)\frac{\{W_i - e(X_i)\}\{Q_{W_i}(\tilde{U}_i \mid X_i) - m(X_i)\}}{S^C_{W_i}(\tilde{U}_i \mid X_i)} \\
&+ \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i - e(X_i))[Q_{W_i}(s \mid X_i) - m(X_i)]\, ds \equiv (II),
\end{aligned}$$

we have the following decomposition of (II);

$$\begin{aligned}
&\frac{\Delta_i^b}{S^C_{W_i}(\tilde{U}_i \mid X_i)}[(e(X_i) - \hat{e}(X_i))(m(X_i) - \hat{m}(X_i)) + (e(X_i) - \hat{e}(X_i))(y(U_i) - m(X_i)) \\
&+ (W_i - e(X_i))(m(X_i) - \hat{m}(X_i))] + \frac{(1-\Delta_i^b)}{S^C_{W_i}(\tilde{U}_i \mid X_i)}[(e(X_i) - \hat{e}(X_i))(m(X_i) - \hat{m}(X_i)) \\
&+ (e(X_i) - \hat{e}(X_i))(Q_{W_i}(\tilde{U}_i \mid X_i) - m(X_i)) + (W_i - e(X_i))(m(X_i) - \hat{m}(X_i))] \\
&- \int_0^{\tilde{U}_i} \frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}[(e(X_i) - \hat{e}(X_i))(m(X_i) - \hat{m}(X_i)) + (e(X_i) - \hat{e}(X_i))(Q_{W_i}(s \mid X_i) - m(X_i)) \\
&+ (W_i - e(X_i))(m(X_i) - \hat{m}(X_i))]\, ds.
\end{aligned}$$

Note that

$$\mathbb{E}\left[\frac{\Delta_i^b}{S^C_{W_i}(\tilde{U}_i \mid X_i)}(e(X_i)-\hat{e}(X_i))(y(U_i)-m(X_i))\right]=0,$$
$$\mathbb{E}\left[\frac{1}{S^C_{W_i}(\tilde{U}_i \mid X_i)}(W_i-e(X_i))(m(X_i)-\hat{m}(X_i))-\int_0^{\tilde{U}_i}\frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(W_i-e(X_i))(m(X_i)-\hat{m}(X_i))\,ds\right]=0,$$

and

$$\mathbb{E}\left[\frac{(1-\Delta_i^b)}{S^C_{W_i}(\tilde{U}_i \mid X_i)}(e(X_i)-\hat{e}(X_i))(Q_{W_i}(\tilde{U}_i \mid X_i)-m(X_i))\right.$$
$$\left.-\int_0^{\tilde{U}_i}\frac{\lambda^C_{W_i}(s \mid X_i)}{S^C_{W_i}(s \mid X_i)}(e(X_i)-\hat{e}(X_i))(Q_{W_i}(s \mid X_i)-m(X_i))\,ds\right]=0.$$

Again, by the law of iterated expectations similar to that of (A3) using cross-fitting technique and Cauchy–Schwarz inequality in (A5), we have that

$$(II)=o_p(b_nc_n). \tag{A8}$$

Combining equations (A7) and (A8), we have that $K_1-K_2=o_p(c_nd_n+c_n^2+b_nc_n)$. Further combining the rates of K_1-K_2 and J_2-J_1 using equation (A6), we have that

$$\hat{\tau}^{I_1}-\tilde{\tau}^{I_1}=o_p(\max((c_n+d_n)c_n,\, b_nc_n,\, b_n^2)),$$

and therefore $\hat{\tau}-\tilde{\tau}=o_p(\max((c_n+d_n)c_n,\, b_nc_n,\, b_n^2))$. Recall that $\tilde{\tau}$ is an i.i.d. average, so we immediately have that by using cross-fitting, we transform any $n^{1/4}$ consistent machine learning method into an efficient estimator of $\tau=\mathbb{E}[y(T(1))-y(T(0))]$. □

