Title
Regression trees and ensembles for cumulative incidence functions

Journal
The International Journal of Biostatistics, 18(2)

ISSN
2194-573X

Authors
Cho, Youngjoo
Molinaro, Annette M
Hu, Chen
et al.

Publication Date
2022-11-01

DOI
10.1515/ijb-2021-0014

Peer reviewed

Youngjoo Cho, Annette M. Molinaro, Chen Hu and Robert L. Strawderman*

Regression trees and ensembles for cumulative incidence functions

https://doi.org/10.1515/ijb-2021-0014
Received February 15, 2021; accepted March 2, 2022; published online March 25, 2022

Abstract: The use of cumulative incidence functions for characterizing the risk of one type of event in the presence of others has become increasingly popular over the past two decades. The problems of modeling, estimation and inference have been treated using parametric, nonparametric and semi-parametric methods. Efforts to develop suitable extensions of machine learning methods, such as regression trees and ensemble methods, have begun comparatively recently. In this paper, we propose a novel approach to estimating cumulative incidence curves in a competing risks setting using regression trees and associated ensemble estimators. The proposed methods use augmented estimators of the Brier score risk as the primary basis for building and pruning trees, and lead to methods that are easily implemented using existing R packages. Data from the Radiation Therapy Oncology Group (trial 9410) is used to illustrate these new methods.

Keywords: Brier score; CART; cause-specific hazard; competing risks; Fine and Gray model; random forests.

1 Introduction

A subject being followed over time may experience several types of events, possibly even fatal. For example, in a Phase III trial of concomitant versus sequential chemotherapy and thoracic radiotherapy for patients with inoperable non-small cell lung cancer (NSCLC) conducted by the Radiation Therapy Oncology Group (RTOG), patients were followed up to 5 years, where both the occurrence of disease progression and death are of particular interest. Such "competing risks" data are often encountered in cancer and other biomedical follow-up studies, in addition to the potential complication of right-censoring on the event time(s) of interest.

Two quantities are often used when analyzing competing risks data: the cause-specific hazard function (CSH) and the cumulative incidence function (CIF). For a given event, the former describes the instantaneous risk of this event at time t, given that no events have yet occurred; the latter describes the probability of occurrence, or absolute risk, of that event across time and can be derived directly from the subdistribution hazard function [1]. Dignam et al. [2] provides a review of methods for handling competing risks data as of 2012, where parametric and semi-parametric approaches to modeling both the CSH and CIF using hazard-type regression modeling are considered. The literature on tree-based methods for estimating the CIF, including ensemble approaches like random forests [RF; 3], remains comparatively under-developed. Indeed, there is no software package currently available that specifically focuses on estimating the CIF using regression tree methods, and ensemble-based methods for estimating the CIF are currently limited to the work of [4, 5]. The methods described in [4] are implemented as part of the `randomForestSRC` package [6], where

*Corresponding author: **Robert L. Strawderman**, Department of Biostatistics & Computational Biology, University of Rochester, Rochester, NY, United States, E-mail: robert_strawderman@urmc.rochester.edu
Youngjoo Cho, Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea, E-mail: yvc5154@konkuk.ac.kr
Annette M. Molinaro, Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA, E-mail: annette.molinaro@ucsf.edu
Chen Hu, Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MA, USA, E-mail: chu22@jhmi.edu

the unpruned regression trees that make up the bootstrap ensemble are built using logrank-type splitting rules appropriate for competing risks [e.g., 7]. The method of [5] instead replaces all censored observations with jackknife pseudovalues (i.e., an imputation) derived from the Aalen-Johansen estimator [8] for a specific cause of failure m at particular follow-up time t; these authors go on to suggest that these imputed responses can then be used in any software package capable of fitting a random forest. Further discussion on both approaches, and contrasts to the methods to be proposed in this paper, can be found in Section 3.3.

In its most general form, the original CART algorithm, and by extension RF, relies on the specification of a loss function that (i) informs all decision-making processes (e.g., what covariate to split on and when/where; when to stop tree growth) and (ii) induces a particular estimator that minimizes the empirical loss. Motivated by the recent work of [9, 10] for right-censored survival data, this paper proposes a direct extension of CART and RF for estimating the CIF in the presence of right-censored competing risks. Specifically, starting with an appropriate version of the Brier loss function [cf., 11] (i.e., squared error loss for a binary outcome), we first develop a simple nonparametric estimate of the CIF for a single event by minimizing this loss function when there is no loss to follow-up (i.e., with full data) and one has specified a fixed partition structure for the covariate space. Estimation in this case may be viewed as a form of binomial regression, where the mean function (i.e., CIF) is piecewise constant on the covariate space. For the case where there is loss to follow-up, we then construct several observed data loss functions that target the same expected loss as the (unobserved) full data Brier loss function. The simplest of these approaches employs inverse probability of censoring weighted estimation (IPCW). Finally, we explain how the development of these new loss functions leads to new splitting and decision rules that can be used by CART and RF algorithms for estimating the CIF, and importantly, show how these new methods can be easily implemented using existing software in combination with a certain form of imputation. The resulting methods may be viewed as nonparametric alternatives to the semiparametric binomial regression approach proposed in [12] for estimating a CIF, differing in the approach to estimation (i.e., through minimizing the Brier loss instead of employing estimating equations). Simulation studies are used to investigate performance of these new methods. In addition, we use these new methods to conduct some secondary analyses for the RTOG 9410 Phase III lung cancer trial mentioned at the beginning of this section. The paper concludes with comments on future work.

2 Estimating a CIF by minimizing squared error loss

2.1 Relevant Data Structures

Let $T^{(m)}$ be the time to event for the event type $m = 1, \dots, K$ where $K \geq 2$ is fixed. Let W be a vector of p covariates, where $W \in S \subset \mathbb{R}^p$. Let $T = \min(T^{(1)}, \dots, T^{(K)})$ be the minimum of all latent event times; it is assumed that T is observed and has a continuous distribution function. Then, in the absence of other loss to follow-up, $F = (T, W, M)$ is assumed to be the fully observed (or full) data for a subject, where M is the observed event type that corresponds to T. The definition of T therefore implies that $(T^{(M)}, M, W)$ is observed. Moreover, it is known for $m \neq M$ that $T^{(m)} > T^{(M)}$ even though $T^{(m)}$ itself is not observed. Define $\mathcal{F} = (F_1, \dots, F_n)$ to be the full data observed on n independent subjects, where $F_i = (T_i, W_i, M_i), i = 1, \dots, n$ are assumed to be identically distributed (i.i.d.).

In the case where there is also potential random loss to follow-up, we suppose that C is a continuous random variable that, given W, is statistically independent of (T, M). Then, for a given subject, we instead observe $O = \{\tilde{T}, \Delta, M\Delta, W\}$, where $\tilde{T} = \min(T, C)$ and $\Delta = I(T \leq C)$ is the (any) event indicator. The observed data on n i.i.d. subjects is $\mathcal{O} = (O_1, \dots, O_n)$. Similarly to the case where $K = 1$, random censoring on T permits estimation of the CIF from the data $\mathcal{O}$. We remark here that the notational set-up intentionally excludes C from the set of possible event times $(T^{(1)}, \dots, T^{(K)})$; the reason for setting the problem up in this way will become clear in Section 2.3.

2.2 CIF estimation via the Brier loss: no loss to follow-up

Let $\psi_{0m}(t;\ w) = P(T \le t, M = m|W = w)$ and define $\Psi_0 = \{\psi_{0m}(t; w), t \ge 0; w \in \mathcal{S}, m = 1, \dots, K\}$. The set of CIFs Ψ_0 can be estimated from the data $\mathcal{F}$ using any suitable parametric or semiparametric method without further assumptions on the data (e.g., independence of $T^{(1)}, \dots, T^{(K)}$). This section describes a simple method for estimating $\psi_{0m}(t;\ w)$ for a fixed cause m and time point $t > 0$ using the Brier loss function. As preparation for Section 3, $\psi_{0m}(t;\ w)$ is assumed to be piecewise constant as a function of W; however, the basic estimation ideas extend to more complex modeling assumptions in a straightforward manner [e.g., 12].

