Statistical Methods in Medical Research
OnlineFirst

https://doi.org/10.1177/09622802221111549

Article

Ensemble methods for survival function estimation with time-varying covariates

Weichi Yao [1], Halina Frydman [1], Denis Larocque [2], and Jeffrey S Simonoff [1]

Abstract

Survival data with time-varying covariates are common in practice. If relevant, they can improve on the estimation of a survival function. However, the traditional survival forests—conditional inference forest, relative risk forest and random survival forest—have accommodated only time-invariant covariates. We generalize the conditional inference and relative risk forests to allow time-varying covariates. We also propose a general framework for estimation of a survival function in the presence of time-varying covariates. We compare their performance with that of the Cox model and transformation forest, adapted here to accommodate time-varying covariates, through a comprehensive simulation study in which the Kaplan-Meier estimate serves as a benchmark, and performance is compared using the integrated L_2 difference between the true and estimated survival functions. In general, the performance of the two proposed forests substantially improves over the Kaplan-Meier estimate. Taking into account all other factors, under the proportional hazard setting, the best method is always one of the two proposed forests, while under the non-proportional hazard setting, it is the adapted transformation forest. K-fold cross-validation is used as an effective tool to choose between the methods in practice.

Keywords

Survival forests, time-varying covariates, survival curve estimate, dynamic estimation, left-truncated right-censored survival data

[1] New York University, New York, NY, USA

[2]HEC Montréal, Montréal, Québec, CA

Corresponding author(s):
Weichi Yao, New York University, New York, NY 10012, USA. Email: wyao@stern.nyu.edu

1 Introduction

Methodology for survival data often assumes that covariate information is time-invariant; that is, only values measured at time 0 are used. In this situation, survival analysis models can provide an estimate of the survival function (the probability of surviving past time t) for a subgroup of the population (i.e. a subpopulation) with a specific set of values for the covariates. Time-varying covariates, however, are common in practice and play an important role in the analysis of censored time-to-event data. For example, in a study of the effect of heart transplant on survival for heart patients, the occurrence of a transplant can be modeled as a time-varying binary covariate,[1] and in a study of the effect of $\mathrm{CD}\,4+$ T-cell counts on the occurrence of AIDS or death for HIV-infected patients, the cell count was used as a time-varying numerical covariate, measured longitudinally.[2] The Cox proportional hazards model[3] has a long history of being used to model and analyze censored survival data. As a semi-parametric model, it assumes that the time-invariant covariates have a proportional effect on the hazard function. The Cox model was extended to fit time-varying covariates using a counting process formulation as follows.[4] Consider continuous-time survival data with time-varying covariates, where each subject may have multiple records of measurements of risk factors at multiple time points. In practice, as the subjects are observed intermittently, the time-varying covariates are assumed constant between observation times. One can then reformat the data structure using the counting process approach by which a data record of a subject becomes a list of pseudo-subjects, that are treated as being independent, left-truncated and right-censored observations.

It is important to recognize that survival regression models, like all regression models, can be used for the two distinct purposes of estimation and prediction. These two purposes, while related (prediction almost always involves estimation of some kind as a first step), are distinctly different from each other. Prediction is meaningful at the level of an individual. In the context of survival data, this would correspond to an estimated survival function for a particular individual. For such an individual, if a change in the value of a covariate at time t^* can potentially impact the future probabilities of survival of that individual, then the estimated survival function for

that individual must be 1 for any time $t < t^*$ (since the covariate couldn't impact the future probabilities of survival unless the individual was alive at that time). A related point is the existence of so-called internal covariates, in which a variable (e.g. blood pressure) can only be measured when an individual is alive, meaning that the act of measuring the covariate implies that the survival function value for that individual must equal 1. These facts, and their implications for prediction, are well-known, and are the reason for the prominence of joint modeling methods[5] in the prediction of survival data with time-varying covariates.

The focus of this paper, however, is exclusively on estimation, rather than prediction. The goal here is not prediction of an individual's survival function, but rather estimation of a population-level survival function. This is an important problem in any situation that involves strategic decision making, such as public health policy or business operations and planning. Civic policymakers are interested in developing strategies that change over time, in response to changing conditions; this does not involve prediction at the individual level, but rather estimation of population probabilities. Such estimation can answer "what if" questions, in order to either be prepared for what might happen, or to try to control what will happen. Recognizing that various aspects of society can change, it can also examine what the effects of such changes might be. Examples of this could include modeling the incidence of COVID-19 infection in the population as vaccination and testing rates change, or estimating what proportion of deliveries will be made by Christmas as various characteristics of the supply chain change. These problems are fundamentally different from the problem of making a prediction for a particular individual, and issues related to joint modeling and the distinction between internal and external covariates are no longer relevant.

When time-varying covariates are available, it is important to use the updated covariate values to dynamically adjust the estimated survival function. We use a simple hypothetical example to fix ideas and illustrate the mechanics of the survival function adjustment. Consider the following simple example, based on the COVID-19 problem noted earlier. The time-to-event is the time to a positive COVID test, and there is a time-varying covariate $\boldsymbol{X(t)}$ that describes the vaccination status of a subject, assuming a two-shot vaccine (unvaccinated, vaccinated with one dose, vaccinated with two doses, vaccinated with two doses and a booster). Starting with a sample of unvaccinated and COVID-free subjects, the changes of $\boldsymbol{X(t)}$ define distinct subgroups of the sample (and hence the population). Estimation of the population

survival function for subjects who have gotten one dose of the vaccine would be based only on those subjects who survived to get that dose, but the overall population survival function estimate would have to be adjusted to account for the probability that unvaccinated subjects did not survive to that time. This same adjustment would be applied at the other times that vaccination status changes, resulting in a survival function estimate that is appropriate for all members of the population who followed the specified temporal vaccination pattern. In the next section we describe how such a survival function estimate can be constructed, and in Appendix A we illustrate the detailed calculations of the dynamically adjusted survival function for this simple example.

The Cox proportional hazards model with time-varying covariates, hereafter referred to as the extended Cox model, relies on restrictive assumptions such as proportional hazards and a log-linear relationship between the hazard function and covariates. Tree-based methods and their ensembles, which are useful non-parametric alternatives to the extended Cox model, also can incorporate time-varying covariates. Recently, two types of survival trees were proposed as extensions of the relative risk tree[6] and of the conditional inference tree,[7] respectively, to left-truncated right-censored (LTRC) data, referred to as LTRC trees.[8] The proposed LTRC tree algorithms allow for time-varying covariate data after the data structure is reformatted using the counting process approach. Another tree-based method that can handle LTRC survival data and therefore potentially be applied to time-varying covariate data is the novel "transformation tree," and the corresponding ensemble is the "transformation forest."[9] These two algorithms are based on a parametric family of distributions characterized by their transformation function and developed to detect distributional alternatives to proportional hazards (PH). None of the above methods have considered the estimation of the survival function. Similarly, recently developed methods for hazard function estimation in the presence of time-varying covariates haven't dealt with survival function estimation in general.[10,11] There exist other survival trees and forest methods that can handle time-varying covariate data, but only for discrete-time survival data.[12–16] In this paper, we focus on forest algorithms for dynamic estimation of the survival function for continuous-time survival data. Ensemble methods like forest algorithms are known to preserve low bias while reducing variance and therefore can substantially improve prediction accuracy, compared to tree algorithms.[17] The most well-known ensemble methods for survival analysis are perhaps the relative risk forest,[18] random survival forest[19] and

conditional inference forest.[20] These forest methods provide estimates of survival functions, but only for right-censored survival data with time-invariant covariates. We propose to extend the relative risk and conditional inference forests, as well as the transformation forest, to allow time-varying (TV) covariates. We refer to them as RRF-TV, CIF-TV, and TSF-TV, respectively.

The proposed methods by design can handle survival data with all combinations of left-truncation and right-censoring in the survival outcome, and with both time-invariant and time-varying covariates. In this paper, we focus on survival data with time-varying covariates. Similar analysis for LTRC data with time-invariant covariates is provided in Section S2 in the Supplemental Material.

2 Proposed forests for time-varying covariate data

Assume p covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ are available, some of which are TV covariates and the others are time-invariant (TI). For example, assume X_1 is the only time-invariant covariate among all p covariates, then at time t, $\boldsymbol{X}(t) = (X_1, X_2(t), \ldots, X_p(t))$. For ease of exposition, we write $\boldsymbol{X}(t) = (X_1(t), X_2(t), \ldots, X_p(t))$ with $X_1(t) \equiv X_1$ for all t. Observations are obtained from N subjects. Note that the subjects are observed only intermittently, for example, $J^{(i)}$ times for subject i, initially observed at $t_0^{(i)}$, and then at $t_j^{(i)}$, $j = 1, 2, \ldots, J^{(i)} - 1$, for follow-up visits, with corresponding observed values $\boldsymbol{x}_j^{(i)} = (x_{j,1}^{(i)}, \ldots, x_{j,p}^{(i)})$. Let $t_{J^{(i)}}^{(i)}$ denote $\tilde{T}^{(i)} = \min(T^{(i)}, C^{(i)})$, the minimum value of the true survival time $T^{(i)}$ and censoring time $C^{(i)}$. We assume non-informative censoring conditional on the covariates. Denote $\Delta = \mathbb{I}\{T^{(i)} \leq C^{(i)}\}$, which indicates whether a subject experienced an event ($\Delta = 1$) or was right-censored ($\Delta = 0$). If $t_0^{(i)} \neq 0$, then we say the survival time is left-truncated. The outcome of interest is the time to the event.

At any time t, let $\mathscr{I}(t)$ denote an arbitrary set of time values up to time t, that is, $\mathscr{I}(t) \subseteq [0, t]$. It could be a finite number of time points, a finite number of intervals, or a disjoint set of time intervals and/or time points. Given the historical data for N subjects observed up to the death or censoring time, the goal is to estimate the conditional survival function $S(t \mid \boldsymbol{X}(u) = \boldsymbol{x}(u), u \in \mathscr{I}(t))$ where $\{\boldsymbol{x}(u), u \in \mathscr{I}(t)\}$ is a particular stream of covariate values. This true survival function is defined as the population proportion of subjects who have the specified covariate values at the specified times up to either time t or their event time

(whichever comes first) that are alive at time t. Note that the population includes all subjects for whom the specified conditions hold up until their time of event if that occurs before the evaluation time t, even if the conditions do not hold after the time of event. This is true if measurement of the covariate is no longer meaningful after the event occurs (as might be the case for a so-called internal covariate, such as blood pressure), or it is meaningful and available but no longer satisfies the conditions (a so-called external covariate, such as pollutant level). The reason is that the influence of the covariate on survival in either case is irrelevant for a subject for whom the event has occurred.