A.2 Proof of Lemma 2

Proof. Note that

$$\hat{\tau}(x)-\tilde{\tau}(x)=\left(\frac{1}{l}\sum_{i=1}^{l}(W_i-\hat{e}(X_i))^2\right)^{-1}$$
$$\times\frac{1}{l}\sum_{i=1}^{l}\left(\frac{\hat{\Upsilon}_i}{\hat{S}^C_{W_i}(U_i\wedge h|X_i)}-\int_0^{U_i\wedge h}\frac{\hat{\lambda}^C_{W_i}(t \mid X_i)}{\hat{S}^C_{W_i}(t \mid X_i)}(W_i-\hat{e}(X_i))[\hat{Q}_{W_i}(t \mid X_i)-\hat{m}(X_i)]\,dt\right)$$
$$-\left(\frac{1}{l}\sum_{i=1}^{l}(W_i-e(X_i))^2\right)^{-1}$$
$$\times\frac{1}{l}\sum_{i=1}^{l}\left(\frac{\Upsilon_i}{S^C_{W_i}(U_i\wedge h|X_i)}-\int_0^{U_i\wedge h}\frac{\lambda^C_{W_i}(t \mid X_i)}{S^C_{W_i}(t \mid X_i)}(W_i-e(X_i))[Q_{W_i}(t \mid X_i)-m(X_i)]\,dt\right),$$

where Υ_i, $\hat{\Upsilon}_i$ are defined in Section A.1 of the Appendix, l is the number of observations falling into the same terminal node as x, and the weights $\alpha_i(x)$ are absorbed in l. Following the same proof of Proposition 1 except the Cauchy–Schwarz

inequality such as equation (A5) becomes

$$\begin{aligned}
&\frac{1}{l}\sum_{i=1}^{l}\Bigg((1-\Delta_i^b)\left(\frac{1}{\hat{S}^C(\tilde{U}_i\,|\,X_i)}-\frac{1}{S^C(\tilde{U}_i\,|\,X_i)}\right)(\hat{Q}(\tilde{U}_i\,|\,X_i)-Q(\tilde{U}_i\,|\,X_i))\\
&\quad-\int_0^{\tilde{U}_i}\left(\frac{\hat{\lambda}^C(s\,|\,X_i)}{\hat{S}^C(s\,|\,X_i)}-\frac{\lambda^C(s\,|\,X_i)}{S^C(s\,|\,X_i)}\right)(\hat{Q}(s\,|\,X_i)-Q(s\,|\,X_i))\,ds\Bigg)\\
&\leq\sqrt{\frac{1}{l}\sum_{i=1}^{l}(1-\Delta_i^b)\sup_{x\in\mathcal{X}}\left(\frac{1}{\hat{S}^C(\tilde{U}_i|x)}-\frac{1}{S^C(\tilde{U}_i|x)}\right)^2}\\
&\quad\times\sqrt{\frac{1}{l}\sum_{i=1}^{l}(1-\Delta_i^b)\sup_{x\in\mathcal{X}}(\hat{Q}(\tilde{U}_i|x)-Q(\tilde{U}_i|x))^2}\\
&\quad+\int_0^{\tilde{U}_i}\sqrt{\frac{1}{l}\sum_{i=1}^{l}\sup_{x\in\mathcal{X}}\left(\frac{\hat{\lambda}^C(s|x)}{\hat{S}^C(s|x)}-\frac{\lambda^C(s|x)}{S^C(s|x)}\right)^2}\times\sqrt{\frac{1}{l}\sum_{i=1}^{l}\sup_{x\in\mathcal{X}}(\hat{Q}(s\,|\,x)-Q(s\,|\,x))^2}\,ds\\
&=o_p(\max(c_n^2,c_nd_n)),
\end{aligned}$$

we have that

$$\hat{\tau}(x)-\tilde{\tau}(x)=o_p(\max((c_n+d_n)c_n,b_nc_n,b_n^2)),$$

which completes the proof. □

A.3 Proof of Theorem 3

Proof. Given the set of forest weights $\alpha_i(x)$ used to define the generalized random forest estimation $\tilde{\tau}(x)$ with unknown true nuisance parameters, we have the following linear approximation:

$$\tilde{\tau}^*(x)=\tau(x)+\sum_{i=1}^{n}\alpha_i(x)\rho_i^*(x),$$

where $\rho_j^*(x)$ denotes the influence function of the jth observation with respect to the true parameter value $\tau(x)$, and $\tilde{\tau}^*(x)$ is a pseudo-forest output with weights $\alpha_i(x)$ and outcomes $\tau(x)+\rho_i^*(x)$.

Note that Assumptions 2–6 in Athey et al. (2019) hold immediately from the definition of the estimating equation $\psi_{\tau(x)}$. In particular, $\psi_{\tau(x)}$ is Lipschitz continuous in terms of $\tau(x)$ for their Assumption 4; The solution of $\sum_{i=1}^{n}\alpha_i\psi^i_{\tau(x)}=0$ always exists for their Assumption 5. By the results shown in Wager and Athey (2018), there exists a sequence $\sigma_n(x)$ for which

$$[\tilde{\tau}^*(x)-\tau(x)]/\sigma_n(x)\to N(0,1),$$

where $\sigma_n^2(x)=\text{polylog}(n/\ell)^{-1}\ell/n$ and $\text{polylog}(n/\ell)$ is a function that is bounded away from 0 and increases at most polynomially with the log-inverse sampling ratio $\log(n/\ell)$.