Let $\mathcal{N}_1, \dots, \mathcal{N}_L$ form a known partition of $\mathcal{S}$. In this section and also in Section 2.3, we assume this partition is given and, consistent with the assumption that $\psi_{0m}(t;\ w)$ is a piecewise constant function of w, that $\psi_{0m}(t; w) = \sum_{l=1}^{L} \beta_{0lm}(t) I\{w \in \mathcal{N}_l\}$, where $\beta_{0lm}(t) = P(T \le t, M = m|W \in \mathcal{N}_l)$ is the same function of t for each $W \in \mathcal{N}_l$. Define $Z_m(t) = I(T \le t, M = m)$ and let

$$\psi_m(t; w) = \sum_{l=1}^{L} \beta_{lm}(t) I\{w \in \mathcal{N}_l\} \tag{1}$$

be a model for $\psi_{0m}(t;\ w)$, $w \in \mathcal{S}$. Then, fixing both $t > 0$ and m, the so-called Brier loss is given by $L_{m,t}^{\text{full}}(F, \psi_m) = \{Z_m(t) - \psi_m(t; w)\}^2 = \sum_{l=1}^{L} I\{w \in \mathcal{N}_l\}\{Z_m(t) - \beta_{lm}(t)\}^2$. Assuming that $\mathcal{F}$ is observed, the corresponding empirical Brier loss is given by

$$L_{m,t}^{\text{emp}}(\mathcal{F}, \psi_m) = \frac{1}{n} \sum_{i=1}^{n} L_{m,t}^{\text{full}}(F_i, \psi_m) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} I\{W_i \in \mathcal{N}_l\}\{Z_{im}(t) - \beta_{lm}(t)\}^2. \tag{2}$$

With t and m fixed and under the assumptions of Section 2.1, $L_{m,t}^{\text{full}}(F, \psi_m)$ is an unbiased estimator of the risk $\Re(t, \psi_m) = E\left[\sum_{l=1}^{L} I\{W \in \mathcal{N}_l\}\{Z_m(t) - \beta_{lm}(t)\}^2\right]$; hence, so is (2). Considered as a function of $\beta_{lm}(t), l = 1, \dots, L$, the risk $\Re(t, \psi_m)$ is minimized when $\beta_{lm}(t) = \beta_{0lm}(t)$ for each l; the loss (2) is minimized when $\psi_m(t; w) = \hat{\psi}_m(t; w) = \sum_{l=1}^{L} I\{w \in \mathcal{N}_l\}\hat{\beta}_{lm}(t)$, where

$$\hat{\beta}_{lm}(t) = \frac{\sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} Z_{im}(t)}{\sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\}} \tag{3}$$

is a nonparametric estimate for $\beta_{0lm}(t)$. By contrast, [12] use a semiparametric binomial regression model to estimate $\psi_{0m}(t, w)$ from $(Z_{im}(t), W_i), i = 1, \dots n$.

2.3 CIF estimation via the Brier loss: random loss to follow-up

In follow-up studies with competing risks outcomes, the full data $\mathcal{F}$ might not be observed due to loss to follow-up. In this case, estimating $\psi_{0m}(t;\ w)$ for a specified m under the loss function (2) is not possible. One way to overcome this challenge is to use a modified loss function that (i) depends only on the observed data $\mathcal{O}$ and (ii) has the same risk as the (unobserved) full data loss [c.f., 14, 13, 9]. Following [9, 10], we propose an appropriate class of inverse probability of censoring weighted (IPCW), and subsequently augmented IPCW (AIPCW), loss functions that share the same risk $\Re(t, \psi_m)$ as the (unobservable) empirical loss (2). This allows us to derive a new observed data estimator of the CIF with both t and m fixed. We then extend this class of losses to the setting of a composite loss function, where the goal is to simultaneously estimate $\psi_{0m}(t_j;\ w)$ at time $t_j, j = 1, \dots, J$. As in the previous section, we assume that $\psi_{0m}(t; w) = \sum_{l=1}^{L} \beta_{0lm}(t) I\{w \in \mathcal{N}_l\}$, where the partition $\{\mathcal{N}_1, \dots, \mathcal{N}_L\}$ of $\mathcal{S}$ is known.

2.3.1 CIF estimation via the IPCW and AIPCW Brier losses

Fix $t > 0$, define $G_0(s|W) = P(C \geq s|W)$ for any $s \geq 0$ and suppose that $G_0(T_i|W_i) \geq \epsilon$ almost surely for some $\epsilon > 0$ ($i = 1, \ldots, n$). Define $\tilde{Z}_{im}(t) = I(\tilde{T}_i \leq t, M_i = m)$, $i = 1, \ldots, n$; easy calculations then show

$$E\left[\frac{\Delta_i}{G_0(\tilde{T}_i|W_i)}(\tilde{Z}_{im}(t) - \psi_m(t; W_i))^2\right] = E\left[(Z_{im}(t) - \psi_m(t; W_i))^2\right] = \mathfrak{R}(t, \psi_m)$$

for a fixed $\psi_m(t;\ w)$. This risk equivalence motivates the construction of an IPCW-type loss function. In particular, define for any suitable survivor function $G(\cdot|\cdot)$

$$L_{m,t}^{ipcw}(\mathcal{O}, \psi_m; G) = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L} I\{W_i \in \mathcal{N}_l\}\left[\frac{\Delta_i\{\tilde{Z}_{im}(t) - \beta_{lm}(t)\}^2}{G(\tilde{T}_i|W_i)}\right]; \tag{4}$$

then, it is easy to see that (4) is minimized by

$$\hat{\beta}_{lm}^{ipcw}(t; G) = \frac{\sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\}\frac{\Delta_i\tilde{Z}_{im}(t)}{G(\tilde{T}_i|W_i)}}{\sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\}\frac{\Delta_i}{G(\tilde{T}_i|W_i)}}, l = 1, \ldots, L, \tag{5}$$

implying that $\hat{\psi}_m(t; w) = \sum_{l=1}^{L} I\{w \in \mathcal{N}_l\}\hat{\beta}_{lm}^{ipcw}(t; G)$ is the corresponding estimator for the CIF at time t for cause m. Moreover, $L_{m,t}^{ipcw}(\mathcal{O}, \psi_m; G_0)$ is an unbiased estimate of $\mathfrak{R}(t, \psi_m)$. Observe that (4) and (5) respectively reduce to (2) and (3) if censoring is absent.

When $K = 1$, the loss (4) is just a special case of that considered in Molinaro et al. [13]; see also Lostritto et al. [14]. In practice, an estimator $\hat{G}(\cdot|\cdot)$ for $G_0(\cdot|\cdot)$ is used in (4); popular approaches here include product-limit estimators derived from the Kaplan–Meier and Cox regression estimation procedures. Of course, other methods could be used, such as regression trees or ensembles for right-censored survival data [e.g., 9, 10, 15].

As in [9], one can use semiparametric estimation theory for missing data to construct an improved estimator of the full data risk $\mathfrak{R}(t, \psi_m)$ by augmenting the IPCW loss function (4) with additional information on censored subjects. In particular, consider the loss function $L_{m,t}^{ipcw}(\mathcal{O}, \psi_m; G)$. Recall that Ψ_0 defines the set of CIFs of interest and let Ψ denote a corresponding model that may or may not contain Ψ_0. Define $V_{lm}(u; t, w, \Psi) = E_\Psi\left[(Z_m(t) - \beta_{lm}(t))^2|T \geq u, W = w\right]$ for any $t, u \geq 0$ and $w \in \mathcal{S}$; it is shown later how this expression specifically depends on Ψ. Then, fixing $\beta_{1m}(t), \ldots, \beta_{Lm}(t)$, the augmented estimator of $\mathfrak{R}(t, \psi_m)$ having the smallest variance that can be constructed from the unbiased estimator $L_{m,t}^{ipcw}(\mathcal{O}, \psi_m; G_0)$ is given by $L_{m,t}^{\mathrm{dr}}(\mathcal{O}, \psi_m; G_0, \Psi_0) = L_{m,t}^{ipcw}(\mathcal{O}, \psi_m; G_0) + L_{m,t}^{\mathrm{aug}}(\mathcal{O}, \psi_m; G_0, \Psi_0)$ where

$$L_{m,t}^{\mathrm{aug}}(\mathcal{O}, \psi_m; G, \Psi) = \frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\}\int_0^{\tilde{T}_i}\frac{V_{lm}(u; t, W_i, \Psi)}{G(u|W_i)}\mathrm{d}M_G(u|W_i) \tag{6}$$

is defined for suitable choices of Ψ, $G(\cdot|\cdot)$ and $M_G(t|w) = I(\tilde{T} \leq t, \Delta = 0) - \int_0^t I(\tilde{T} \geq u)d\Lambda_G(u|w)$, where $\Lambda_G(\cdot|\cdot)$ denotes the cumulative hazard function corresponding to the model $G(\cdot|\cdot)$ [cf. 16, Section 9.3 and 10.4]. The "doubly robust" loss $L_{m,t}^{\mathrm{dr}}(\mathcal{O}, \psi_m; G, \Psi)$ reduces to a special case of the class of loss functions proposed in Steingrimsson et al. [9] when $K = 1$.

The loss function $L_{m,t}^{\mathrm{dr}}(\mathcal{O}, \psi_m; G, \Psi)$ can be simplified further: because $Z_m(t)$ is binary,

$$V_{lm}(u; t, w, \Psi) = y_m(u; t, w, \Psi) - 2y_m(u; t, w, \Psi)\beta_{lm}(t) + \beta_{lm}^2(t) \tag{7}$$

for any suitable Ψ (e.g., Ψ_0), where $y_m(u;\ t, w, \Psi) = E_\Psi\{Z_m(t)|T \geq u, W = w\}$ reduces to

$$y_m(u; t, w, \Psi) = \begin{cases} \dfrac{P_\Psi(u \leq T \leq t, M = m|W = w)}{P_\Psi(T \geq u|W = w)} & \text{if} \quad u \leq t \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

The notation E_Ψ and P_Ψ means that these quantities are calculated under the CIF model specification Ψ. Hence, under a model Ψ, the calculation of $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ requires estimating both the CIF for cause m and the all-cause survival probability $P_\Psi(T \geq u | W = w)$.