The proposed forest methods provide the survival function estimate by following three steps. First, we adopt the counting process approach to reformat the data structure. This approach assumes that the time-varying covariates are constant between the observed time points, that is,

$$\boldsymbol{X}^{(i)}(t) = \boldsymbol{x}_j^{(i)}, \quad t \in \left[t_j^{(i)}, t_{j+1}^{(i)}\right), \quad j = 0, 1, \ldots, J^{(i)} - 1$$

It then splits the i-th subject observation into $J^{(i)}$ pseudo-subject observations: $(t_j^{(i)}, t_{j+1}^{(i)}, \delta_j^{(i)}, \boldsymbol{x}_j^{(i)})$ with LTRC times $t_j^{(i)}$, $t_{j+1}^{(i)}$, and event indicator $\delta_j^{(i)} = \mathbb{I}\{j = J^{(i)} - 1\}$, $j = 0, 1, \ldots, J^{(i)} - 1$. The multiple records from N subjects now become a list of pseudo-subjects,

$$\left\{\left\{\left(t_j^{(i)}, t_{j+1}^{(i)}, \delta_j^{(i)}, \boldsymbol{x}_j^{(i)}\right)\right\}_{j=0}^{J^{(i)}-1}\right\}_{i=1}^{N}$$

The set of pseudo-subjects is treated as if they were independent in the following form

$$\left\{\left(L_l', R_l', \delta_l', \boldsymbol{x}_l'\right)\right\}_{l=1}^{n}, \quad n = \sum_{i=1}^{N} J^{(i)} \tag{1}$$

where $\boldsymbol{x}_l' = (x_{l,1}', \ldots, x_{l,p}')$ is the vector of the observed values of p covariates from the l-th pseudo-subjects in the reformatted dataset. The second step is to apply the forest algorithms on the reformatted dataset given in (1), to fit a model. Finally, in the

third step, given a particular stream of covariate values $\boldsymbol{x}_j^*$ at the corresponding time values t_j^*, $j = 0, 1, \ldots$, a survival function estimate is constructed based on the outputs of the proposed forest algorithms. More specifically, at any time t, with $\mathscr{X}^*(t)$ denoting the covariate information up to time t,

$$\mathscr{X}^*(t) = \left\{ \boldsymbol{x}_j^*, \forall j : 0 \leq t_j^* \leq t \right\} \tag{2}$$

we compute the estimated survival probability $\widehat{S}(t \mid \mathscr{X}^*(t))$.

2.1 Extending right-censored TI survival forests to the proposed TV forests

The conditional inference forest and the relative risk forest are both tree-based ensemble methods, where B individual trees are grown from B bootstrap samples drawn from the original data. Randomness is induced into each node of each individual tree when selecting a variable to split on. Only a random subset I of the total p covariates is considered for splitting at each node. The node is then split using the candidate covariates based on different criteria for different forest methods. To extend the two forest methods for right-censored survival data with time-invariant covariates to the forests for (left-truncated) right-censored survival data with time-varying covariates, the splitting criteria are modified.

2.1.1 Recursive partitioning in the proposed CIF-TV forest. Consider right-censored survival time data of the form $(\tilde{T}, \Delta, \boldsymbol{X})$, with survival/censored time $\tilde{T}$, event indicator Δ ($\tilde{T}$ denotes the survival time if $\Delta = 1$, or censored time if $\Delta = 0$), and p time-invariant covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$. In each node, the recursive partitioning in the conditional inference forest algorithm is based on a test of the global null hypothesis of independence between the response variable in the right censored case $\boldsymbol{V} = (\tilde{T}, \Delta)$ and any of the covariates in the random subset I. It is formulated in terms of $|I|$ partial hypotheses, $H_0 = \cap_{k=1}^{|I|} H_0^k$ with

$$H_0^k : D(\boldsymbol{V} \mid X_k) = D(\boldsymbol{V}), \quad k = 1, \ldots, |I| \tag{3}$$

where $D(\boldsymbol{V} \mid \boldsymbol{X_k})$ denotes the conditional distribution of $\boldsymbol{V}$ given the covariate $\boldsymbol{X_k}$. The independence is measured by linear statistics incorporating the log-rank scores that take censoring into account. In the extension of conditional inference tree to LTRC conditional inference tree, the log-rank score can be modified as follows for LTRC data.[8] Given the list of pseudo-subject observations with LTRC survival times as in (1), the response variable now becomes $\boldsymbol{V} = (L_l', R_l', \delta_l')$ in the test of partial null hypothesis of independence (3) for the $\boldsymbol{l}$-th observation $(L_l', R_l', \delta_l', \boldsymbol{x}_l')$. The corresponding log-rank score is defined as

$$U_l = \begin{cases} 1 + \log \widehat{S}(R_l') - \log \widehat{S}(L_l'), & \text{if } \delta_l' = 1 \\ \log \widehat{S}(R_l') - \log \widehat{S}(L_l'), & \text{otherwise} \end{cases} \tag{4}$$

Note that $\widehat{S}$ is the nonparametric maximum likelihood estimator (NPMLE) of the survival function which takes into account left-truncation. Such an estimator can be constructed using the product-limit estimator, that is, Kaplan-Meier estimator with pseudo-subjects that fall into the current node.[21,22] We similarly use the log-rank score U_l in the proposed extension of conditional inference forest to LTRC conditional inference forest.

2.1.2 Recursive partitioning in the proposed RRF-TV forest. The relative risk forest combines the use of relative risk trees[6] with random forest methodology[17] as a way to reliably estimate relative risk values. The Classification and Regression Tree (CART) paradigm[23] is used to produce a relative risk forest by exploiting an equivalence with Poisson tree likelihoods.

The splitting criterion under the relative risk framework is to maximize the reduction in the one-step deviance between the log-likelihood of the saturated model and the maximized log-likelihood. At node $\boldsymbol{h}$, let $\mathscr{R}_h$ denote the set of labels of those observations that fall into the region corresponding to node $\boldsymbol{h}$, and let $\lambda_h(t)$ and Λ_h denote the corresponding hazard and cumulative hazard function, respectively. Under the assumption of PHs,

$$\lambda(t) = \lambda_0(t)\varphi_h$$

where λ_0 is the baseline hazard and φ_h is the nonnegative relative risk of the node h. Given the right-censored observations $(\tilde{t}_l, \delta_l)$, $l \in \mathscr{R}_h$, the maximum likelihood estimate of φ_h is

$$\widehat{\varphi}_h = \frac{\sum_{l \in \mathscr{R}_h} \delta_l}{\sum_{l \in \mathscr{R}_h} \Lambda_0(\tilde{t}_l)}$$

where the Nelson-Aalen estimator using all of the data at the root node $\widehat{\Lambda}_0$ is used for Λ_0.[6] The full likelihood deviance residual for node h is defined as

$$d_h = \sum_{l \in \mathscr{R}_h} 2\left[\delta_l \log\left(\frac{\delta_l}{\widehat{\Lambda}_0(\tilde{t}_l)\widehat{\varphi}_h}\right) - \left(\delta_l - \widehat{\Lambda}_0(\tilde{t}_l)\widehat{\varphi}_h\right)\right] \tag{5}$$

For a Poisson regression model, let ϱ_h denote the event rate, s_l and c_l be the exposure time and the event count for observation l, respectively, then (5) is equivalent in form to the deviance residual based on the Poisson regression model,

$$d_h^{\text{Pois}} = \sum_{l \in \mathscr{R}_h} 2\left[c_l \log\left(\frac{c_l}{s_l \hat{\varrho}_h}\right) - \left(c_l - s_l \hat{\varrho}_h\right)\right] \tag{6}$$

with $\hat{\varrho}_h = \frac{\sum_{l \in \mathscr{R}_h} c_l}{\sum_{l \in \mathscr{R}_h} s_l}$, by replacing $\hat{\varrho}_h$ with $\widehat{\varphi}_h$, s_l with $\widehat{\Lambda}_0(\tilde{t}_l)$, and c_l with δ_l.[6] To adapt the Poisson regression tree approach for left-truncated right-censored survival observations $\{(L_l', R_l', \delta_l')\}$, the key is to modify the estimated $\widehat{\Lambda}_0(\tilde{t}_l)$ and δ_l to replace s_l, c_l and $\hat{\varrho}_h$ in (6). First, compute the estimated cumulative hazard function $\widehat{\Lambda}_0(\cdot)$ based on all (pseudo-subject) observations. The exposure time s_l and the event count c_l for observation l in (6) are then replaced by $\widehat{\Lambda}_0(R_l') - \widehat{\Lambda}_0(L_l')$ and δ_l' to obtain the deviance residual appropriate for LTRC data.[8]

2.1.3 Implementation of the proposed forests. To implement the CIF-TV and RRF-TV algorithms, we make use of the fast algorithms provided in the packages partykit[24] and randomForestSRC,[25] respectively. The RRF-TV building architecture is based on employing the fast C code from randomForestSRC. The Poisson splitting

rule[6] is coded in C and is incorporated by exploiting the custom splitting rule feature in the *rfsrc* function. The CIF-TV is built by extending the survival forest algorithms coded in the *cforest* function from partykit with the log-rank score adapted for LTRC data.

2.2 Bootstrapping subjects versus bootstrapping pseudo-subjects

In forest-like algorithms, bootstrapped samples are typically used to construct each individual tree to increase independence between these base learners. The nonparametric bootstrap approach is used in all three types of forests being considered here (CIF-TV, RRF-TV, and TSF-TV). It places positive integer weights that sum to the sample size on approximately 63% of the observations in any given bootstrap sample, and the 37% of the data excluded during this procedure is called out-of-bag data (OOB data). As we split each subject into several pseudo-subjects and treat these pseudo-subjects as independent observations on which to build the forests, we have two bootstrapping options: we can bootstrap subjects or bootstrap pseudo-subjects.

Bootstrapping pseudo-subjects is used for some discrete survival forest methods.[12,13] Since all pseudo-subjects are treated as independent observations in the recursive partitioning process,[26,8] bootstrapping pseudo-subjects is just bootstrapping "independent" observations as the first step of any forest algorithm. On the other hand, bootstrapping subjects is a natural approach, as it keeps all of the pseudo-subjects for each subject in the bootstrap sample. In fact, simulations have shown that the two different bootstrapping mechanisms do not result in fundamentally different levels of performance; see Section S1.5 in the Supplemental Material for more details. This paper will focus on forests based on bootstrapping subjects.

2.3 Regulating the construction of individual trees in the proposed forests

In a forest algorithm, only a random subset of covariates is considered for splitting at each node. The size of this random set is denoted by *mtry*. In addition to *mtry*, many other parameters play an important role in establishing a split in the individual tree. In both the *cforest*[24] function for the conditional inference forest and the *traforest*[27] function for the transformation forest algorithms, *minsplit* (the minimum sum of weights in a node in order to be considered for splitting), *minprob* (the minimum proportion of observations needed to establish a terminal node), and *minbucket* (the minimum sum of weights in a terminal node) control whether or not to implement a split; in the *rfsrc*[28] function for the relative risk forest algorithms, *nodesize* controls

the average terminal node size. These tuning parameters thereby regulate the size of the individual trees. The recommended values for these parameters are usually given as defaults to the algorithm. For example, *mtry* is usually set to be $\sqrt{p}$, where p is the total number of covariates,[20,19] *nodesize* to be 15 in the *rfsrc* function, (*minsplit*, *minbucket*) to be $(20, 7)$ in the *cforest* function and the *traforest* function, which we refer to as the *default parameter settings*.

The best values for these parameters would be expected to depend on the problem and they should be treated as tuning parameters.[29] It has been shown for conditional inference forests for interval-censored data[30] that these parameters have a non-negligible effect on the overall performance of the forest algorithm. As we extend the forest framework to allow for left truncation, and from time-invariant covariate data to time-varying covariate data, we should also consider rules for choosing tuning parameters.