Furthermore, by Lemma 4 in Athey et al. (2019),

$$(n/\ell)^{1/2}[\tilde{\tau}(x)-\tilde{\tau}^*(x)]=O_p\left(\max\left(\ell^{-\frac{\pi\log((1-\nu)^{-1})}{2\log(\nu^{-1})}},\left(\frac{\ell}{n}\right)^{1/6}\right)\right).$$

Following Lemma 2, as long as $o_p(\max((c_n + d_n)c_n, b_n c_n, b_n^2))$ goes faster than $\text{polylog}(n/\ell)^{-1/2}(\ell/n)^{1/2}$, we have

$$[\hat{\tau}(x) - \tau(x)]/\sigma_n(x) \to N(0, 1).$$

□

A.4 Figures

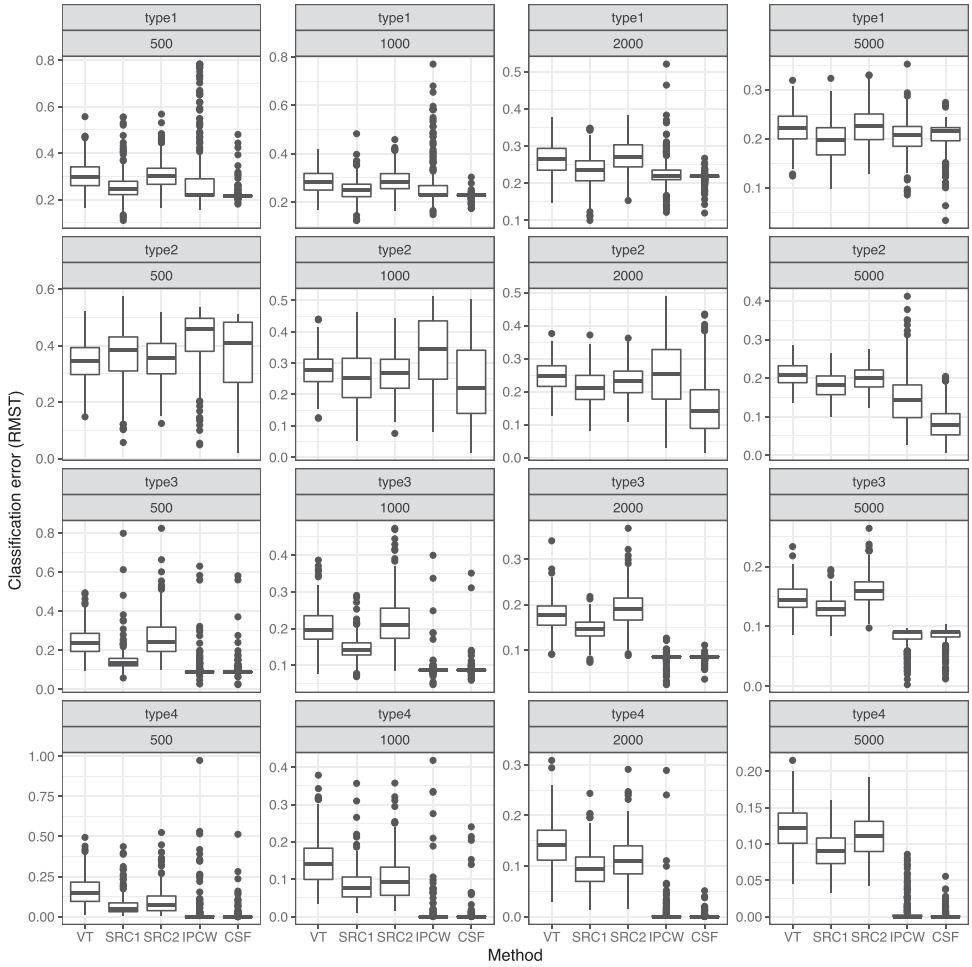

Figure A1. Classification error of different methods with restricted mean survival time (RMST) estimand. From left to right, top to bottom, the plots correspond to Scenarios 1–4, respectively. 'VT' denotes the virtual twin method with random survival forests; 'SRC1' denotes random survival forests using covariates (X, W); 'SRC2' denotes random survival forests using covariates (X, W, XW). 'IPCW' denotes a causal forest with Inverse Probability of Censoring Weighting; 'CSF' denotes a causal survival forest. Training size is (500, 1,000, 2,000, 5,000), the number of covariates 15, the size of the test set 2,000, and the number of repetitions 250.