Considering $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ as a function of the L scalar parameters $\beta_{1m}(t), \ldots, \beta_{Lm}(t)$ only and differentiating with respect to each one, it can be shown that

$$\tilde{\beta}^{\mathrm{dr}}_{lm}(t; G, \Psi) = \frac{\sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} \left[\widetilde{TS}^1_{1,im}(t) + \widetilde{TS}^1_{2,im}(t)\right]}{\sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} \left[\widetilde{TS}^0_{1,im} + \widetilde{TS}^0_{2,im}\right]}, \; l = 1, \ldots, L \tag{9}$$

minimize $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$, where

$$\widetilde{TS}^0_{1,im} = \frac{\Delta_i}{G(\tilde{T}_i | W_i)} \qquad \widetilde{TS}^0_{2,im} = \int_0^{\tilde{T}_i} \frac{\mathrm{d}M_G(u|W_i)}{G(u|W_i)}$$

$$\widetilde{TS}^1_{1,im}(t) = \frac{\tilde{Z}_{im}(t)\Delta_i}{G(\tilde{T}_i | W_i)} \qquad \widetilde{TS}^1_{2,im}(t) = \int_0^{\tilde{T}_i} \frac{y_m(u; t, W_i, \Psi)}{G(u|W_i)} \mathrm{d}M_G(u|W_i). \tag{10}$$

The validity of this result relies on the assumption that $G(\tilde{T}_i | W_i) \geq \epsilon > 0$ for some ϵ and each $i = 1, \ldots, n$. Under this same assumption, Lemma 1 of [17] implies

$$\widetilde{TS}^0_{1,im} + \widetilde{TS}^0_{2,im} = \frac{\Delta_i}{G(\tilde{T}_i | W_i)} + \frac{1 - \Delta_i}{G(\tilde{T}_i | W_i)} - \int_0^{\tilde{T}_i} \frac{\mathrm{d}\Lambda_G(u|W_i)}{G(u|W_i)} = 1;$$

letting $N_l = \sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\}$, $l = 1, \ldots, L$, it follows that (9) can be rewritten as

$$\hat{\beta}^{\mathrm{dr}}_{lm}(t; G, \Psi) = \frac{1}{N_l} \sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} \left[\widetilde{TS}^1_{1,im}(t) + \widetilde{TS}^1_{2,im}(t)\right], \; l = 1, \ldots, L. \tag{11}$$

Similarly to Section 2.3.1, $\hat{\psi}_m(t; w) = \sum_{l=1}^{L} I\{w \in \mathcal{N}_l\} \hat{\beta}^{\mathrm{dr}}_{lm}(t; G, \Psi)$ now generates the corresponding CIF estimate at time t for cause m and, in addition, $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ and (11) respectively reduce to (2) and (3) when censoring is absent.

The specification $G(t|w) = \tilde{G}(t|w) = 1$ for all $t \geq 0$ and $w \in \mathcal{S}$ generates an interesting special case of $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ despite $\tilde{G}(\cdot|\cdot)$ being incorrectly modeled in the presence of censoring. In particular, for suitable Ψ, (i) $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; \tilde{G}, \Psi) = \sum_{l=1}^{L} L^{bj}_{ml,t}(\mathcal{O}, \psi_m; \Psi)$ where

$$L^{bj}_{ml,t}(\mathcal{O}, \psi_m; \Psi) = \frac{1}{n} \sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} \left[\Delta_i \{\tilde{Z}_{im}(t) - \beta_{lm}(t)\}^2 + (1 - \Delta_i) V_{lm}(\tilde{T}_i; t, W_i, \Psi)\right];$$

and, (ii) for $\Psi = \Psi_0$, $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; \tilde{G}, \Psi_0)$ is an unbiased estimator of the risk $\mathfrak{R}(t, \psi_m)$. Noting that (7) implies $V_{lm}(\tilde{T}_i; t, W_i, \Psi)$ can be rewritten in terms of $y_m(\tilde{T}_i; t, w, \Psi)$ for every i, the minimizer of $L^{bj}_{ml,t}(\mathcal{O}, \psi_m; \Psi)$ is given by

$$\tilde{\beta}^{bj}_{lm}(t; \Psi) = \frac{1}{N_l} \sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} [\Delta_i \tilde{Z}_{im}(t) + (1 - \Delta_i) y_m(\tilde{T}_i; t, W_i, \Psi)].$$

That is, under the loss $L^{bj}_{ml,t}(\mathcal{O}, \psi_m; \Psi)$, the estimator for $\beta_{lm}(t)$ is the Buckley–James (BJ) estimator of the mean response within the partition $\mathcal{N}_l$ [18], an estimator that can also be derived directly from (11) by setting $G = \tilde{G}$. For this reason, we refer to $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; \tilde{G}, \Psi)$ as the Buckley–James loss function. For a fixed value of m and l, the function $\beta_{lm}(t)$ (i.e., the cumulative incidence for type m within node $\mathcal{N}_l$) is monotone increasing in t.

In contrast to the doubly robust loss, the Buckley–James loss function therefore preserves monotonicity; this property is useful when considering multiple time points, as considered in the next section.

2.3.2 Composite AIPCW loss functions: the case of multiple time points

Under the piecewise constant model (1), the quantity being estimated within each partition depends on t; however, the set of partitions remains the same across time. As a result, for a given m, it is reasonable to anticipate reduced variability when estimating $\psi_{0m}(t;\ w)$ by considering losses constructed from $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ that incorporate information across several time points.

Recall that $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi) = L^{ipcw}_{m,t}(\mathcal{O}, \psi_m; G) + L^{\mathrm{aug}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ where $L^{ipcw}_{m,t}(\mathcal{O}, \psi_m; G)$ is given by (4) and $L^{\mathrm{aug}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ is given by (6). For a given set of time points $0 < t_1 < t_2 < \cdots < t_J < \infty$, a simple composite loss function for a given event type m can be formed by calculating

$$L^{\mathrm{mult,dr}}_{m,\mathbf{t}}(\mathcal{O}, \psi_m; G, \Psi) = \sum_{j=1}^{J} \alpha_j L^{\mathrm{dr}}_{m,t_j}(\mathcal{O}, \psi_m; G, \Psi), \tag{12}$$

where $\alpha_j > 0, j = 1, \ldots, J$ are pre-specified weights such that $\sum_{j=1}^{J} \alpha_j < \infty$. We obtain

$$\tilde{\beta}^{\mathrm{mult,dr}}_{lm}(t_j; G, \Psi) = \frac{1}{N_l} \sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} \left[\widetilde{TS}^1_{1,im}(t_j) + \widetilde{TS}^1_{2,im}(t_j)\right], \tag{13}$$

as the minimizers of (12) with respect to $\beta_{lm}(t_j)$ for $j = 1, \ldots, J;\ l = 1, \ldots, L$. In the absence of censoring, the indicated composite loss function and partition-specific estimators reduce to that which would be computed by extending the loss function introduced in Section 2.2 in the manner described above.

One question that might be asked at this stage is how one should choose $0 < t_1 < t_2 < \cdots < t_J < \infty$. In many biomedical applications, there will be one or more specific time points of interest; in this case, the choice of $t_1, \ldots, t_J$ is reasonably clear. In the absence of interest in specific time points, there is less formal guidance available. In part, this is because the loss function (12) is well-defined, and can be used, in its own right for any given sequence of time points. For example, one could simply choose $0 < t_1 < t_2 < \cdots < t_J < \infty$ according to specified quantiles of the marginal (i.e., all-cause) distribution of T and specify weights to equally or unequally weight loss contributions across time. Using equal weights, for example, limited simulation experiments (not shown) have shown that using $J = 3, 4$, or 5 time points spread across quantiles that are not too extreme generally results in very similar estimators. A more objective perspective for choosing $t_1, \ldots, t_J$ would instead be to regard the weighted composite loss as an approximation to the integrated Brier score on $[0, t_J]$ [e.g., 19]. For example, taking $0 < t_1 < t_2 < \cdots < t_J < \infty$ to be equally spaced with grid spacing $h > 0$, one can set $\alpha_j = h$ to obtain a trapezoidal rule approximation to the integrated Brier score on $[0, t_J]$; more generally, one can take $0 < t_1 < t_2 < \cdots < t_J < \infty$ to be unequally spaced and obtain a different trapezoidal rule approximation by selecting the α_js accordingly. Under certain smoothness assumptions on the Brier score, it is then possible to select J to approximate the integrated Brier score on an finite interval to within a certain degree of error [e.g., 20]. However, in view of comments made earlier on the sensitivity of the results to the choice of J, the use of highly accurate integral approximations seems unnecessary.

Thus far, we have assumed the existence of a fixed partition $\{\mathcal{N}_1, \ldots, \mathcal{N}_L\}$ of $\mathcal{S}$. In this situation, and regardless of how $0 < t_1 < t_2 < \cdots < t_J < \infty$ are chosen, the use of a composite loss like (13) actually yields no extra efficiency gain for estimating the CIF for cause m within each partition. This can be seen from (13), which is exactly equal to (11) computed for $t = t_j$; that is, the partition-specific estimators for $t_1, \ldots, t_J$ do not depend on $\alpha_1, \ldots, \alpha_J$. We stress here that this is a direct consequence of the absence of further modeling assumptions that restrict the relationship between $\beta_{lm}(t)$ (i.e., the CIF when $W \in \mathcal{N}_l$) and $\beta_{l'm}(t)$ (i.e., the CIF when $W \in \mathcal{N}_{l'}$) when $l \neq l'$.