The algorithm designed in survival forests for interval-censored data with time-invariant covariates[30] tunes the value of *mtry* on the "out-of-bag observations." To adapt the same idea to survival forests based on bootstrapping subjects on a dataset with time-varying covariates, we define the "out-of-bag observations" for the b-th tree to be the observations from those subjects that are left out of the b-th bootstrap sample and not used in the construction of the b-th tree. The survival curve can be estimated by using each of the B trees in which that subject was "out-of-bag," denoted as $\widehat{S}^{\mathrm{OOB}}$. To evaluate the fit of the out-of-bag estimate $\widehat{S}^{\mathrm{OOB}}$ with a specific value of *mtry*, we compute the estimation error defined as the integrated Brier score designed for time-invariant covariate data,[31] adapted here for time-varying covariate data as follows.

For a given dataset $\mathscr{D}$, define the integrated Brier score $\widehat{\mathrm{IBS}}(\widehat{S}; \mathscr{D})$ for the estimated survival function $\widehat{S}$ as

$$\widehat{\mathrm{IBS}}(\widehat{S}; \mathscr{D}) = \frac{1}{|\mathscr{D}|} \sum_{i \in \mathscr{D}} \frac{1}{\tau^{(i)}} \int_0^{\tau^{(i)}} \widehat{W}^{(i)}(t) \left[\tilde{Y}^{(i)}(t) - \widehat{S}\left(t \mid \mathscr{X}^{(i)}(t)\right) \right]^2 dt \tag{7}$$

where $\tau^{(i)}$ determines the length of difference evaluation time span for subject i, $\tilde{Y}^{(i)}(t) = \mathbb{I}\{\tilde{T}^{(i)} > t\}$ is the observed status ($\tilde{T}^{(i)}$ is the survival/censored time), $\widehat{W}^{(i)}(t)$ is the inverse probability of censoring weights,

$$\widehat{W}^{(i)}(t) = \frac{\left(1 - \tilde{Y}^{(i)}(t)\right)\Delta^{(i)}}{\widehat{G}\left(\tilde{T}^{(i)}\right)} + \frac{\tilde{Y}^{(i)}(t)}{\widehat{G}(t)}$$

with $\widehat{G}$ the Kaplan-Meier estimate of the censoring distribution based on $\{(\tilde{T}^{(i)}, 1 - \Delta^{(i)})\}_{i \in \mathscr{D}}$.[31] The corresponding Brier score $\widehat{\mathrm{BS}}(t, \widehat{S}; \mathscr{D})$ at time t is defined as

$$\widehat{\mathrm{BS}}(t; \widehat{S}; \mathscr{D}) = \frac{1}{|\mathscr{D}|} \sum_{\mathrm{i} \in \mathscr{D}} \widehat{\mathrm{W}}^{(\mathrm{i})}(\mathrm{t}) \left[\tilde{\mathrm{Y}}^{(\mathrm{i})}(\mathrm{t}) - \widehat{\mathrm{S}}\left(\mathrm{t} \mid \mathscr{X}^{(\mathrm{i})}(\mathrm{t})\right)\right]^2 \tag{8}$$

The resulting estimation error for the ensemble method with a specific value of *mtry* can then be computed by setting $\widehat{S} = \widehat{S}^{\mathrm{OOB}}$ in (7). An appropriate value of *mtry* is the one that minimizes the "out-of-bag" estimation error.

Regarding the values of other tuning parameters, the optimal values that determine the split vary from case to case. As fixed numbers, the default values may not affect the splitting at all when the sample size is large, while having a noticeable effect in smaller data sets. This inconsistency can potentially result in good performance in some data sets and poor performance in others. In the simulations, we set *minsplit*, *minbucket*, and *nodesize* to be the maximum of the default value and the square root of the number of pseudo-subject observations n. This set of values can automatically adjust to the change in size of the data set. We refer to the above choice of *mtry*, *minsplit*, *minbucket*, and *nodesize* as the *proposed parameter settings*, as opposed to the default settings.

2.4 Constructing a survival function estimate for time-varying covariate data

Consider a particular stream of covariate values $\boldsymbol{x}_j^*$ at time t_j^*, for $j = 0, 1, \ldots, J-1$. Denote $\mathscr{X}^*(u) = \{\boldsymbol{x}_j^* : \forall j,\ t_j^* \leq u\}$ the set of the covariate values up to time u. At time $t \in [t_j^*, t_{j+1}^*)$, we derive a recursive computation as follows

(9)

$$\widehat{S}\big(t \mid \mathscr{X}^*(t)\big) = \begin{cases} 1, & t = t_0^* \\ \frac{\widehat{S}_{A,j}(t)}{\widehat{S}_{A,j}(t_j^*)} \widehat{S}\big(t_j^* \mid \mathscr{X}^*(t)\big), & t \in [t_j^*, t_{j+1}^*) \end{cases}$$

where $\widehat{S}_{A,j}(t) \triangleq \widehat{\mathbb{P}}(T > t \mid \boldsymbol{x}_j^*)$ denotes the output of the algorithm for the input with time-invariant covariate value $\boldsymbol{x}_j^*$. See Appendix B for details of derivation.

Further, by expansion, (9) is equivalent to

$$\widehat{S}\big(t \mid \mathscr{X}^*(t)\big) = \frac{\widehat{S}_{A,j}(t)}{\widehat{S}_{A,j}(t_j^*)} \prod_{l=0}^{j-1} \frac{\widehat{S}_{A,l}(t_{l+1}^*)}{\widehat{S}_{A,l}(t_l^*)} \tag{10}$$

for $t \in [t_j^*, t_{j+1}^*)$. The formulation in (10) provides another perspective to view the resulting survival curve estimate—it is constructed by combining the pseudo-subject-specific ensemble estimates of the survival function with multiplicative correction factors. These correction factors ensure monotonicity of the overall curve.

Note that the construction in (10) coincides with what the function *survfit* in the R package survival[32] uses to give a subject's survival function estimate from a *coxph* fit using the same counting process approach.

Figure 1 gives an illustration of the estimated survival functions with or without "updating" the covariate values at time $t_1^*, t_2^*, \ldots, t_5^*$. For each j, the "update" in the estimated survival probability starting from time t_j^* for all the future time $t > t_j^*$ reflects the difference in the estimated surviving proportions of two subpopulations of subjects with their covariate trajectories diverging from the shared past before t_j^*. One can see that the change in covariate information at each time point of change can make a huge impact on the future path of the estimated survival function.

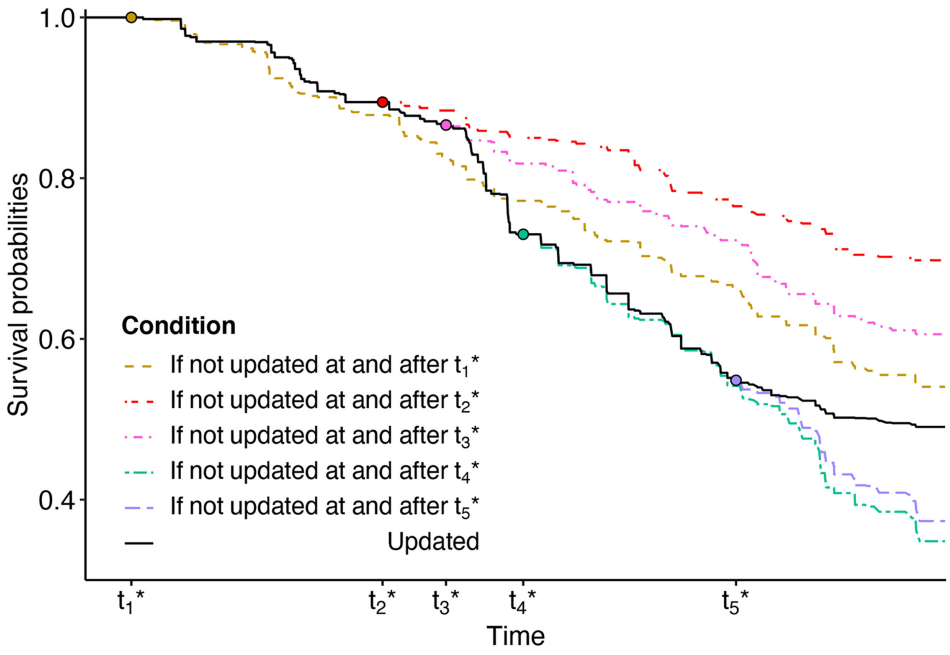

Figure 1. Illustration of estimated survival functions with or without changing the covariate values at t_j^*, $j = 1, 2, \ldots, 5$. At t_j^*, the dot on the curve shows the estimated survival function at the time of change (having been updated at all of the previous time points $t_1^*, \ldots, t_{j-1}^*$). If not updated with the latest change, the estimated survival function at $t > t_j^*$ is shown as the dashed line with the same color as the dot. The solid black line shows the one estimated by CIF-TV and constructed as given in (9). It tracks all of the changes in covariate values and updates the estimated survival probabilities at each time step of change.

3 Simulation study

3.1 Data generation scheme

In the simulation study, observations from N subjects are generated independently with p covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$. We set $p = 20$. Eight of these covariates are time-invariant: $X_1, X_{11} \sim \text{Bern}(0.5)$, $X_2, X_7, X_{10} \sim \text{Unif}(0, 1)$, $X_8 \sim \text{Unif}(1, 2)$, X_9 follows a categorical distribution with possible outcomes $\{1, 2, 3, 4, 5\}$ with equal probability, X_{12} follows a categorical distribution with possible outcomes $\{0, 1, 2\}$ with equal probability. The others are time-varying, whose values are obtained at m randomly generated time points, different for each subject. In the simulations, we set $m = 11$. At each of these preset time points, for some time-varying covariates, the value is randomly resimulated from its distribution: $X_3, X_{19} \sim \text{Bern}(0.5)$, $X_4, X_{15}, X_{17} \sim \text{Unif}(0, 1)$, X_5 and X_{14} both follow a categorical distribution with possible outcomes $\{1, 2, 3, 4, 5\}$ with equal probability; for other time-varying covariates, the value is resimulated following particular

patterns: X_6, whose initial value is randomly generated from $\{0, 1, 2\}$ with equal probability, which will choose to stay at the original value or move one level up but the largest value can only be 2; the changing pattern of X_{13} is always $0 \to 1$; the changing pattern of X_{16} is either $0 \to 1$ or $1 \to 2$; the changing pattern of X_{18} is $0 \to 1 \to 2$; value of X_{20} is a linear function of the left-truncated time point of the interval with slope and intercept follows $\text{Unif}(0, 1)$. Further details of the changing pattern of X_6, X_{13}, X_{16}, X_{18} and X_{20} can be found in Section S1.2 in the Supplemental Material.

After the time-varying covariates' values are generated at each of those m preset time points, the true survival time T is then computed under different model setups and the right-censoring time C is generated independently.

3.2 Model setup

We consider the following factors for different variations of data generating models:

a. Different proportions of time-varying covariates in the true model (scenario).
b. Different signal-to-noise ratios (SNRs) labeled as "High" and "Low," constructed by choosing different coefficients in the true model.
c. Different hazard function settings: a PH and a non-PH (non-PH) setting.
d. Different survival relationships between the hazards and covariates: a linear, a nonlinear, or an interaction model.
e. Different censoring rates: 20% and 50%.
f. Different sample sizes: $N = 100$, 300, and 500.
g. Different amount of knowledge of history of changes in covariates' values: Case I —When all changes in values of covariates are known, labeled as "Full," and Case II—When only half of the changes in values of the covariates are known, labeled as "Half."