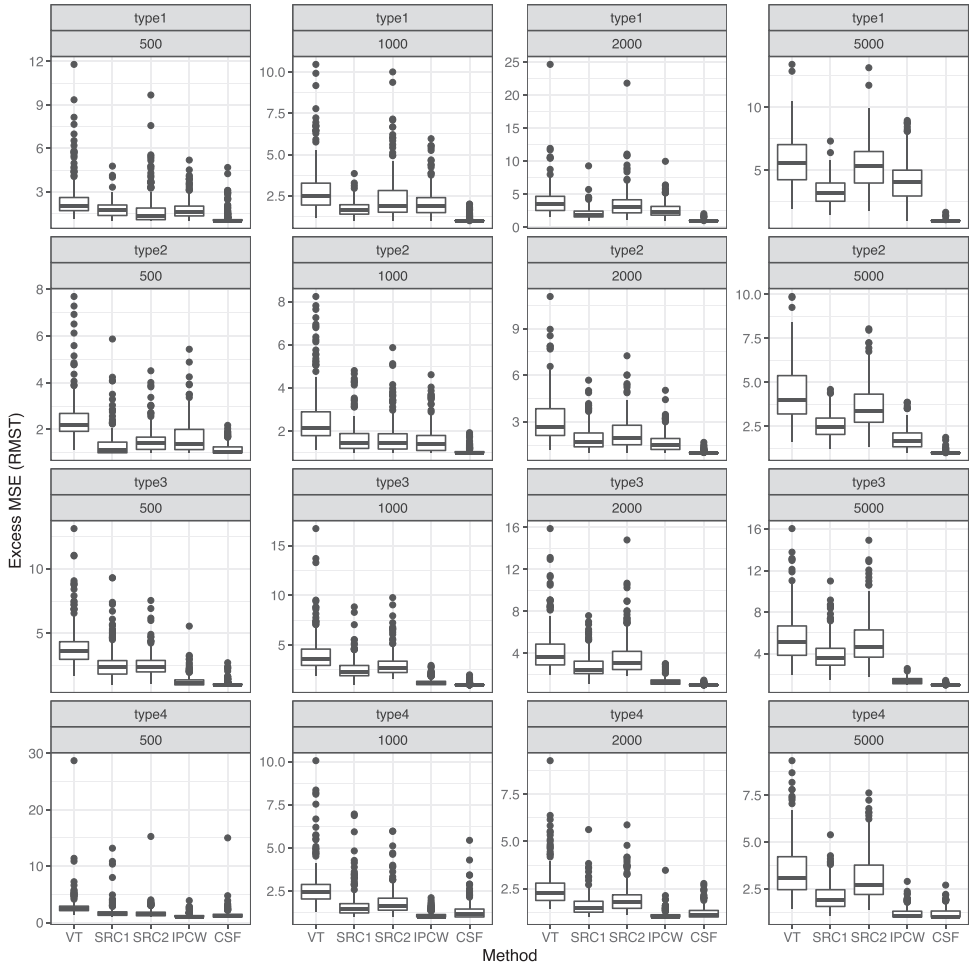

Figure A2. Excess MSE (defined as $(1/B)\sum_{b=1}^{B} \mathrm{MSE}_b(\text{method})/\min(\mathrm{MSE}_b(m)) : m \in \{\text{all methods}\}$) for different methods with restricted mean survival time (RMST) estimand in the four scenarios. From left to right, top to bottom, the plots correspond to Scenarios 1–4, respectively. 'VT' denotes the virtual twin method with random survival forests; 'SRC1' denotes random survival forests using covariates (X, W); 'SRC2' denotes random survival forests using covariates (X, W, XW). 'IPCW' denotes a causal forest with Inverse Probability of Censoring Weighting; 'CSF' denotes a causal survival forest. Training size is (500, 1,000, 2,000, 5,000), the number of covariates 15, the size of the test set 2,000, and the number of repetitions 250.