However, in the case of regression trees, and by extension ensembles of trees (e.g., RF), the partition $\{\mathcal{N}_1, \ldots, N_L\}$ for every tree is estimated adaptively from the data. In this situation, the use of a weighted composite loss (13) directly influences both the selection of L and the chosen partition boundaries, hence results in some implicit dependence between $\beta_{lm}(t)$ and $\beta_{l'm}(t)$ for $l \neq l'$. Consequently, performance gains may still be expected when estimating $\psi_{m0}(t_j;\ w), j = 1, \ldots, J$ using a composite loss function whether one uses trees or ensembles of trees. We consider such methods further in the next section.

3 CIF regression trees and ensembles

The developments in Section 2 provide an important building block for developing new splitting and evaluation procedures when using CART to build regression trees for estimating the CIF, with or without loss to follow-up. Because RF relies on bootstrapped ensembles of CART trees, the loss-based estimation procedures have similarly important implications for RF. In the coming sections, we propose several variants on CART and RF for competing risks data that use the loss functions introduced in previous section.

3.1 Estimating a CIF via CART or RF: no loss to follow-up

Given a specified loss function, CART [21] fits a regression tree as follows:

(1) Using recursive binary partitioning, grow a maximal tree by selecting a (covariate, cutpoint) combination at every stage that minimizes the chosen loss function;
(2) Using cross-validation, select the best tree from the sequence of candidate trees generated by Step 1 via cost complexity pruning (i.e., using penalized loss).

In its most commonly used form for regression problems with a continuous outcome, CART estimates the conditional mean response as a piecewise constant function on $\mathcal{S}$, making all decisions on the basis of minimizing squared error loss. The resulting tree-structured regression function estimates the predicted response within each terminal node (i.e., partition of $\mathcal{S}$) using the sample mean of the observations falling into that node. The set of terminal nodes (i.e., the partition structure) is determined adaptively from the data as a result of steps 1 and 2 above.

The random forests algorithm [RF; 3] is an ensemble-based extension of CART:

(1) Bootstrap the data; that is, draw B random samples with replacement from $\mathcal{F}$.
(2) For each bootstrapped dataset, run Step 1 of the CART algorithm above, possibly randomly selecting a set of $p^* \leq p$ candidate covariates when determining a (covariate, cutpoint) combination at each possible splitting stage.
(3) Compute the terminal node estimators for each subject for each of the B trees and average these to obtain an ensemble predictor.

The critical step that underpins both CART and RF is Step 1 of the CART algorithm, where connections to the developments of Section 2 should now be evident. In particular, in the absence of censoring and under the piecewise constant model (1) for $\psi_{0m}(t;\ w)$, Section 2.2 shows that a nonparametric estimate for $\psi_{0m}(t;\ w)$ at t can be obtained by minimizing the loss (2). This basic estimation problem is equivalent to estimating the conditional mean response using the modified dataset $\mathcal{F}_{red,t} = \left\{(Z_{im}(t), W_i')', i = 1, \ldots n\right\}$ by minimizing the squared error loss (2). Therefore, any implementation of CART or RF for squared error loss applied to $\mathcal{F}_{red,t}$ will produce a corresponding CART- or RF-based estimate of $\psi_{0m}(t;\ w)$. For example, CART estimates L and the associated set of terminal nodes $\{\mathcal{N}_1, \ldots, \mathcal{N}_L\}$ from the data $\mathcal{F}_{\text{red},t}$, and within each terminal node, estimates $\psi_{0m}(t;\ w)$ by (3).

For the case of multiple time points, a composite loss function analogous to (indeed, a special case of) (12) is given by

$$L_{m,\mathbf{t}}^{\text{emp}}(\mathcal{F}, \psi_m) = \sum_{l=1}^{L} \sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} \left(\sum_{j=1}^{J} \frac{\alpha_j}{n} \{Z_{im}(t_j) - \beta_{lm}(t_j)\}^2 \right)$$
$$= \sum_{l=1}^{L} \sum_{i=1}^{n} I\{W_i \in \mathcal{N}_l\} \{\mathbf{Z}_{im} - \boldsymbol{\beta}_{lm}\}^\top \mathbf{D}^{-1} \{\mathbf{Z}_{im} - \boldsymbol{\beta}_{lm}\} \tag{14}$$

where $\mathbf{Z}_{im} = (Z_{im}(t_1), \ldots, Z_{im}(t_J))^\top$, $\boldsymbol{\beta}_{lm} = (\beta_{lm}(t_1), \ldots, \beta_{lm}(t_J))^\top$, and $\mathbf{D}$ is a diagonal matrix with $D_{jj} = n\alpha_j^{-1}$, $j = 1, \ldots, J$. One can therefore estimate the desired CIF either by a tree or random forest directly from the data $\left\{ \left(\mathbf{Z}_{im}, W_i'\right)', i = 1, \ldots n \right\}$ using the `MultivariateRandomForest` package [22], which builds regression trees using a Mahalanobis loss function of the form (14); see also [23]. The `randomForestSRC` package also accommodates multivariate response data, but currently uses loss functions in the case of squared-error loss that involve repeatedly (i.e., dynamically) standardizing the outcomes when determining where and when to split. Since this process of repeated standardization has implications for certain equivalences on which we later rely, we use the `MultivariateRandomForest` package to implement our methods in the next subsection.

3.2 Estimating a CIF via CART or RF: loss to follow-up

The CART and RF algorithms as outlined in the previous subsection extend easily to more general loss functions, where decisions and predictions are instead derived from minimizing the chosen loss function. In particular, the loss $L_{m,t}^{\text{dr}}(\mathcal{O}, \psi_m; G, \Psi)$ or its composite extension $L_{m,\mathbf{t}}^{\text{mult,dr}}(\mathcal{O}, \psi_m; G, \Psi)$ could be used in place of (2) in either algorithm in the presence of censoring. A detailed description of such an algorithm in the case of CART (i.e., for building regression trees) can be found in [24], along with extensive simulation studies evaluating the performance of several different loss functions derived from (12). There, it was found that the use of the doubly robust and Buckley–James forms of the loss function, with the augmentation term parameter Ψ estimated using the methods of [4], provided the second best performance overall, being inferior only to the "oracle" setting where a correctly specified family of parametric models is chosen for Ψ.

Although the algorithms extend quite easily to more general loss functions, existing software may not be able to easily accommodate these extensions. As will be shown below, however, algorithms that use the loss function $L_{m,t}^{\text{dr}}(\mathcal{O}, \psi_m; G, \Psi)$ or $L_{m,\mathbf{t}}^{\text{mult,dr}}(\mathcal{O}, \psi_m; G, \Psi)$ can be implemented easily with existing software using a certain form of response imputation; see [10] for related results in the case where $K = 1$.

Recall that $L_{m,t}^{\text{dr}}(\mathcal{O}, \psi_m; G, \Psi) = L_{m,t}^{ipcw}(\mathcal{O}, \psi_m; G) + L_{m,t}^{\text{aug}}(\mathcal{O}, \psi_m; G, \Psi)$ where $L_{m,t}^{ipcw}(\mathcal{O}, \psi_m; G)$ is given by (4) and $L_{m,t}^{\text{aug}}(\mathcal{O}, \psi_m; G, \Psi)$ is given by (6). Using the results in (7) and (8) and notation defined in (10), calculations similar to [10] show that

$$L_{m,t}^{\text{dr}}(\mathcal{O}, \psi_m; G, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} I(W_i \in \mathcal{N}_l) \left[\beta_{ml}^2(t) - 2H_m^{\text{dr}}(t; O_i, G, \Psi)\beta_{ml}(t) + H_m^{\text{dr}}(t; O_i, G, \Psi) \right]$$

where $H_m^{\text{dr}}(t; O_i, G, \Psi) = \widetilde{TS}_{1,im}^1(t) + \widetilde{TS}_{2,im}^1(t)$. Define the modified "imputed" loss function

$$L_{m,t}^{\text{dr},*}(\mathcal{O}, \psi_m; G, \Psi) = \sum_{i=1}^{n} \sum_{l=1}^{L} I(W_i \in \mathcal{N}_l)(H_m^{\text{dr}}(t; O_i, G, \Psi) - \beta_{ml}(t))^2$$
$$= \sum_{i=1}^{n} \sum_{l=1}^{L} I(W_i \in \mathcal{N}_l) \left\{ \beta_{ml}^2(t) - 2H_m^{\text{dr}}(t; O_i, G, \Psi)\beta_{ml}(t) + \left[H_m^{\text{dr}}(t; O_i, G, \Psi)\right]^2 \right\}.$$

Importantly, observe $L_{m,t}^{\text{dr}}(\mathcal{O}, \psi_m; G, \Psi) - L_{m,t}^{\text{dr},*}(\mathcal{O}, \psi_m; G, \Psi)$ does not depend on $\beta_{ml}(t), l = 1, \ldots, L$ when $H_m^{\text{dr}}(t; O_i, G, \Psi), i = 1, \ldots, n$ does not depend on these terms. Hence, a CART tree or RF built using $L_{m,t}^{\text{dr}}(\mathcal{O}, \psi_m; G, \Psi)$ will be identical to that built using $L_{m,t}^{\text{dr},*}(\mathcal{O}, \psi_m; G, \Psi)$ [see also 10, Thm. 4.1].