Scenario. We consider two different proportions of time-varying covariates in the true model: 2TI + 1TV, and 2TI + 4TV; see Table 1. Only the first six covariates are given in the table since X_7 to X_{20} are never involved in the true DGP.

Scenario	Time-invariant		Time-varying			
	X_1	X_2	X_3	X_4	X_5	X_6
2TI + 1TV	✓	✓			✓	
2TI + 4TV	✓	✓	✓	✓	✓	✓

Survival relationships. Given a survival relationship, the survival time T depends on $\vartheta(t) = \vartheta(\boldsymbol{X}(t))$. Here we use scenario 2TI + 4TV for illustration.

For a linear survival relationship, $\vartheta(t) = \beta_0 + \sum_{k=1}^{6} \beta_k X_k(t)$ with constants β_k, $k = 0, \ldots, 6$. For a nonlinear survival relationship, $\vartheta(t) = \phi_1 \cos\left(\sum_{k=1}^{6} X_k(t)\right) + \phi_2 \log\left(\psi_0 + \sum_{k=1}^{6} \psi_k X_k(t)\right) + \phi_3 X_1(t)(2X_2(t))^{4X_4(t)}$ with some constants ϕ_1, ϕ_2, ϕ_3, and ψ_k, $k = 0, \ldots, 6$. For an interaction model, ϑ is determined by the value of time-varying covariate X_4 and the value of time-varying covariate X_5. Figure 2 gives an example of the structure of the covariates driving the interaction survival relationship, where $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4$ correspond to

(i) $\vartheta(t) = \eta_1[X_1(t)X_2(t) - \log(X_3(t) + X_4(t)) - X_6(t)/X_5(t)] + \eta_2$

(ii) $\vartheta(t) = \gamma_0 + \sum_{k=1}^{6} \gamma_k X_k(t)$

(iii) $\vartheta(t) = \eta_3[\cos(\pi(X_1(t) + X_5(t))) + \sqrt{X_2(t) + X_6(t)} - X_3(t)] + \eta_4$

(iv) $\vartheta(t) = \alpha_0 + \sum_{k=1}^{6} \alpha_k X_k(t)$

with some constants $\{\alpha_k\}_{k=0}^{6}$, $\{\gamma_k\}_{k=0}^{6}$ and $\{\eta_k\}_{k=1}^{4}$.

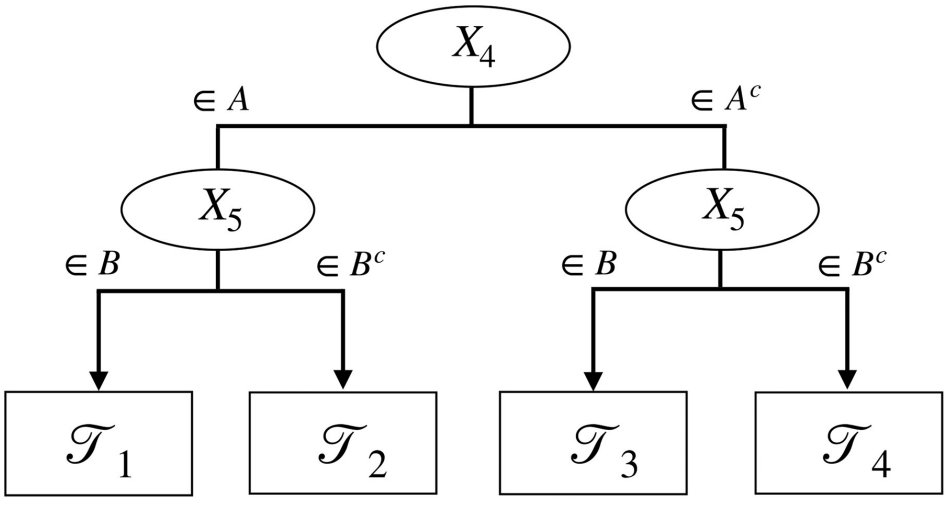

Figure 2. An example of the structure of the covariates driving the interaction survival relationship, where set A and B are some sets of values of X_4 and X_5, respectively.

Survival distributions under the PH and the non-PH setting. Given a survival relationship model, the survival time T depends on ϑ via a Weibull distribution.

For PH models, a closed-form solution can be derived to generate survival times with time-varying covariates for the Weibull distribution.[33] For non-PH models, a closed-form solution exists for the Weibull distribution, with its nonconstant shape term a function of the covariates (note that the PH relationship is on the scale parameter for the Weibull distribution).

To be more specific, for the PH setting, we consider the underlying hazard function

$$h(t) = h_0(t)\exp(\vartheta(t)) \tag{11}$$

where the baseline hazard function is given by $h_0(t) = \lambda\nu t^{\nu-1}$ with $\lambda > 0$ and $\nu > 0$. For the non-PH setting, the hazard function is set to be

$$h(t) = \lambda\exp(\vartheta(t))(\lambda t)^{\exp(\vartheta(t))-1} \tag{12}$$

where $\lambda > 0$. Values of $\vartheta(t)$ have been scaled to be between -3 and 3. Note that, compared with the Weibull distribution under the PH setting, now the time-varying effects appear in the shape term instead of the scale term. The survival function is then given by $S(t) = \exp(-\int_0^t h_0(s)\exp(\vartheta(s))ds)$. Further details of simulating the survival time T can be found in Section S1.1 in the Supplemental Material.

Histograms of survival times for typical samples with the number of subjects $N = 500$ in each scenario are provided in Section S1.3 in the Supplemental Material to illustrate the data generating processes. The parameters set in the simulation study can be found in Section S1.4 in the Supplemental Material.

Knowledge of history of changes in covariates' values. In practice, it is likely that not all of the changes in the covariates' values are known to the data analyst. For example, suppose that a patient's blood pressure is to be measured at regularly scheduled examination times. If a patient obeys the schedule then, from the doctor's point of view, all changes in blood pressure are known. However, if a patient skips some scheduled examination times, then not all changes in the blood pressure are known. In the latter case, this means that whatever modeling method is used to

estimate the survival curves, it is operating with incorrect values as inputs and therefore its performance would be expected to deteriorate. Of course, the fact that blood pressure is actually changing continuously is an extreme example of this phenomenon; in these simulations we limit ourselves to changes at a finite number of time points. The simulations are designed to investigate the performance of different modeling methods in this situation under the following two circumstances:

- Case I. When all changes in covariate values are known;
- Case II. When only half of the changes in covariate values are known.

The missing changes are selected completely at random. To generate a dataset under Case II, one can start with the dataset generated under Case I. The following example is given to illustrate how to construct such datasets. Suppose the baseline covariates' values of the subject is $\boldsymbol{X}(t_0) = \boldsymbol{x}_0$ and the covariates values $\boldsymbol{X}$ change $J-1$ times at time $t_1, \ldots, t_{J-1}$, before the subject is censored or the event occurs at $t_J = \tilde{T}$. For $J = 3$, $\boldsymbol{X}(t_1) = \boldsymbol{x}_1$ and $\boldsymbol{X}(t_2) = \boldsymbol{x}_2$. The counting process approach assumes

$$\boldsymbol{X}(t) = \boldsymbol{x}_{j-1}, \quad t_{j-1} \leq t < t_j, \quad j = 1, 2, 3 \tag{13}$$

The information of the subject under Case I, displays exactly as in (13). For a dataset under Case II when only half of the changes are known, only one of $\{t_1, t_2\}$ is known. If only the change at t_k ($k = 1, 2$) is known, the observed information for the same subject is then

$$\begin{aligned} \boldsymbol{X}(t) &= \boldsymbol{x}_0, \quad 0 \leq t < t_k \\ \boldsymbol{X}(t) &= \boldsymbol{x}_k, \quad t_k \leq t \leq \tilde{T} \end{aligned} \tag{14}$$

Note that for both (13) and (14), the true survival curve is constructed using the information as in (13), when all history of changes in values are known.

3.3 Evaluation measures

Since the goal is to estimate the survival function, we evaluate estimation performance using the average integrated L_2 difference between the true and the estimated survival curves $\widehat{S}$. Given a dataset $\mathscr{D}$, containing N subjects, each with

pseudo-subject information up to the survival/censored time $\tilde{T}^{(i)}$, $\mathscr{X}^{(i)}(\tilde{T}^{(i)})$, $i = 1, 2, \ldots, N$,

$$L_2\,(\widehat{S}) = \frac{1}{N}\sum_{i\in\mathscr{D}}\frac{1}{\tilde{T}^{(i)}}\int_0^{\tilde{T}^{(i)}}[S^{(i)}(t) - \widehat{S}(t|\mathscr{X}^{(i)}(t))]^2 dt \tag{15}$$

Note that we evaluate the integrated L_2 difference only up to $\tilde{T}^{(i)}$, the last time point where the survival status is known. In the simulations, as we generate the true survival time $T^{(i)}$, we have the trajectory of covariate values up to time $T^{(i)}$ for any given subject i even when it is censored at time $C^{(i)} < T^{(i)}$. However, here we intend to match the scenario in real world applications where the covariate information is usually no longer recorded after the event occurs (e.g. the patient dies) or the subject is censored (e.g. lost contact). Thus, we define the best modeling method to be the one that gives us the lowest integrated L_2 difference, which is an average value from all subjects; for each subject, the difference between an estimated survival curve and the true survival curve up to its last observed time is measured.

3.4 Simulation results

The extended Cox model is included as a benchmark method, since it is one of the most commonly used methods in practice. Another benchmark method used in this paper is the Kaplan-Meier method, which uses no covariates' values to construct the survival function estimate; this helps illustrate the improvement in estimation from incorporating the covariate information.

In this section, we present simulation results based on 500 simulation trials. The number of trees for bootstrap samples is set to be 100 for all forest methods. We also only focus on the Weibull-Increasing distribution, and omit discussion of the Weibull-Decreasing distribution, since results for the latter distribution are similar. Detailed results are given in Section S1.7 of the Supplemental Material.

3.4.1 Regulating the construction of trees in forests. Figure 3 gives an example of how CIF-TV performs with different values of *mtry* in the scenario 2TI + 4TV, when the censoring rate is 20%, and the signal-to-noise ratio is low. The *mtry* values are

tuned based on the "out-of-bag observations." Similar results for RRF-TV and TSF-TV can be found in Section S1.6 in the Supplemental Material.