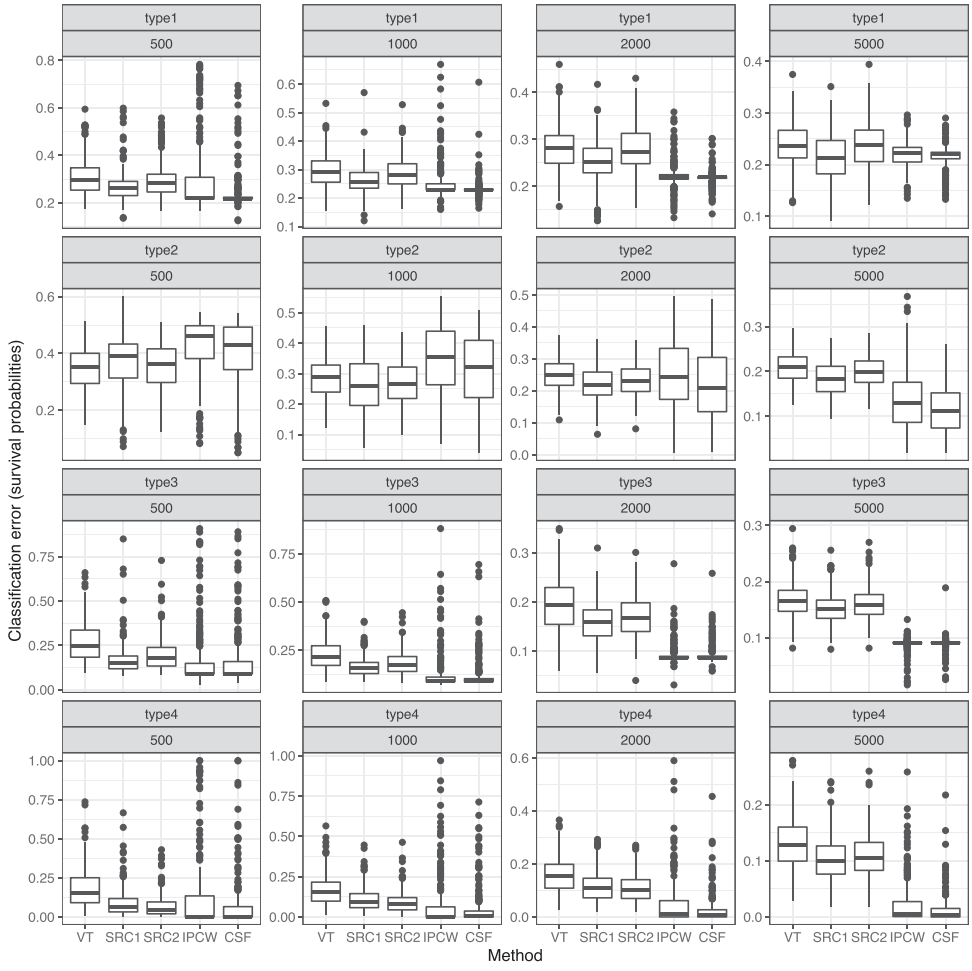

Figure A3. Classification error of different methods with survival probability estimand. From left to right, top to bottom, the plots correspond to Scenarios 1–4, respectively. 'VT' denotes the virtual twin method with random survival forests; 'SRC1' denotes random survival forests using covariates (X, W); 'SRC2' denotes random survival forests using covariates (X, W, XW). 'IPCW' denotes a causal forest with Inverse Probability of Censoring Weighting; 'CSF' denotes a causal survival forest. Training size is (500, 1,000, 2,000, 5,000), the number of covariates 15, the size of the test set 2,000, and the number of repetitions 250.