A similar correspondence can be established between the composite loss $L_{m,\mathbf{t}}^{\text{mult,dr}}(\mathcal{O}, \psi_m; G, \Psi)$ in (12) and the Mahalanobis-type loss function

$$L_{m,\mathbf{t}}^{\text{mult,dr},*}(\mathcal{O}, \psi_m; G, \Psi) = \sum_{l=1}^{L} \sum_{i=1}^{n} I(W_i \in \mathcal{N}_l) \left(\mathbf{H}_{im}^{\text{dr}} - \boldsymbol{\beta}_{lm}\right)^{\top} \mathbf{D}^{-1} \left(\mathbf{H}_{im}^{\text{dr}} - \boldsymbol{\beta}_{lm}\right), \tag{15}$$

where $\mathbf{H}_{im}^{\text{dr}} = (H_m^{\text{dr}}(t_1; O_i, G, \Psi), \ldots, H_m^{\text{dr}}(t_J; O_i, G, \Psi))^{\top}$, $j = 1, \ldots, J$; compare with (14). Hence, a CART tree or RF built using (12) is identical to that built using (15).

Critically, these results imply that a CART or RF algorithm that uses the loss function $L_{m,\mathbf{t}}^{\text{mult,dr}}(\mathcal{O}, \psi_m; G, \Psi)$ can be implemented by applying a version of that algorithm designed for squared error loss to the modified dataset $\left(\mathbf{H}_{im}^{\text{dr}}, W_i\right), = 1 \ldots, n$, where $\mathbf{H}_{im}^{\text{dr}}$ is an imputed univariate ($J = 1$) or multivariate ($J > 1$) response. This includes the case of Buckley–James loss, which results as a special case upon setting $G(t|w) = 1$ for all $t > 0$ and $w \in \mathcal{S}$. Specifically, for a fixed set of times $0 < t_1 < \cdots < t_J < \infty$ and event type m, the relevant RF estimation algorithm is as follows:

Algorithm M_0:

1. Compute $\hat{G}$ and $\hat{\Psi}$ by appropriate modeling;
2. Compute $H_m^{\text{dr}}(t_j; O_i, \hat{G}, \hat{\Psi})$ for $i = 1, \ldots, n$, $j = 1, \ldots, J$;
3. Run `MultivariateRandomForest` on the imputed dataset $\left(\hat{\mathbf{H}}_{im}^{\text{dr}}, W_i\right), = 1 \ldots, n$, where

$$\hat{\mathbf{H}}_{im}^{\text{dr}} = (H_m^{\text{dr}}(t_1; O_i, \hat{G}, \hat{\Psi}), \ldots, H_m^{\text{dr}}(t_J; O_i, \hat{G}, \hat{\Psi}))^{\top}, \ j = 1, \ldots, J.$$

The above procedure extends in an obvious way to other tree- and forest-based algorithms that make all decisions on the basis of minimizing squared error loss.

3.2.1 A modified imputation approach for doubly robust losses

Recall that $\mathbf{H}_{im}^{\text{dr}} = (H_m^{\text{dr}}(t_1; O_i, G, \Psi), \ldots, H_m^{\text{dr}}(t_J; O_i, G, \Psi))^{\top}$ where $H_m^{\text{dr}}(t; O_i, G, \Psi) = \widetilde{TS}_{1,im}^1(t) + \widetilde{TS}_{2,im}^1(t)$ and $\widetilde{TS}_{r,im}^1(t), r = 1, 2$ are defined as (10). Observing that

$$\widetilde{TS}_{2,im}^1(t) = (1 - \Delta_i) \frac{y_m(\tilde{T}_i; t, W_i, \Psi)}{G(\tilde{T}_i|W_i)} - \int_0^{\tilde{T}_i} \frac{y_m(u; t, W_i, \Psi)}{G^2(u|W_i)} Y_i(u) \mathrm{d}\bar{G}(u|W_i),$$

it can be seen that this term has the potential to be estimated with undesirably high variability due to the presence of $G^2(\cdot|W_i)$ in the denominator of the second term.

In the context of devising testing procedures for the Cox regression model, Lin, Wei and Ying [25] proposed approximating certain martingale integrals using a simple but effective simulation technique. The basic idea, applied here, involves replacing $M_G(u|W_i)$ by $M_G^*(u) = \xi_i I(\tilde{T}_i \le u, \Delta_i = 0)$, where $\xi \sim N(0, 1)$. In particular, suppose that $\widetilde{TS}_{2,im}^1(t)$ in the term $H_m^{\text{dr}}(t; O_i, G, \Psi)$ is replaced by

$$\widetilde{TS}_{2,im}^1(t, \xi_i) = \int_0^{\tilde{T}_i} \frac{y_m(u; t, W_i, \Psi)}{G(u|W_i)} \mathrm{d}M_G^*(u) = \xi_i (1 - \Delta_i) \frac{y_m(\tilde{T}_i; t, W_i, \Psi)}{G(\tilde{T}_i|W_i)}.$$

Defining $H_m^{\text{dr}}(t, \xi_i; O_i, G, \Psi) = \widetilde{TS}_{1,im}^1(t) + \widetilde{TS}_{2,im}^1(t, \xi_i)$, we obtain the alternative loss function

$$L_{m,t}^{\text{dr},*}(\mathcal{O}, \psi_m, \xi; G, \Psi) = \sum_{i=1}^{n} \sum_{l=1}^{L} I(W_i \in \mathcal{N}_l)(H_m^{\text{dr}}(t, \xi_i; O_i, G, \Psi) - \beta_{ml}(t))^2.$$

A straightforward conditioning argument shows that each of $E\left[L_{m,t}^{\text{dr}}(\mathcal{O}, \psi_m; G, \Psi)\right]$, $E\left[L_{m,t}^{\text{dr},*}(\mathcal{O}, \psi_m; G, \Psi)\right]$ and $E\left[L_{m,t}^{\text{dr},*}(\mathcal{O}, \psi_m, \xi; G, \Psi)\right]$ have the same minimizers; however, a CART tree or RF built using $L_{m,t}^{\text{dr},*}(\mathcal{O}, \psi_m, \xi; G, \Psi)$

is no longer guaranteed to be identical to that built using either $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ or $L^{\mathrm{dr},*}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ because $L^{\mathrm{dr},*}_{m,t}(\mathcal{O}, \psi_m, \xi; G, \Psi)$ contains an extra mean zero term involving $\beta_{ml}(t)$.

Define the vector $\mathbf{H}^{\mathrm{dr}}_{im}(\xi_i) = (H^{\mathrm{dr}}_m(t_1, \xi_i; O_i, G, \Psi), \ldots, H^{\mathrm{dr}}_m(t_J, \xi_i; O_i, G, \Psi))^\top$. Then, for a fixed set of times $0 < t_1 < \cdots < t_J < \infty$ and event type m, we obtain a modified version of the algorithm presented in Section 3.2:

Algorithm M_1:

(1) Compute $\hat{G}$ and $\hat{\Psi}$ by appropriate modeling;
(2) Loop over $r = 1, \ldots, R$, where $R \geq 1$ is set by the user:
   (a) Generate $\xi_i^{(r)} \sim N(0, 1)$ for $i = 1, \ldots, n$.
   (b) Compute $H^{\mathrm{dr}}_m(t_j, \xi_i^{(r)}; O_i, \hat{G}, \hat{\Psi})$ for $i = 1, \ldots, n$, $j = 1, \ldots, J$.
   (c) Run `MultivariateRandomForest` on the modified imputed dataset $(\hat{\mathbf{H}}^{\mathrm{dr}}_{im}(\xi_i^{(r)}), W_i), = 1, \ldots, n$, where $\hat{\mathbf{H}}^{\mathrm{dr}}_{im}(\xi_i^{(r)}) = (H^{\mathrm{dr}}_m(t_1, \xi_i^{(r)}; O_i, \hat{G}, \hat{\Psi}), \ldots, H^{\mathrm{dr}}_m(t_J, \xi_i^{(r)}; O_i, \hat{G}, \hat{\Psi}))^\top$, $j = 1, \ldots, J$.
   (d) Record rth result.
(3) Average the R ensemble estimates to obtain a final ensemble predictor.

Analogously to Algorithm M_0, Step 2(c) of the above algorithm involves bootstrapping the rth modified version of the input dataset B times to obtain a RF predictor for the rth modified dataset. Step 3 then averages these R different RF estimates to produce a single ensemble predictor. As a computationally efficient version of this algorithm, Step 2(c) could be run with $B = 1$ only; that is, instead of generating a full RF at this stage, one builds a single tree using random feature selection without pruning. The resulting algorithm then reduces to a RF-type algorithm based on R bootstrap samples, but where there is an extra component of randomization used in the generation of each bootstrap sample.