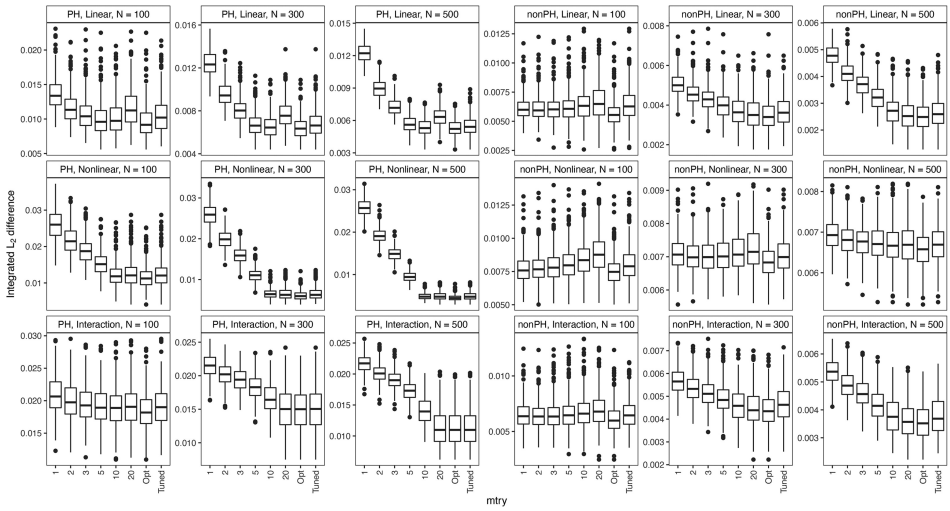

Figure 3. Integrated L_2 difference of CIF-TV with different $mtry$ values. Datasets are generated with a light right-censoring rate (20%), survival times following a Weibull-Increasing distribution. From the top row to the bottom, are given results for the linear, nonlinear, and interaction survival relationship; the first three columns show results under the PH setting for the number of subjects $N = 100, 300, 500$, respectively, and the last three columns for results under the non-PH setting. In each plot, 1—CIF-TV with $mtry = 1$; 2—CIF-TV with $mtry = 2$; 3—CIF-TV with $mtry = 3$; 5—CIF-TV with $mtry = 5$; 10—CIF-TV with $mtry = 10$; 20—CIF-TV with $mtry = 20$; Opt—CIF-TV with value of *mtry* that gives the smallest Integrated L_2 difference in each round; Tuned—CIF-TV with the value of $mtry$ tuned by the "out-of-bag" tuning procedure. The default value in conditional inference forest is $mtry = 5$.

In these examples, one can see that the forests using the "out-of-bag" tuning procedure give relatively good performance overall. In fact, results from other model setups are broadly similar, in the sense that this tuning procedure provides a relatively reliable choice of *mtry* and it gains in accuracy as the number of subjects N increases. In contrast, the default value of *mtry* does not always perform well, and choosing a different value can sometimes significantly improve performance.

Table 2 gives examples under the PH setting to show the performance comparison between each forest with its default parameter settings and with the proposed parameter settings.

Proportional hazards setting							
	Case I. All changes in covariates' values are known						
N	Extended Cox	CIF-TV(D)	CIF-TV(P)	RRF-TV(D)	RRF-TV(P)	TSF-TV(D)	TSF-TV(P)
100	0.57 ± 0.15	0.17 ± 0.26	0.46 ± 0.15	0.31 ± 0.20	0.43 ± 0.16	0.12 ± 0.27	0.36 ± 0.13
300	0.86 ± 0.04	0.24 ± 0.16	0.65 ± 0.07	0.37 ± 0.13	0.63 ± 0.08	0.15 ± 0.18	0.56 ± 0.07
500	0.92 ± 0.02	0.27 ± 0.12	0.71 ± 0.05	0.38 ± 0.10	0.69 ± 0.05	0.23 ± 0.16	0.63 ± 0.05
	Case II. Half of changes in covariates' values are unknown						
N	Extended Cox	CIF-TV(D)	CIF-TV(P)	RRF-TV(D)	RRF-TV(P)	TSF-TV(D)	TSF-TV(P)
100	0.30 ± 0.17	0.30 ± 0.19	0.37 ± 0.14	0.34 ± 0.17	0.35 ± 0.16	0.26 ± 0.17	0.27 ± 0.10
300	0.55 ± 0.06	0.39 ± 0.11	0.50 ± 0.06	0.44 ± 0.10	0.50 ± 0.08	0.37 ± 0.11	0.42 ± 0.06
500	0.59 ± 0.04	0.42 ± 0.08	0.54 ± 0.04	0.46 ± 0.07	0.53 ± 0.05	0.42 ± 0.08	0.47 ± 0.05
Non-proportional hazards setting							
	Case I. All changes in covariates' values are known						
N	Extended Cox	CIF-TV(D)	CIF-TV(P)	RRF-TV(D)	RRF-TV(P)	TSF-TV(D)	TSF-TV(P)
100	-0.55 ± 0.28	-0.39 ± 0.38	0.11 ± 0.21	-0.27 ± 0.33	0.12 ± 0.18	-0.28 ± 0.42	0.35 ± 0.21
300	-0.27 ± 0.10	-0.27 ± 0.27	0.44 ± 0.12	-0.14 ± 0.23	0.33 ± 0.12	-0.40 ± 0.31	0.62 ± 0.12
500	-0.23 ± 0.07	-0.24 ± 0.21	0.59 ± 0.09	-0.08 ± 0.18	0.47 ± 0.11	-0.30 ± 0.28	0.72 ± 0.08
	Case II. Half of changes in covariates' values are unknown						
N	Extended Cox	CIF-TV(D)	CIF-TV(P)	RRF-TV(D)	RRF-TV(P)	TSF-TV(D)	TSF-TV(P)
100	-0.51 ± 0.28	-0.10 ± 0.26	0.08 ± 0.16	-0.11 ± 0.26	0.09 ± 0.18	0.04 ± 0.29	0.21 ± 0.17
300	-0.16 ± 0.10	0.00 ± 0.19	0.22 ± 0.10	0.01 ± 0.18	0.20 ± 0.08	0.09 ± 0.22	0.37 ± 0.13
500	-0.11 ± 0.07	0.05 ± 0.16	0.28 ± 0.10	0.05 ± 0.15	0.24 ± 0.08	0.15 ± 0.17	0.43 ± 0.12

*Given a method *A*, each cell value are given as mean $\pm$ one standard deviation of $(L_2(\text{KM}) - L_2(\text{A}))/L_2(\text{KM})$ based on all simulations.
For similar results under other model setups, please refer to Section S1.6 in the Supplemental Material.

In Table 2, positive numbers indicate a decrease in integrated L_2 difference compared to a Kaplan-Meier fit on the dataset, while negative numbers indicate an increase. The absolute value of the numbers represents the size of the difference between the integrated L_2 difference of the candidate and that of a Kaplan-Meier fit. The table shows that forests with the proposed parameter settings can provide improved performance over those with default parameter settings across all different number of subjects N by a substantial amount. Note that, under the non-PH setting, for datasets with all of the changes in covariates' history known, the negative numbers indicate the poor performance of forests with default parameter settings even compared to a simple Kaplan-Meier curve, showing that the default methods can fail miserably. In contrast, for all forests with the proposed parameter settings, as N increases, the change in sign and in the absolute value of the numbers indicates better and better performance in general. Overall, the performance of the proposed parameter setting is relatively stable and better than that of the default values.

In the following discussion, we therefore only focus on the forest methods with the proposed parameter settings.

3.4.2 Properties of the proposed forest methods. Using factorial designs, we study the difference between each of the proposed forest methods and a simple Kaplan-Meier fit under the effects of the following factors: censoring rate, amount of knowledge, survival relationship, training sample size, scenario, hazards setting, and SNR. The effects are estimated based on an analysis of variance model fit with these factors as main effects. Figure 4 provides the main effects plots for the integrated L_2 difference improvement from the proposed forest methods over a simple Kaplan-Meier fit.

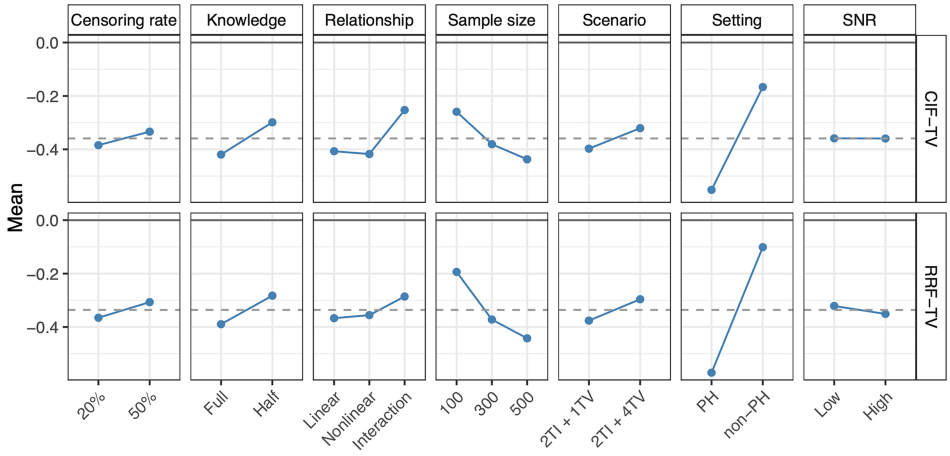

Figure 4. Main effects plots of integrated L_2 difference improvement from the proposed forests over a simple Kaplan-Meier fit. Given a method A (CIF-TV or RRF-TV), the difference improvement is computed as $(L_2(\mathrm{A}) - L_2(\mathrm{KM}))/L_2(\mathrm{KM})$. The solid line gives the zero value and the dashed line gives the mean value over all effects for a reference.

In Figure 4, the overall center of location is negative, highlighting that both of the proposed forest methods perform better than a simple Kaplan-Meier fit. The overall mean integrated L_2 difference of CIF-TV is slightly smaller than that of RRF-TV. The relative performance of the proposed forest methods can vary with changes in factors. Note that the p-values of the hypothesis testing on the simple main effect of SNR is insignificant at a 0.10 level, which suggests that the impact of the change of its level on the performance of the proposed forest methods over a Kaplan-Meier fit is negligible (more details can be found in Section S1.8 in the Supplemental Material).

For the other factors, the fewer the number of changes in values of covariates that are known, higher censoring rate, smaller training sample size, larger portion of covariates being time-varying, more relaxed assumption on the hazard setting, and more complicated structure of the survival relationship (all reflecting more difficult estimation tasks), the less the proposed forest methods improve over a simple Kaplan-Meier fit. Conversely, in the opposite situations, the stronger the ability of the proposed forest methods to estimate the underlying survival relationship and therefore bring a greater improvement.

In particular, the two proposed forest methods win by a much larger margin under the PH setting, which is expected since the log-rank-type splitting procedures used in the proposed forest methods rely to some extent on the PH assumption.

It is also clear that the difference between the number of time-invariant (TI) and the number of time-varying (TV) covariates is driving the scenario effect. When #TV − #TI increases, the relative performance of the proposed forest methods deteriorates. Presumably, this is because the increasing level of local time-varying effects makes the underlying relationship more difficult to estimate.

Note that the improvement from the proposed forest methods over a Kaplan-Meier fit remains relatively stable to the change of levels in both the censoring rate and the scenario factors compared to the change in other factors. In the following discussion, we mainly focus on the factors that are more influential based on our previous study: the number of changes in values of covariates that are known, the underlying survival relationship, the sample size, and the hazard function setting. The simulation results presented are based on the datasets generated under the scenario 2TI + 4TV, the lower SNR, with 20% censoring rate.

3.4.3 Estimation performance comparison. Figure 5 gives side-by-side boxplots of integrated L_2 difference defined in (15) for performance comparison under different model setups. It shows that for the linear survival relationship under the PH setting, the extended Cox model performs the best. This is expected as the extended Cox model relies exactly on the assumption of PH and a log-linear relationship between the hazard function and covariates. For nonlinear and interaction survival relationships, all forests outperform the extended Cox model, showing their advantage in dealing with a relatively complex situation. More specifically, for nonlinear setups, CIF-TV performs the best and RRF-TV the second, while for interaction model setups, RRF-TV performs the best and CIF-TV the second. In addition, CIF-TV and RRF-TV outperform TSF-TV across all different number of subjects and survival

relationships. Under the non-PH setting, the extended Cox model cannot even outperform a simple Kaplan-Meier fit on the dataset, whether all changes in values of covariates are known or not. As discussed, the presence of non-PH settings poses great challenges to modeling methods that assume PH; not just Cox, but also the survival forests like CIF-TV and RRF-TV that use a log-rank splitting rule. On the other hand, TSF-TV, which is specifically designed to detect non-PH deviations, performs the best across all different setups under the non-PH setting.