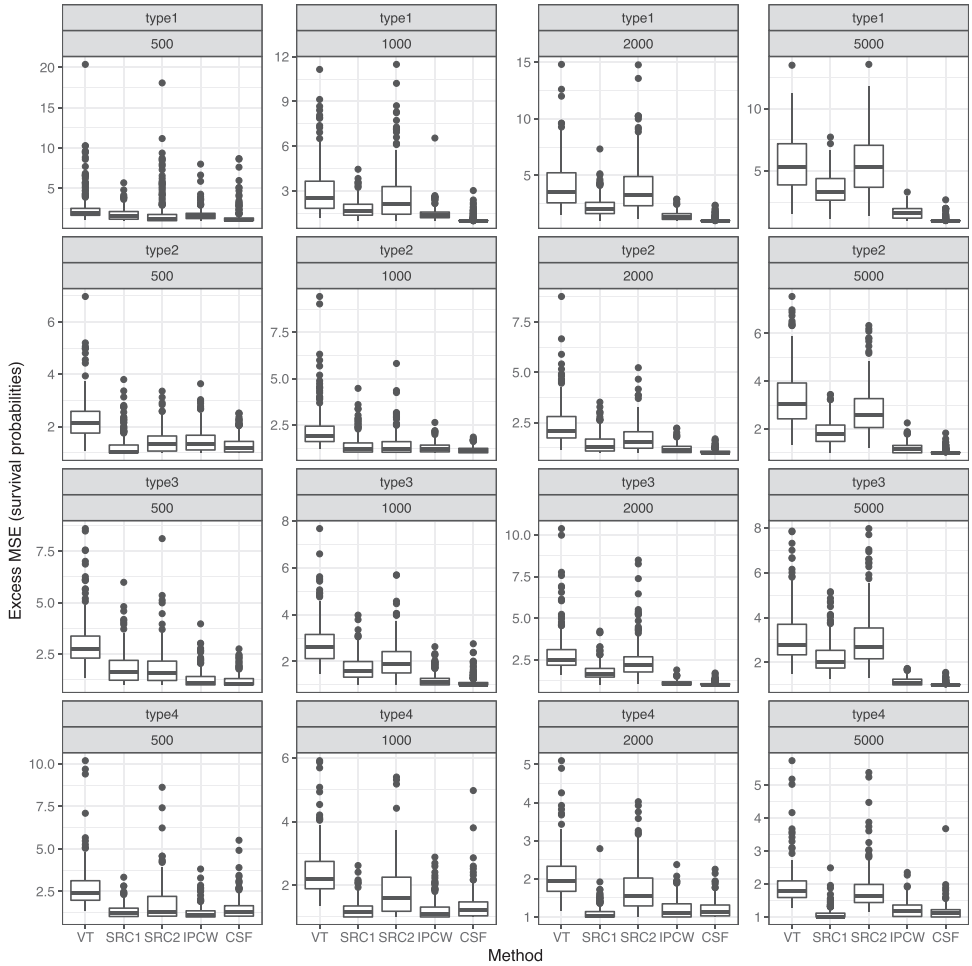

Figure A4. Excess MSE (defined as $(1/B)\sum_{b=1}^{B} \mathrm{MSE}_b(\text{method})/\min(\mathrm{MSE}_b(m)) : m \in \{\text{all methods}\}$) for different methods with survival probability estimand in the four scenarios. From left to right, top to bottom, the plots correspond to Scenarios 1–4, respectively. 'VT' denotes the virtual twin method with random survival forests; 'SRC1' denotes random survival forests using covariates (X, W); 'SRC2' denotes random survival forests using covariates (X, W, XW). 'IPCW' denotes a causal forest with Inverse Probability of Censoring Weighting; 'CSF' denotes a causal survival forest. Training size is (500, 1,000, 2,000, 5,000), the number of covariates 15, the size of the test set 2,000, and the number of repetitions 250.

References

Andrews D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, *61*(4), 821–856. http://doi.org/10.2307/2951764

Arlot S., & Genuer R. (2014). 'Analysis of purely random forests bias', arXiv, arXiv:1407.3939, preprint: not peer reviewed.

Athey S., & Imbens G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. http://doi.org/10.1073/pnas.1510489113

Athey S., Tibshirani J., & Wager S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178. http://doi.org/10.1214/18-AOS1709

Athey S., & Wager S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, *5*(2), 37–51. http://doi.org/10.1353/obs.2019.0001

Athey S., & Wager S. (2021). Policy learning with observational data. *Econometrica*, *89*(1), 133–161. http://doi.org/10.3982/ECTA15732

Beran R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, *5*(3), 445–463. https://doi.org/10.1214/aos/1176343842

Biau G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063–1095. doi:10.5555/2188385.2343682

Biau G., Devroye L., & Lugosi G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, *9*, 2015–2033. doi:10.5555/1390681.1442799

Breiman L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Breiman L., Friedman J., Stone C. J., & Olshen R. A. (1984). *Classification and regression trees*. CRC Press.