3.3 Some further remarks

We now provide some high-level comparisons between the proposed methods and the approaches of [4, 5] to building CIF ensemble estimators. Our proposed methods share greater similarity with the methods proposed in [5], as will be discussed below; hence, we consider that work first.

In [5], each censored outcome is replaced with a jackknife pseudovalue derived from the marginal Aalen-Johansen estimator [8]. These authors then build a random forest by bootstrapping these pseudovalues (i.e., computed for a specific cause of failure m at particular follow-up time t, treating each as a fully observed response) and constructing an ensemble estimator from the bootstrapped regression trees. In the presence of censoring, their specific proposal amounts to running RF with squared error loss on a set of imputed response variables (i.e., pseudovalues). The justification for this approach is that the ith jackknife pseudovalue reduces to $Z_{im}(t)$ if the ith response is uncensored and it gives an approximately unbiased estimate of the marginal CIF for cause m at time t if the ith observation is censored. Hence, averaging these values in a terminal node in a given tree should result in a sensible estimator of the CIF within any given terminal node provided that censoring is independent of failure, cause of failure and all covariates.

Similarly to [5], the methods described in [4] also make direct use of the Aalen-Johansen estimator [8]. Specifically, for each nonparametric bootstrap sample drawn from the original training dataset, an unpruned regression tree is built using logrank-type splitting rules appropriate for competing risks [e.g., 7]. One can then respectively form the Aalen-Johansen estimate using all observations that fall into each terminal node, and average those estimates over all bootstrap samples to obtain an estimated CIF. The approach taken in [4] to building the individual trees that make up the ensemble predictor involves maximizing heterogeneity between terminal nodes (i.e., "goodness of split") using a statistic specifically designed for testing for differences between CIFs; this differs from RF, which in standard form intends to maximize homogeneity within nodes through minimizing a specified loss function (e.g., squared error).

The methods described in [4] differ from those proposed here in several ways. Most importantly, the trees that make up the ensemble are built using a "goodness of split," not loss-based, criterion function. In addition, response imputation is not used (or needed); as such, an ensemble estimate is constructed from

averages of Aalen–Johansen estimates of the CIF within each terminal node (hence a function of time) rather than from the mean of imputed response variables. Focusing on the specific case of regression tree algorithms that use squared error loss, the most important differences between the approach taken in [5] and our proposed methods stem from (i) our adherence to a principled loss-based approach to estimation that directly generalizes that used by CART and RF for uncensored data; (ii) a rigorous justification for implementation when using the imputed responses defined in Section 3.2; and (iii) the actual form of the imputed responses. More specifically, in the setting where the loss function $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ is constructed using a single time point t, the corresponding equivalent squared error loss $L^{\mathrm{dr},*}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ utilizes imputed responses of the form $H^{\mathrm{dr}}_m(t; O_i, G, \Psi) = \widetilde{TS}^1_{1,im}(t) + \widetilde{TS}^1_{2,im}(t)$ for $i = 1, \ldots, n$. These imputed responses differ from, but may be considered analogous to, the jackknife pseudovalues of [5]. Indeed, similarly to [5], $H^{\mathrm{dr}}_m(t; O_i, G, \Psi)$ reduces to $Z_{im}(t)$ if the ith response is uncensored and is an unbiased estimate of the conditional CIF for cause m at time t (i.e., for a correct censoring distribution G). In contrast to [5], our approach only requires censoring to be independent of failure and cause of failure conditional on covariates. Use of the doubly robust loss is also robust to misspecification of one of G and Ψ (but not both). In addition, using the jackknife pseudovalues of [5] as the observed responses in a RF algorithm that employs squared error loss loses the direct connection being exploited in Section 3.2, that is, the equivalence between the imputed loss $L^{\mathrm{dr},*}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ (i.e., squared error loss derived from the imputed responses $H^{\mathrm{dr}}_m(t; O_i, G, \Psi), i = 1, \ldots, n$) and the observed data loss $L^{\mathrm{dr}}_{m,t}(\mathcal{O}, \psi_m; G, \Psi)$ (i.e., an approximately unbiased estimator of the desired full data risk function).

The methodology proposed in [5] does not consider the possibility of estimating the CIF at multiple time points, though their estimator can be extended to this setting in a manner similar to that done here. The methodology proposed in [4] does not require pre-specification of $t_1, \ldots, t_J$ and can be used to estimate the CIF at any time point, hence the CIF curve, within the range of the observed data. In this regard, the need to specify a grid of times is a disadvantage of methods that average terminal node estimates constructed directly from imputed time-specific outcomes. In the case where the full CIF curve is desired, one easy solution is to generate the desired CIF estimate on a modest number of time points (e.g., $J \in \{5, \ldots, 10\}$) and then construct a curve using monotone interpolation. In settings where software provides easy access to each individual tree that makes up the ensemble estimate of the CIF, alternative approaches are possible that use terminal node estimates (e.g., Aalen–Johansen estimator) that differ from the natural mean estimate induced by the composite loss.

4 Simulation study: CIF estimation via RF

4.1 Main simulation setting

In this section, we will evaluate the performance of estimators derived using Algorithms M_0 and M_1 and compare the prediction errors to the RF procedure for CIF estimation proposed by [4], which is implemented in the R package `randomForestSRC`. The methods of [5] are not considered here due to the absence of an explicit proposal for handling multiple time points.

Let $W_i \sim N(0,1), i = 1, \ldots, 20$ be independent predictor variables. Define the true CIFs $\psi_{0m}(t;\, \boldsymbol{W}), m = 1, 2$ as follows:

$$\psi_{01}(t; \boldsymbol{W}) = 1 - (1 - p(1 - e^{-t}))^{\exp(\boldsymbol{\beta}_1^T \boldsymbol{Z}(\boldsymbol{W}))} \tag{16}$$

$$\psi_{02}(t; \boldsymbol{W}) = (1-p)^{\exp(\boldsymbol{\beta}_1^T \boldsymbol{Z}(\boldsymbol{W}))} \times (1 - \exp(-t\, \exp(\boldsymbol{\beta}_2^T \boldsymbol{Z}(\boldsymbol{W})))), \tag{17}$$

where $\boldsymbol{Z}(\boldsymbol{W}) = (\sin(\pi W_1 W_2), W_3^2, W_{10}, I(W_{11} > 0), W_{12}, \exp(W_{15}))$ and the regression coefficients are given by $\boldsymbol{\beta}_1 = (0.5, 0.5, 0.5, 0.5, 0.6, -0.3)^\top$ and $\boldsymbol{\beta}_2 = (0, -0.5, -0.5, -0.5, 0.5, 0.1)^\top$. Random censoring is generated from log normal distribution with mean $0.1 + 0.1 \cdot |W_1 + W_3 + W_5| + 0.1 \cdot |W_{11} + W_{13} + W_{15}|$ and variance 1. In this setting, the overall censoring rate is approximately 28.1%.

4.2 Simulation results

4.2.1 Algorithms to be compared

We focus on estimation of the CIF at the 25th, 50th and 75th time points of the marginal failure time distribution T; these are approximated outside the main simulation using a single, very large random sample. These time points are also used in the computation of all composite loss functions. CIF estimates are obtained using both Algorithms M_0 (Section 3.2) and M_1 (Section 3.2.1), and compared to those produced by `rfsrc` in the `randomForestSRC` package [6]. For Algorithm M_1, we set $R = 1$ and $B = 500$; for each simulated dataset; this corresponds to generating a single set of n independent standard normal random variables to be used in the computation of the modified imputed loss $L^{\mathrm{dr},*}_{m,t}(\mathcal{O}, \psi_m, \xi; G, \Psi)$, and then using 500 bootstrap samples to generate a RF predictor.

For calculating $y_m(u;\ t, w, \Psi)$ in (8), we estimate Ψ using `rfsrc`. We denote the resulting Buckley–James (*BJ-RF*) and doubly robust (*DR-RF*) transformations $H^{bj*}_m(t, O_i; \hat{\Psi})$ and $H^{\mathrm{dr}*}_m(t, O_i; \hat{G}, \hat{\Psi}), i = 1, \ldots, n$, where the censoring distribution estimate $\hat{G}$ is obtained using the methods of [26]. Specifically, $\hat{G}$ is estimated using the `rpart` package [27] with the minimum number of observations in each node (i.e., `minbucket`) set to 30. For comparison, we also compute (a) versions of these same estimators using correctly specified parametric models derived directly from (16), with relevant parameters estimated using the maximum likelihood approach detailed in Jeong and Fine [28]; these results for the parametric Fine-Gray-type model for the CIF are denoted *BJ-FG (true)* and *DR-FG (true)*, respectively, and, (b) the RF estimator of the CIF obtained using `rfsrc`.

Tuning parameters play an important role in the performance of ensemble estimators. In the case of `rfsrc`, two key tuning parameters are (i) the minimum number of observations in each terminal node (*nodesize*) and (ii) the number of candidate variables selected for consideration at each split (*mtry*). There are identical parameters with different names to be selected for use with the package `MultivariateRandomForest`, specifically through the use of the required function `build_single_tree`; for simplicity, we present and summarize results using the labels *nodesize* and *mtry*. For each algorithm, we calculate the relevant ensemble estimators by setting these tuning parameters as follows:

- Tuning Set 1: *nodesize* $= 20$ and *mtry* $= \lfloor \sqrt{p} \rfloor$.
- Tuning Set 2: *nodesize* and *mtry* are selected to minimize the out-of-bag (OOB) error.