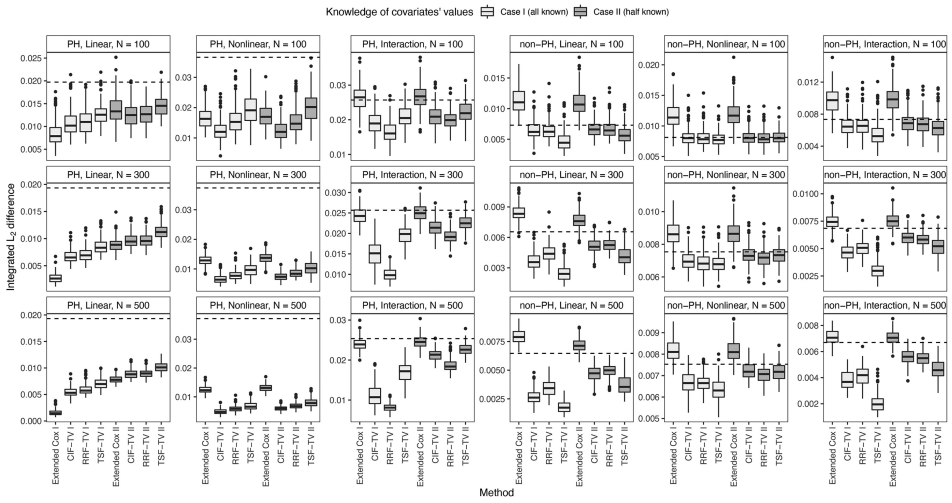

Figure 5. Boxplots of integrated L_2 difference for performance comparison. Datasets are generated with survival times following a Weibull distribution, light right-censoring rate (20%). The three rows show results for the number of subjects $N = 100, 300, 500$, respectively; the first three columns show results under the PH setting for survival relationship linear, nonlinear, and interaction, respectively, and the last three columns for results under the non-PH setting. The horizontal dashed line shows the median integrated L_2 difference of a Kaplan-Meier fit on the datasets. In each of the plots, the set of boxplots lightly shaded shows the performance of different methods on datasets with history of changes in covariates' values known; the set heavily shaded shows the performance on datasets with half of the changes in covariates' values unknown.

It is not surprising that having all changes in values known gives increasingly better performance compared to only having half of the changes known as the sample size increases. As N increases, false information due to unknown changes has a negative effect on performance of all modeling methods. In particular, this affects the extended Cox model more than the forest methods when the underlying survival relationship is linear under the PH setting, while it affects the extended Cox model less for all other cases. This is simply because the extended Cox model already performs poorly in

nonlinear and non-PH situations, so the misleading information from incorrect knowledge of covariates' values cannot hurt performance very much.

Generally, if the true underlying model setup is known, one should choose CIF-TV or RRF-TV under the PH setting, and TSF-TV under the non-PH setting. However, none of the forest methods can perform well all of the time. In the next section, we provide guidance on how to choose among these forest methods.

3.4.4 Guidance for choosing the modeling method. Cross-validation methods have been used in the past for the error estimation of survival models.[34] We propose to use one of the most common methods, K-fold cross-validation, implemented with integrated Brier scores for survival data, to select the "best" modeling method for a given dataset $\mathscr{D}$ as follows.

For a given survival curve estimate $\widehat{S}$,

i. Split the dataset into K non-overlapping subsets $\mathscr{D}_k$ $(k = 1, 2, \ldots, K)$, each containing (roughly) equal number of subjects;
ii. For each $k = 1, 2, \ldots, K$
   a. Modeling methods $\widehat{S}_k$ are then trained with the data $\mathscr{D} \setminus \mathscr{D}_k$ where the k-th subset is removed;
   b. Test $\widehat{S}_k$ on data in the k-th test set $\mathscr{D}_k$ and compute the corresponding integrated Brier score $\widehat{\mathrm{IBS}}(\widehat{S}_k; \mathscr{D}_k)$ as given in (7);
iii. Average over all K subsets and obtain

$$\mathrm{IBSCVErr}(\widehat{S}) = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathrm{IBS}}(\widehat{S}_k; \mathscr{D}_k) \tag{16}$$

We then choose the modeling method that gives the smallest $\mathrm{IBSCVErr}(\widehat{S})$ in (16).

For the simulated data sets, we use 10-fold cross-validation to choose between modeling methods. The measures p_{B}, r_{B}, and r_{W} are used to evaluate the performance,

$$p_{\mathrm{B}} = \#\{x_{\mathrm{CV}} = x_{\min}\}/n_{\mathrm{rep}} \tag{17}$$

$$r_{\mathrm{B}} = |x_{\min} - x_{\mathrm{CV}}|/x_{\min} \tag{18}$$

$$r_{\mathrm{W}} = |x_{\max} - x_{\mathrm{CV}}|/x_{\max} \tag{19}$$

where n_{rep} denotes the number of simulations ($n_{\mathrm{rep}} = 500$), x_{CV} denotes the integrated L_2 difference of the method chosen by cross-validation, and $x_{\min}$ and $x_{\max}$ denote the lowest and highest integrated L_2 differences among all modeling methods, respectively. In each round of simulation, we call the method that gives $x_{\min}$ the best modeling method and the method that gives $x_{\max}$ the worst modeling method. By definition, p_{B} provides the proportion of the times IBS-based 10-fold CV selects the best modeling method, and r_{B} and r_{W} compute the relative errors from the best and the worst modeling method, respectively. The smaller r_{B} is, or the larger r_{W} is, the better IBS-based 10-fold CV works.

Table 3 presents the summary of the performance of the IBS-based 10-fold CV rule. It is not surprising that IBS-based 10-fold CV works better under Case I where all changes in covariate values are known in general, with larger values of p_{B}, smaller values of r_{B} and larger values of r_{W}. That is, the incorrect knowledge of covariate values also hurts the performance of the selection procedure. In general, as the number of subjects N increases, the value of p_{B} gets larger for most of the scenarios, indicating IBS-based 10-fold CV is able to pick up the best modeling method at a higher frequency; even under those scenarios where p_{B} is lower than 50%, the relative error from the best modeling method r_{B} remains within 10% for most of the cases. Note that more than half of the cases for $N = 100$ have the relative error from the best modeling method r_{B} less than 10% and almost all of the cases for $N = 500$ have r_{B} under 5%. That means that even when the IBS-based 10-fold CV does not pick the best modeling method, it is still able to pick a method that works reasonably well, resulting in the integrated L_2 difference being not much higher than that of the best method.

Sample size	Setting	Relationship	Case I			Case II		
			p_B^{*a}	$r_B^{\dagger b}$	$r_W^{\dagger b}$	p_B^{*a}	$r_B^{\dagger b}$	$r_W^{\dagger b}$
$N = 100$	PH	Linear	0.34	0.35 ± 0.44	0.21 ± 0.18	0.35	0.14 ± 0.16	0.16 ± 0.12
		Nonlinear	0.66	0.11 ± 0.22	0.34 ± 0.18	0.69	0.09 ± 0.21	0.35 ± 0.16
		Interaction	0.71	0.03 ± 0.08	0.33 ± 0.11	0.33	0.08 ± 0.10	0.21 ± 0.09
	Non-PH	Linear	0.87	0.04 ± 0.16	0.56 ± 0.14	0.63	0.06 ± 0.13	0.45 ± 0.14
		Nonlinear	0.53	0.04 ± 0.07	0.33 ± 0.10	0.47	0.04 ± 0.07	0.32 ± 0.10
		Interaction	0.79	0.04 ± 0.12	0.42 ± 0.14	0.55	0.05 ± 0.10	0.34 ± 0.12
$N = 300$	PH	Linear	0.99	0.03 ± 0.40	0.69 ± 0.11	0.49	0.07 ± 0.11	0.19 ± 0.10
		Nonlinear	0.83	0.03 ± 0.14	0.49 ± 0.12	0.76	0.04 ± 0.12	0.45 ± 0.12
		Interaction	0.95	0.00 ± 0.02	0.58 ± 0.07	0.58	0.06 ± 0.09	0.19 ± 0.08
	Non-PH	Linear	0.95	0.01 ± 0.05	0.70 ± 0.09	0.77	0.04 ± 0.10	0.44 ± 0.13
		Nonlinear	0.59	0.02 ± 0.03	0.21 ± 0.06	0.42	0.02 ± 0.03	0.16 ± 0.05
		Interaction	0.97	0.01 ± 0.05	0.59 ± 0.11	0.73	0.03 ± 0.08	0.30 ± 0.11
$N = 500$	PH	Linear	1.00	0.00 ± 0.00	0.78 ± 0.07	0.79	0.03 ± 0.06	0.22 ± 0.08
		Nonlinear	0.80	0.02 ± 0.07	0.60 ± 0.08	0.78	0.03 ± 0.07	0.54 ± 0.07
		Interaction	0.84	0.01 ± 0.05	0.65 ± 0.05	0.86	0.03 ± 0.07	0.22 ± 0.07
	Non-PH	Linear	0.94	0.01 ± 0.06	0.77 ± 0.06	0.80	0.03 ± 0.10	0.48 ± 0.12
		Nonlinear	0.74	0.01 ± 0.02	0.22 ± 0.06	0.34	0.02 ± 0.02	0.12 ± 0.04
		Interaction	0.99	0.00 ± 0.01	0.71 ± 0.09	0.82	0.02 ± 0.06	0.34 ± 0.11

*a p_B is computed as in (17), with mean value over all simulations.
†b r_B and r_W are computed as in (18) and (19), respectively, with mean value $\pm$ one standard deviation over all simulations.

3.5 Proposed forest methods for time-invariant covariate data

We have focused on ensemble methods for survival data with time-varying covariates, as we feel that this is a very common and important situation that has been understudied in the past. Having said that, there are certainly many situations in which only time-invariant (baseline) covariate information is available, and understanding the properties of different methods in that situation is important. Section S2 in the Supplemental Material describes the results of simulations related to this question. In those simulations, datasets with left-truncated and right-censored survival times are generated based on time-invariant covariates.

In fact, the simulation results of all comparative estimation performance in the case of time-invariant covariates are broadly similar to those in the time-varying covariates cases. That is,

1. The "out-of-bag" tuning procedure can provide a reliable choice of *mtry* that gives relatively good performance in general. One should also consider adjusting other tuning parameters such as *minsplit*, *minbucket* in the conditional inference forest and the transformation forest, and *nodesize* in random survival forest, as the size of dataset grows.
2. Taking into an account all other factors, under the PH setting, the best method is always one of the two proposed forests, while under the non-PH setting, it is

the transformation forest method.

3. The IBS-based CV rule is a good option for choosing among the various methods, as the comparative performance of methods appears to be different from setting to setting.