Buja A., Brown L., Kuchibhotla A. K., Berk R., George E., & Zhao L. (2019). Models as approximations II: A model-free theory of parametric regression. *Statistical Science*, *34*(4), 545–565. http://doi.org/10.1214/18-STS694

Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., Newey W., & Robins J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), 1–68. http://doi.org/10.1111/ectj.12097

Ciampi A., Thiffault J., Nakache J.-P., & Asselain B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, *4*(3), 185–204. http://doi.org/10.1016/0167-9473(86)90033-2

Cui Y. (2021). Individualized decision-making under partial identification: Three perspectives, two optimality results, and one paradox. *Harvard Data Science Review*, *3*(3), 1–19. https://doi.org/10.1162/99608f92.d07b8d16

Cui Y., Zhu R., & Kosorok M. (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic Journal of Statistics*, *11*(2), 3927–3953. http://doi.org/10.1214/17-EJS1305

Cui Y., Zhu R., Zhou M., & Kosorok M. (2022). Consistency of survival tree and forest models: Splitting bias and correction. *Statistica Sinica*, *32*, 1245–1267. doi:10.5705/ss.202020.0263

Fan C., Lu W., Song R., & Zhou Y. (2017). Concordance-assisted learning for estimating optimal individualized treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(5), 1565–1582. http://doi.org/10.1111/rssb.2017.79.issue-5

Fan Q., Hsu Y.-C., Lieli R. P., & Zhang Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, *40*(1), 313–327. http://doi.org/10.1080/07350015.2020.1811102

Fleming T. R., & Harrington D. P. (2011). *Counting processes and survival analysis(Vol. 169)*. John Wiley & Sons.

Foster D. J., & Syrgkanis V. (2019). 'Orthogonal statistical learning', arXiv, arXiv:1901.09036, preprint: not peer reviewed.

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

Friedberg R., Tibshirani J., Athey S., & Wager S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, *30*(2), 1–15. http://doi.org/10.1080/10618600.2020.1831930

Hahn P. R., Murray J. S., & Carvalho C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding and heterogeneous effects. *Bayesian Anal*, *15*(3), 965–1056. https://doi.org/10.1214/19-BA1195

Hammer S. M., Katzenstein D. A., Hughes M. D., Gundacker H., Schooley R. T., Haubrich R. H., Keith Henry W., Lederman M. M., Phair J. P., & Niu M., *et al.* (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081–1090. http://doi.org/10.1056/NEJM199610103351501

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

Hothorn T., Bühlmann P., Dudoit S., Molinaro A., & van der Laan M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355–373. http://doi.org/10.1093/biostatistics/kxj011

Hothorn T., Hornik K., & Zeileis A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. http://doi.org/10.1198/106186006X133933

Hothorn T., Lausen B., Benner A., & Radespiel-Tröger M. (2004). Bagging survival trees. *Statistics in Medicine*, *23*(1), 77–91. http://doi.org/10.1002/(ISSN)1097-0258

Imai K., & Li M. L. (2021). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, 1–15. doi:10.1080/01621459.2021.1923511

Imbens G. W., & Rubin D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ishwaran H., & Kogalur U. B. (2019). *Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 2.8.0. https://cran.r-project.org/package=randomForestSRC

Ishwaran H., Kogalur U. B., Blackstone E. H., & Lauer M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860. http://doi.org/10.1214/08-AOAS169

Kennedy E. H. (2020). 'Optimal doubly robust estimation of heterogeneous causal effects', arXiv, arXiv:2004.14497, preprint: not peer reviewed.