Results with the suffix *-opt* correspond to parameters selected under Tuning Set 2. Hence, results are reported for 4 cases:

(i) Fixed *mtry* and *nodesize* with Algorithm M_0;
(ii) Optimized *mtry* and *nodesize* with Algorithm M_0;
(iii) Fixed *mtry* and *nodesize* with Algorithm M_1;
(iv) Optimized *mtry* and *nodesize* with Algorithm M_1.

4.2.2 Summary of results

For all simulation settings, results are obtained for 400 independent (estimation, test) dataset pairs. The estimation dataset is generated as in Section 4.1, with $n = 250$; an independent test dataset of size $n_{\text{test}} = 2000$ consisting only of the covariates is generated similarly. For all simulation settings, we compute the mean square error

$$\frac{1}{n_{\text{test}}} \sum_{r=1}^{n_{\text{test}}} \{\hat{\psi}_m(t|\boldsymbol{W}_r) - \psi_{0m}(t|\boldsymbol{W}_r)\}^2$$

to compare the performance of different algorithms at different values of t. Figures 1–4 show the results from the simulation setting with forests with multiple time points and comparisons with results obtained using `rfsrc` using the same approaches to selecting *mtry* and *nodesize*. Figures 1 and 2 focus on $m = 1$ and respectively compare the results for Algorithms M_0 and M_1 with fixed and optimized tuning parameters;

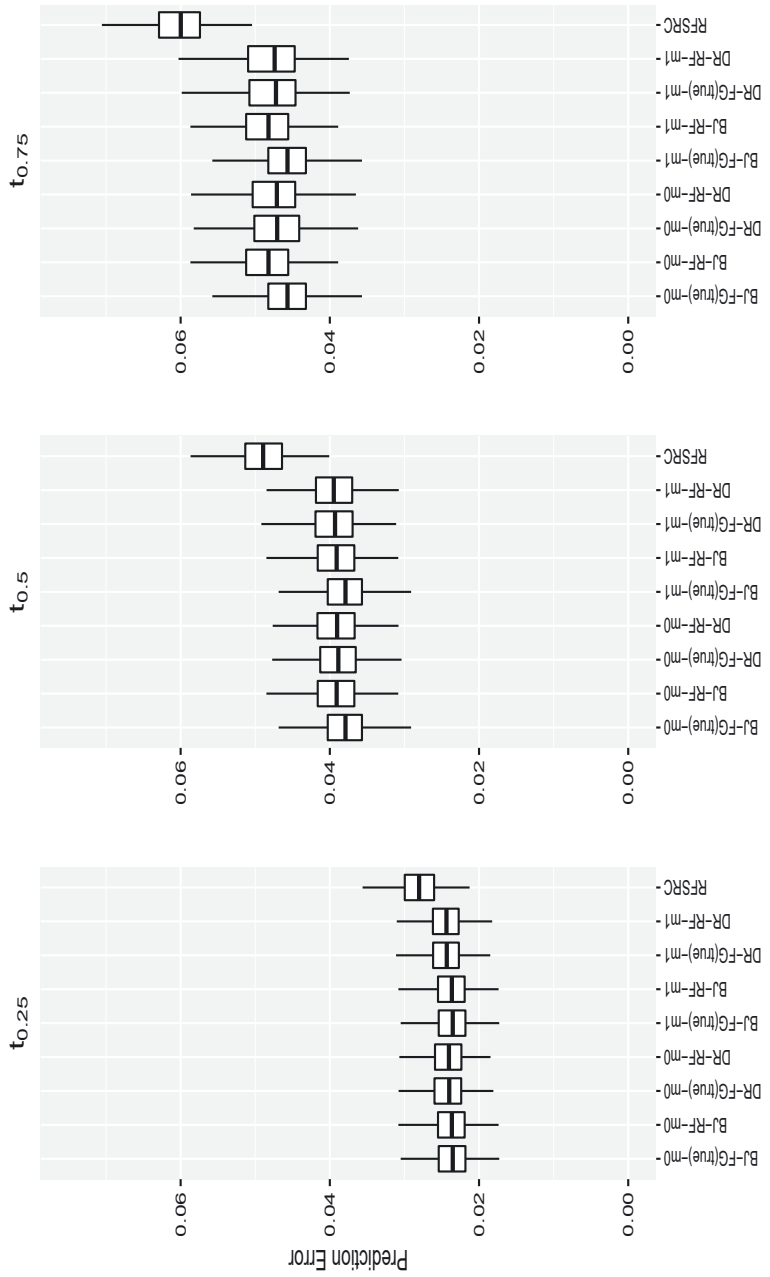

Figure 1: Comparing algorithms M_0 and M_1 with fixed tuning parameters for event $m = 1$. Results for rfsrc are labeled RFSRC. Results with Buckley–James and doubly robust losses are respectively prefaced by BJ- and DR-; the methods for $\hat{\Psi}$ as required by algorithms M_0 and M_1 are respectively denoted by RF- and FG(true)-; and, the use of the imputation algorithms M_0 and M_1 in generating the ensemble estimator is respectively denoted by *m*0 and *m*1.

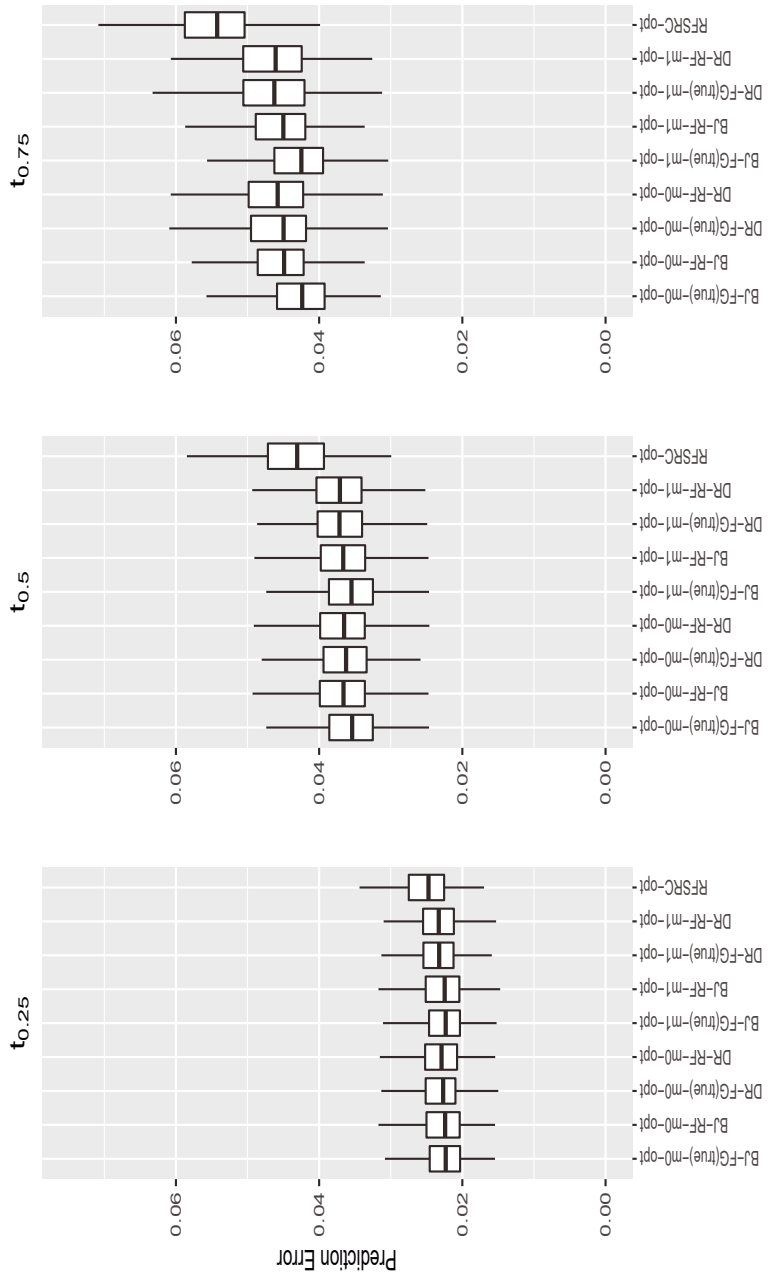

Figure 2: Comparing algorithms M_0 and M_1 with optimized tuning parameters for event $m = 1$. Results with Buckley–James and doubly robust losses are respectively prefaced by BJ- and DR-; the methods for $\hat{\Psi}$ as required by algorithms M_0 and M_1 are respectively denoted by RF- and FG(true)-; and, the use of the imputation algorithms M_0 and M_1 in generating the ensemble estimator is respectively denoted by *m0*– opt and *m1*– opt, the opt indicating use of optimized tuning parameters.