3.6 Real data application

We now illustrate application of the proposed time-varying covariates forests to a real data set, the Mayo Clinic Primary Biliary Cirrhosis Data, available in the R package survival. To study the effectiveness of using D-penicillamine as treatment, 312 patients with primary biliary cirrhosis (PBC) were enrolled in a randomized medical trial at the Mayo Clinic from January 1974 to May 1984.[35] The outcome of interest is the time to death for these patients. In this study, medical measurements and other patient information were recorded as covariates' values at entry and at yearly intervals. The Cox model was used to estimate the survival function for patients with primary biliary cirrhosis based on 12 noninvasive, easily collected covariates that require only a blood sample and clinical evaluation.[35] These 12 covariates include age at entry, alkaline phosphotase (U/L), logarithm of serum albumin (g/dL), presence of ascites, aspartate aminotransferase (U/mL), logarithm of serum bilirubin (mg/dL), serum cholesterol (mg/dL), condition of edema, presence of hepatomegaly or enlarged liver, platelet count, logarithm of prothrombin time and presence or absence of spiders. As the study was extended for another four years, a total of 1945 visits were generated. All of the 12 covariates except age become time-varying covariates in the follow-up data. At the end of the follow-up study, 169 of the 312 patients were still alive, 140 had died, and three had been lost contact with. The extended study allows the researchers to study the effects of the changes in the prognostic variables, as time-varying covariates.[36] We therefore fit the proposed forest methods as well as the extended Cox model[36] on the dataset with time-varying covariates from the extended study. Note that estimates of this kind for different sets of covariate values over time can be useful in providing guidance from a public policy point of view, as they highlight the different average survival experiences of different subpopulations with different covariate paths. To better illustrate the effects of the time-varying covariates, we also consider the corresponding time-invariant dataset where the covariate values are never updated after the initial observation. For performance comparison, the Brier score-based 10-fold cross-validation error at the t-th recording day is computed for each method on both the time-varying and time-invariant covariate datasets, as shown

in Figure 6. The corresponding integrated Brier score cross-validation results are given in Table 4.

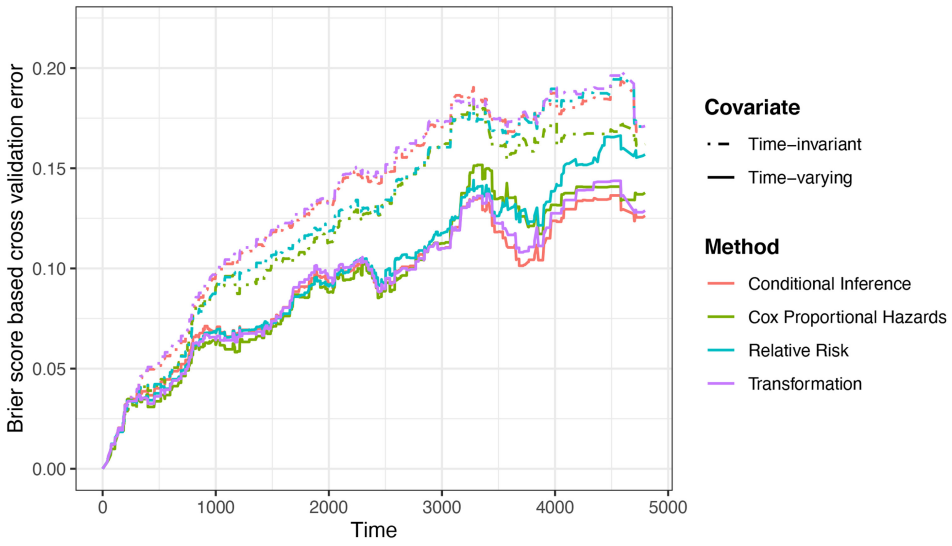

Figure 6. Brier score-based 10-fold cross-validation errors at t-th recording day provided for (1) the extended Cox model, CIF-TV, RRF-TV, and TSF-TV on the PBC data with time-varying covariate values obtained in the extended study; (2) Cox model, CIF, RRF, and TSF on the PBC dataset with only the initial covariate values (i.e. with time-invariant covariates). The results are shown up to time point where only 5% of the subjects are still at risk.

Covariate	Cox	CIF	RRF	TSF
Time-invariant	0.1245	0.1354	0.1302	0.1371
Time-varying*	0.0984	0.0952	0.1049	0.0961

*For time-varying covariate data, the results are shown for the extended Cox model, CIF-TV, RRF-TV, and TSF-TV, respectively.

Figure 6 shows that the cross-validation errors from all methods on the time-varying covariate dataset are lower than the corresponding ones in the time-invariant covariate dataset after $t > 300$, which suggests that the updated covariate information can significantly improve performance. Given the results on the time-varying covariate data in Figure 6, one can see that the differences in performance are relatively small. The extended Cox model slightly outperforms the others before $t = 3000$, where around 30% of the subjects are still at risk. CIF-TV and TSF-TV give the best performance between $t = 3000$ and $t = 4200$. After $t = 4200$, since the results are based on very few failures (given only 10% of subjects are still at risk), the Brier scores are highly variable and not as reliable. Note that the Brier score results also suggest that the survival relationship between the hazards and covariates is relatively (log-)linear for $t < 3000$, but not for $t > 3000$. Based on these Brier score-based

cross-validation results, we would recommend an extended Cox model for $0 < t < 3000$, and CIF-TV for $t > 3000$. Table 4 shows that CIF-TV gives the lowest corresponding integrated Brier score cross-validation errors, suggesting that CIF-TV provides the best estimated accuracy overall. In practice, where the underlying true survival distribution may possess a complex structure with time-varying features, a plot of the cross-validation-based Brier scores as shown in Figure 6 can potentially help data analysts make better choices for survival estimation at different time points, compared to a universal choice of method for all time points.

4 Discussion

The estimation of a population-level survival probabilities for time-varying covariate data are useful in many survival analysis settings, causal inference research being an important example. Additional challenges arise when considering time-varying treatment regimens. To draw real-world evidence about the effectiveness of such regimens on patient survival, the key is to account for the time-varying confounding effects and one way to address this issue is by using the inverse probability of treatment weighting.[37,38] For survival data, the estimation of the time-varying weights can potentially be improved by using flexible tree-based methods allowing time-varying covariates (confounders). A recent work[39] that uses LTRC forests shows that the use of more flexible models for the estimation of time-varying weights can lead to more accurate treatment effect estimation.

In this paper, we propose extending the relative risk, conditional inference, and transformation forests to provide survival function estimation with time-varying covariates. The extension of random survival forests[19] for cumulative hazard function estimation with time-varying covariates based on log-rank-type splitting rules requires reformulation of the log-rank test statistics[40,41] and/or the log-rank scores.[42] For a log-rank score rule, one can make use of the log-rank score specified in (4). However, in this case one cannot construct the extension by modifying the custom splitting rule feature and employing the fast C code from randomForestSRC, since its basic structure only allows for right-censored survival. A fast and efficient algorithm to construct random survival forests for left-truncated right-censored survival data is needed, and is a goal of future work.

5 Conclusion

In this paper, we have proposed two new ensemble algorithms, CIF-TV and RRF-TV, and adapted the transformation algorithm, TSF-TV. These three forest algorithms can handle (left-truncated) right-censored survival data with time-varying covariates and provide dynamic estimation. The tuning parameters in the proposed forest methods for survival data with time-varying covariates affect their overall performance. Guidance on how to choose those parameters is provided to improve on the potentially poor performance of forests with the default parameter settings.

The estimation performance of the proposed forest methods is investigated to understand how the improvement over a Kaplan-Meier fit is related to changes in different factors. Focusing on the more influential factors, the estimation performance comparison against other methods shows that the proposed forest methods outperform others under certain circumstances, while no method can dominate in all cases. We then provide guidance for choosing the modeling method in practice, showing that cross-validation is able to pick the best modeling method most of the time, or at least select a method that performs not much worse than the best method.

Our developed methodology and algorithms allow for estimation using the proposed forests for (left-truncated) right-censored data with time-invariant covariates. The same data-driven guidance for tuning the parameters or selecting a modeling method also applies to the time-invariant covariates case (for both left-truncated right-censored survival data and right-censored survival data), which implies its broad effectiveness regardless of additional left-truncation and regardless of the presence of time-varying effects.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Weichi Yao https://orcid.org/0000-0002-3412-5317

Denis Larocque https://orcid.org/0000-0002-7372-7943

Notes

Code availability	R scripts for reproducibility of the simulations and real dataset illustrative example analysis are available from the github repository, https://github.com/WeichiYao/TimeVaryingData_LTRCforests. An R package, LTRCforests, which implements CIF-TV and RRF-TV for LTRC data with application to time-varying data, is available on CRAN.
Supplemental material	Supplemental material for this article is available online.

References

1. Crowley J, Hu M. Covariance analysis of heart transplant survival data. *J Am Stat Assoc* 1977; 72: 27–36.
2. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. *J Am Stat Assoc* 1995; 90: 27–37.
3. Cox DR. Regression models and life-tables. *J R Stat Soc, Ser B* 1972; 34: 187–202.
4. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat* 1982; 10: 1100–1120.
5. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*. Boca Raton, FL, USA: CRC Press, 2012.
6. LeBlanc M, Crowley J. Relative risk trees for censored survival data. *Biometrics* 1992; 48: 411–425.
7. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* 2006; 15: 651–674.
8. Fu W, Simonoff JS. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics* 2017; 18: 352–369.
9. Hothorn T, Zeileis A. Predictive distribution modeling using transformation forests. *J Comput Graph Stat* 2021; 30: 1181–1196.
10. Sun Y, Chiou SH, Wang MC. ROC-guided survival trees and ensembles. *Biometrics* 2020; 76: 1177–1189.
11. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med Res Methodol* 2020; 20: (1). Crossref.
12. Bou-Hamad I, Larocque D, Ben-Ameur H. Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. *Stat Modelling* 2011; 11: 429–446.
13. Schmid M, Welchowski T, Wright MN et al. Discrete-time survival forests with hellinger distance decision trees. *Data Min Knowl Discov* 2020; 34: 812–832.

14. Kretowska M. Oblique survival trees in discrete event time analysis. *IEEE J Biomed Health Inform* 2020; 24: 247–258.

15. Puth MT, Tutz G, Heim N et al. Tree-based modeling of time-varying coefficients in discrete time-to-event models. *Lifetime Data Anal* 2020; 26: 545–572.

16. Moradian H, Yao W, Larocque D et al. Dynamic estimation with random forests for discrete-time survival data. *Canadian J Stat* 2022; 50: 533–548.

17. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–22.

18. Ishwaran H, Blackstone EH, Pothier C et al. Relative risk forests for exercise heart rate recovery as a predictor of mortality. *J Am Stat Assoc* 2004; 99: 591–600.

19. Ishwaran H, Kogalur UB, Blackstone EH et al. Random survival forest. *Ann Appl Stat* 2008; 2: 841–860.

20. Hothorn T, Bühlmann P, Dudoit S et al. Survival ensembles. *Biostatistics* 2006; 7: 355–373.

21. Gross ST, Lai TL. Nonparametric estimation and regression analysis with left-truncated and rightcensored data. *J Am Stat Assoc* 1996; 91: 1166–1180.

22. Tsai WY, Jewell NP, Wang MC. A note on the product-limit estimator under right censoring and left truncation. *Biometrika* 1987; 74: 883–886.