Künzel S. R., Sekhon J. S., Bickel P. J., & Yu B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. http://doi.org/10.1073/pnas.1804597116
Leblanc M., & Crowley J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, *88*(422), 457–467. http://doi.org/10.1080/01621459.1993.10476296
Lin Y., & Jeon Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, *101*(474), 578–590. http://doi.org/10.1198/016214505000001230
Lu M., Sadiq S., Feaster D. J., & Ishwaran H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, *27*(1), 209–219. http://doi.org/10.1080/10618600.2017.1356325
Lu W., Zhang H. H., & Zeng D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, *22*(5), 493–504. http://doi.org/10.1177/0962280211428383
Luedtke A. R., & van der Laan M. J. (2016a). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, *44*(2), 713. http://doi.org/10.1214/15-AOS1384
Luedtke A. R., & van der Laan M. J. (2016b). Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, *12*(1), 305–332. http://doi.org/10.1515/ijb-2015-0052
MacKinnon J. G., & White H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325. http://doi.org/10.1016/0304-4076(85)90158-7
Manski C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, *72*(4), 1221–1246. http://doi.org/10.1111/ecta.2004.72.issue-4
Meinshausen N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(34), 983–999. doi:10.5555/1248547.1248582
Murphy S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 331–355. http://doi.org/10.1111/rssb.2003.65.issue-2
Neugebauer R., & van der Laan M. (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, *137*(2), 419–434. http://doi.org/10.1016/j.jspi.2005.12.008
Nie X., & Wager S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. http://doi.org/10.1093/biomet/asaa076
Oprescu M., Syrgkanis V., & Wu Z. S. (2019). Orthogonal random forest for causal inference. In *International Conference on Machine Learning* (pp. 4932–4941).
Qian M., & Murphy S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, *39*(2), 1180. http://doi.org/10.1214/10-AOS864
R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Robins J. M., Rotnitzky A., & Zhao L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866. http://doi.org/10.1080/01621459.1994.10476818
Robinson P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, *56*(4), 931–954. http://doi.org/10.2307/1912705
Rosenbaum P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. http://doi.org/10.1093/biomet/70.1.41
Schick A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, *14*(3), 1139–1151. http://doi.org/10.1214/aos/1176350055
Segal M. R. (1988). Regression trees for censored data. *Biometrics*, *44*(1), 35–47. http://doi.org/10.2307/2531894
Semenova V., & Chernozhukov V. (2021). Estimation and inference about conditional average treatment effect and other structural functions. *The Econometrics Journal*, *24*(2), 264–289. http://doi.org/10.1093/ectj/utaa027
Sexton J., & Laake P. (2009, January). Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, *53*(3), 801–811. http://doi.org/10.1016/j.csda.2008.08.007
Steingrimsson J. A., Diao L., Molinaro A. M., & Strawderman R. L. (2016). Doubly robust survival trees. *Statistics in Medicine*, *35*(20), 3595–3612. http://doi.org/10.1002/sim.v35.20
Steingrimsson J. A., Diao L., & Strawderman R. L. (2019). Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, *114*(525), 370–383. http://doi.org/10.1080/01621459.2017.1407775
Sun Q., Zhu R., Wang T., & Zeng D. (2019, January). Counting process-based dimension reduction methods for censored outcomes. *Biometrika*, *106*(1), 181–196. http://doi.org/10.1093/biomet/asy064
Tian L., Alizadeh A. A., Gentles A. J., & Tibshirani R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532. http://doi.org/10.1080/01621459.2014.951443

Tibshirani J., Athey S., Friedberg R., Hadad V., Hirshberg D., Miner L., Sverdrup E., Wager S., & Wright M. (2022). *grf: Generalized random forests*. R package version 2.1.0. https://github.com/grf-labs/grf

Tsiatis A. (2007). *Semiparametric theory and missing data*. Springer Series in Statistics. Springer. https://books.google.com/books?id=xqZFi2EMB40C

Tsiatis A. A., Davidian M., Zhang M., & Lu X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, *27*(23), 4658–4677. http://doi.org/10.1002/sim.v27:23

van der Laan M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, *2*(1), 1008. http://doi.org/10.2202/1557-4679.1008

van der Laan M. J., & Robins J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

van der Laan M. J., & Rose S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media.

Vansteelandt S., & Dukes O. (2022). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society: Series B*. Forthcoming.

Wager S., & Athey S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. http://doi.org/10.1080/01621459.2017.1319839

Wager S., & Walther G. (2015). 'Adaptive concentration of regression trees, with application to random forests', arXiv, arXiv:1503.06388, preprint: not peer reviewed.

White H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. http://doi.org/10.2307/1912934

White H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. http://doi.org/10.2307/1912526

Yang J., Dahabreh I. J., & Steingrimsson J. A. (2021). Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics*.

Zeileis A. (2005). A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, *24*(4), 445–466. http://doi.org/10.1080/07474930500406053

Zeileis A., Hothorn T., & Hornik K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. http://doi.org/10.1198/106186008X319331

Zhang B., Tsiatis A. A., Laber E. B., & Davidian M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, *68*(4), 1010–1018. http://doi.org/10.1111/j.1541-0420.2012.01763.x

Zhang M., Tsiatis A. A., & Davidian M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, *64*(3), 707–715. http://doi.org/10.1111/biom.2008.64.issue-3

Zhao Y., Zeng D., John Rush A., & Kosorok M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, *107*(499), 1106–1118. http://doi.org/10.1080/01621459.2012.695674

Zhao Y.-Q., Zeng D., Laber E. B., Song R., Yuan M., & Kosorok M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, *102*(1), 151–168. http://doi.org/10.1093/biomet/asu050

Zhu R., & Kosorok M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, *107*(497), 331–340. http://doi.org/10.1080/01621459.2011.637468

Zhu R., Zhao Y.-Q., Chen G., Ma S., & Zhao H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, *73*(2), 391–400. http://doi.org/10.1111/biom.v73.2

Zimmert M., & Lechner M. (2019). 'Nonparametric estimation of causal heterogeneity under high-dimensional confounding', arXiv, arXiv:1908.08779, preprint: not peer reviewed.