23. Breiman L, Friedman JH, Olshen RA et al. *Classification and Regression Trees*. Wadsworth, Belmont, California: Taylor & Francis, 1984.

24. Hothorn T, Seibold H, Zeileis A. *partykit: A toolkit with infrastructure for representing, summarizing, and visualizing tree-structured regression and classification models*, 2020. R package version 1.2-7.

25. Ishwaran H, Kogalur UB. *Fast unified random forests for survival, regression, and classification (RF-SRC)*, 2020. LR package version 2.9.3.

26. Bacchetti P, Segal MR. Survival trees with time-dependent covariates: application to estimating changes in the incubation period of AIDS. *Lifetime Data Anal* 1995; 1: 35–47.

27. Hothorn T. *trtf: Transformation trees and forests*, 2020. R package version 0.3-7.

28. Breiman L, Cutler A, Liaw A et al. *randomForest: Breiman and Cutler's random forests for classification and regression*., 2018. R package version 4.6-14.

29. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, NY, USA: Springer Series in Statistics Springer New York Inc., 2001.

30. Yao W, Frydman H, Simonoff JS. An ensemble method for interval-censored time-to-event data. *Biostatistics* 2021; 22: 198–213.

31. Graf E, Schmoor C, Sauerbrei W et al. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999; 18: 2529–2545.

32. Therneau TM, Lumley T, Elizabeth A et al. *survival: Survival analysis*, 2020. R package version 3.1-12.
33. Austin PC. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Stat Med* 2012; 31: 3946–3958.
34. Gerds TA, Schumacher M. Efron-type measures of prediction error for survival analysis. *Biometrics* 2007; 63: 1283–1287.
35. Dickson ER, Grambsch PM, Fleming TR et al. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* 1989; 10: 1–7.
36. Murtaugh PA, Dickson ER, van Dam GM et al. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology* 1989; 20: 126–134.
37. Barber JS, Murphy SA, Verbitsky N. Adjusting for time-varying confounding in survival analysis. *Sociol Methodol* 2004; 34: 163–192.
38. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiolog. *Epidemiology* 2000; 11: 550–560.
39. Hu L, Li F, Ji J et al. Estimating the causal effects of multiple intermittent treatments with application to COVID-19, 2022. arXiv:2109.13368v2.
40. LeBlanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc* 1993; 88: 457–457.
41. Segal MR. Regression trees for censored data. *Biometrics* 1988; 44: 35–47.
42. Hothorn T, Lausen B. On the exact distribution of maximally selected rank statistics. *Comput Stat Data Anal* 2003; 43121–137.

Appendix

A. Dynamically adjusted survival function for the hypothetical COVID-19 example

The time-to-event is the time to a positive COVID test. There is a time varying covariate, $X(t)$, which describes the vaccination status of a subject with values in $\{0, 1, 2, 3\}$ where $0 =$ unvaccinated, $1 =$ vaccinated with a single dose and $2 =$ vaccinated with two doses, and $3 =$ vaccinated with two doses and a booster. The covariate may change values at discrete time points; in this example at times 1 and 2.

We construct the hypothetical tree for COVID progression, see Figure 7. At time $t = 0$, there is a sample of $n = 100$ unvaccinated and COVID-free subjects at the root of the tree, that is, every subject has $X(0) = 0$. Thus, the estimated survival function at time $t = 0$, $\widehat{S}(0) = 1$. Each of the branches on the tree are labeled b_0, b_{00}, etc. The initial branch is b_0; it represents a subject not vaccinated in the time interval $[0, 1)$. Similarly, branch b_{00} represents a subject unvaccinated over the time interval $[0, 2)$; branch b_{01}, a subject vaccinated with at

single dose at time $t = 1$; b_{011}, a subject vaccinated with a single dose at time $t = 1$ and not receiving second dose at time $t = 2$, etc. For simplicity, we will assume that there is no censoring.

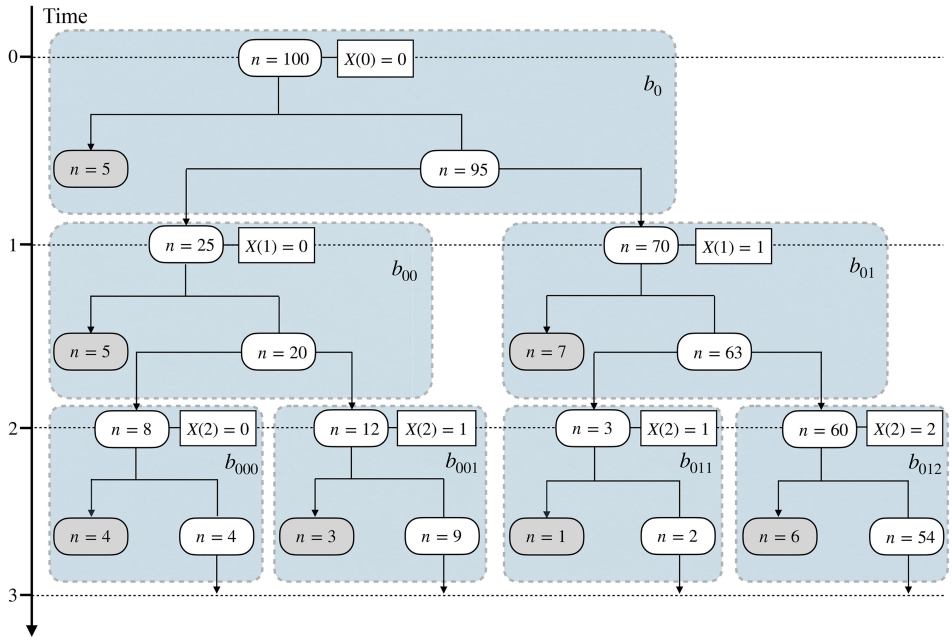

Figure 7. The hypothetical tree for COVID progression during $[0, 3)$. The tree nodes are shown as ovals with values indicating the number of subjects in the corresponding group. Darker nodes stand for the groups of subjects infected with COVID, and the lighter ones for COVID-free. At time $t = 0$, 1, and 2, the updated vaccination status $X(t)$ are shown in the squares alongside with the nodes. Each of the seven gray shaded areas corresponds to the group with a different vaccination status. For example, the b_0 area indicates the group unvaccinated in time interval $[0, 1)$, b_{01} the group vaccinated at time 1, and b_{012} the group vaccinated at times $t = 1$ and 2. Similar interpretations hold for the other gray areas.

At time $t = 1$, there are five subjects infected with COVID and 95 COVID-free. Hence, the estimated survival function at time $t = 1$ is

$$\widehat{S}(1) = \widehat{S}(1 \mid X(u) = 0, 0 \leq u < 1) = \widehat{S}(1 \mid b_0) = 0.95$$

Of the 95 COVID-free subjects at time $t = 1$, 70 get their first dose of vaccine and 25 remain unvaccinated. At time $t = 2$ of the 25 unvaccinated subjects, 20 are COVID-free and of 70 subjects who received a single vaccine, 63 are COVID-free. Thus, the estimated survival function at time $t = 2$ on b_{00} $(X(u) = 0,\ 0 \leq u < 2)$ is

$$\widehat{S}(2 \mid b_{00}) = \widehat{S}(1 \mid b_0)(20/25) = 0.95(0.8) = 0.76$$

and on b_{01} is

$$\widehat{S}(2 \mid b_{01}) = \widehat{S}(1 \mid b_0)(63/70) = 0.95(0.9) = 0.855$$

Of the 63 COVID-free vaccinated subjects at time 1, 60 receive the second dose of vaccine at time $t = 2$ and 3 do not. Of those 60, 54 are COVID-free at time $t = 3$. Thus, the estimated survival probability at $t = 3$ for this group is

$$\widehat{S}(3 \mid b_{012}) = \widehat{S}(2 \mid b_{01})(54/60) = 0.855(0.9) = 0.7695$$

For the group of three subjects with no second vaccine at time $t = 2$, the estimated survival probability at time $t = 3$ is lower:

$$\widehat{S}(3 \mid b_{011}) = \widehat{S}(2 \mid b_{01})(2/3) = 0.855(2/3) = 0.57$$

At time $t = 2$, of the 20 unvaccinated COVID-free subjects, 12 receive their first dose of vaccine and eight do not. Of those 12 subjects, nine are COVID-free at time $t = 3$. For this group, which gets a single vaccine at time $t = 2$, the estimated survival probability at $t = 3$ is

$$\widehat{S}(3 \mid b_{001}) = \widehat{S}(2 \mid b_{00})(9/12) = 0.76(0.75) = 0.57$$

which happens in this example to be the same as the estimated survival probability of the group that got the single vaccine at time $t = 1$.

Finally, for the non-vaccinated group (b_{000}), the estimated survival probability at time $t = 3$ is the lowest:

$$\widehat{S}(3 \mid b_{000}) = \widehat{S}(2 \mid b_{00})\,(4/8) = 0.76(0.5) = 0.38$$

The subjects who did or did not receive a booster at the next time period would be handled in a similar way.

B. Derivation of the survival estimate

Recall that by definition of survival functions, $S(t|\mathcal{X}^*(t)) = \mathbb{P}(T > t|\mathcal{X}^*(t))$. At given time $t \in [t_j^*, t_{j+1}^*)$, note that $\mathbb{P}(T > t|\mathcal{X}^*(t)) = \mathbb{P}(T > t,\ T > t_j^* \mid \mathcal{X}^*(t))$, we apply the conditional probability and obtain

$$S(t|\mathcal{X}^*(t)) = \mathbb{P}\big(T > t \mid T > t_j^*,\ \mathcal{X}^*(t)\big) S\big(t_j^* \mid \mathcal{X}^*(t_j^*)\big) \tag{20}$$

In constructing the survival function estimate, we assume that the hazard at time t is a function only of the current covariate values at time t (but these covariates can include lagged values of some covariates). This allows us to construct the estimate at time t using any subjects in the population with the specified value at that precise time point; that is, we estimate $\mathbb{P}\big(T > t \mid T > t_j^*,\ \mathcal{X}^*(t)\big)$ by computing

$$\widehat{\mathbb{P}}\big(T > t \mid T > t_j^*,\ \boldsymbol{x}_j^*\big) = \frac{\widehat{\mathbb{P}}\big(T > t \mid \boldsymbol{x}_j^*\big)}{\widehat{\mathbb{P}}\big(T > t_j^* \mid \boldsymbol{x}_j^*\big)}$$

where both the numerator and the denominator are the values of the estimated survival function in the hypothetical case with the covariate $\boldsymbol{x}_j^*$ at t and t_j^*, respectively. The risk sets that are used to compute these two quantities consider all subjects with covariate values $\boldsymbol{x}_j^*$ at t_j^*, regardless of their covariate paths before t_j^*. Note that this hypothetical estimated survival function is in fact the output of the algorithm for the input with covariate value $\boldsymbol{x}_j^*$.

Therefore, by substituting $\widehat{S}_{A,j}(t) \triangleq \widehat{\mathbb{P}}\big(T > t \mid \boldsymbol{x}_j^*\big)$, we can then approximate (20) as

$$\widehat{S}\big(t \mid \mathcal{X}^*(t)\big) = \frac{\widehat{S}_{A,j}(t)}{\widehat{S}_{A,j}(t^*)} \widehat{S}\big(t_j^* \mid \mathcal{X}^*(t_j^*)\big)$$

which gives us the formula in (9).