Figure 3: Comparing algorithms M_0 and M_1 with fixed tuning parameters for event $m = 2$. Results for rfsrc are labeled RFSRC. Results with Buckley–James and doubly robust losses are respectively prefaced by BJ- and DR-; the methods for $\hat{\Psi}$ as required by algorithms M_0 and M_1 are respectively denoted by RF- and FG(true)-; and, the use of the imputation algorithms M_0 and M_1 in generating the ensemble estimator is respectively denoted by *m*0 and *m*1.

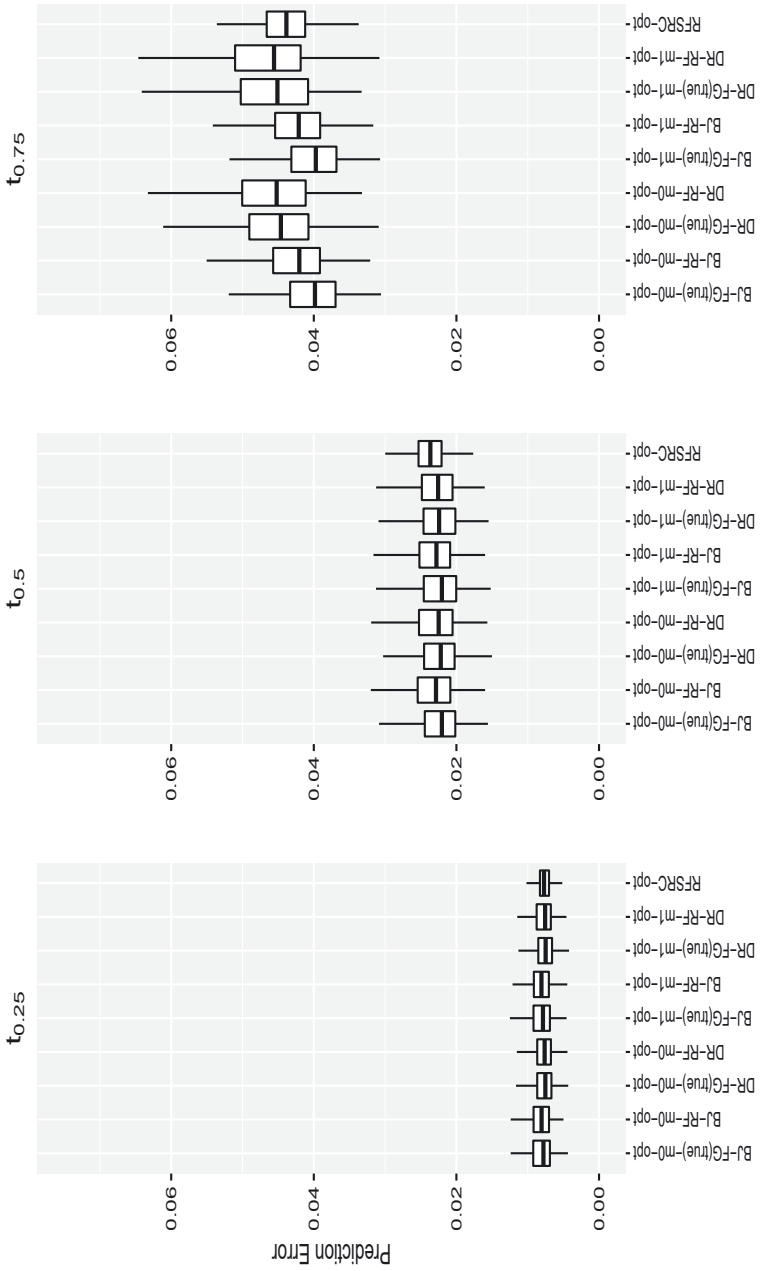

Figure 4: Comparing algorithms M_0 and M_1 with optimized tuning parameters for event $m = 2$. Results with Buckley–James and doubly robust losses are respectively prefaced by BJ- and DR-; the methods for $\hat{\Psi}$ as required by algorithms M_0 and M_1 are respectively denoted by RF- and FG(true)-; and, the use of the imputation algorithms M_0 and M_1 in generating the ensemble estimator is respectively denoted by *m*0— opt and *m*1— opt, the opt indicating use of optimized tuning parameters.

Figures A.1 and A.2 in Section A.1.1 of the Supplementary Materials respectively compares the results with fixed and optimized tuning parameters for each algorithm. Figures 3, 4, A.3 and A.4 repeat these results for $m = 2$. We give an overall summary of these results below:

- Algorithms M_0 and M_1 exhibit similar performance for the 25th and 50th percentile time points; however, Algorithm M_1 tends to perform better for the 75th percentile, where the impact of the censoring rate is higher.
- With optimization of tuning parameters, all methods demonstrate similar or slightly improved performance than the same approach using fixed choices of *nodesize* and *mtry*, at least for the 25th and 50th percentile. For the 75th percentile, the effects are somewhat less evident, and in the case of the event $m = 2$, slightly worse for algorithms M_0 and M_1.
- The best overall performance is observed for BJ (FG-true) with optimization of tuning parameters, followed by BJ-RF with optimization of tuning parameters. The BJ approach has the advantage of not needing to estimate the censoring distribution at all. In general, algorithms that use the approach of [28] for estimating the Ψ required for computing the augmentation term (i.e., a parametric model that agrees with data generating mechanism) perform somewhat better than those that use `rfsrc` for this same limited purpose; however, the results are not dramatically different.
- The proposed algorithms perform as well, and often somewhat better, than `rfsrc` in terms of minimizing $MSE(t)$, whether or not tuning parameters are optimized.

The results using Algorithm M_1 were also re-run with $R = 500$ and $B = 1$, and were indistinguishable from those summarized here (results not shown).

4.3 Other simulation results

The Supplementary Material contains additional simulation results. In Section A.1.2 results are provided for the same model as described in Section 4.1, but where the covariates are correlated with each other; see Figures A.5–A.8 which repeat the figures generated for the main simulation study in this alternative setting for $m = 1$. The simulations summarized in the main paper, and in Section A.1.2, are then repeated for $n = 500$; see Figures A.9–A.16 in Section A.1.3. Finally, a different simulation model based on the accelerated failure time model for competing risks described in [29], where we consider $n = 250$ as in the main paper; the model and results can be found in Section A.1.4. In all cases, the results follow the same general patterns as summarized in the previous section.

5 Example: lung cancer treatment trial

5.1 Main results

We illustrate our methods using data from the RTOG 9410, a randomized trial of patients with locally advanced inoperable non-small cell lung cancer. The motivation for this trial was to ascertain whether sequential or concurrent delivery of chemotherapy and thoracic radiotherapy (TRT) is a better treatment strategy. The original RTOG 9410 study randomized 610 patients to three treatment arms: sequential chemotherapy followed by radiotherapy (RX = 1); once-daily chemotherapy concurrent with radiotherapy (RX = 2); and, twice-daily chemotherapy concurrent with radiotherapy (RX = 3). The primary endpoint of interest was overall survival and the main trial analysis results were published in [30], demonstrating a survival benefit of concurrent delivery of chemotherapy and TRT compared with sequential delivery. Secondary analyses of the data using the time from randomization to the first occurrence of three possible outcomes are considered: in-field failure (cancer recurrence within the treatment field for TRT); out-field failure (cancer recurrence and distant metastasis outside of the treatment field for TRT); and, death without documented in-field or out-field failure (i.e., without observed cancer progression). Among these event types, those that first experienced

out-field failures are of particular interest since these patients typically have suboptimal prognosis and may be candidates for more intensified treatment regimens intended to prevent distant metastasis, including but not limited to consolidative chemotherapy, prophylactic cranial irradiation (for brain metastases), and so on. As such, patients that experienced both in-field failure and out-field failure were considered to be out-field failures for purposes of this analysis.

At the time the study database was last updated in 2009, there were 577 patients, with approximately 4% censoring on the aforementioned outcomes. Our methods could be applied to directly analyze this final dataset. However, because the censoring rate is so low, we have decided to take a more illustrative approach. Specifically, we first create a "fully observed" dataset by removing the 23 censored observations. We then compare the results of analyses of the resulting uncensored dataset of 554 patients to analyses of data that were created from this uncensored dataset using an artificially induced censoring mechanism. The main purpose of doing this analysis is two-fold; first, with such a low censoring rate, the results for the uncensored dataset should largely reflect an analysis that would be done for the full dataset of 577 patients; second, we are now able to study how the introduction of (artificial) censoring affects the results and, in particular, illustrate how well the various procedures recover the estimator that would be obtained had outcomes been fully observed (i.e., no random loss to follow-up).

We focus on building forests for each outcome (i.e., out-field failure, death) using a composite loss function with three time points (5.2, 8.5, 15.9 months), selected as the 25th, 50th and 75th percentiles of the observed "all cause" event time distribution (i.e., T). Some related analyses using regression trees alone may be found in [24]. Baseline covariates included in this analysis are RX (Treatment), Age, Stage (American Joint Committee on Cancer [AJCC] stage IIIB vs. IIIA or II), Gender, KPS (Karnofsky performance score of either 70, 80, 90 or 100), Race (White vs. non-White), and Histology (Squamous vs. non-Squamous). Censoring is created according to a Uniform $[0, 50]$ distribution, generating approximately 29% censoring on T. In addition to building forests using the uncensored version of the dataset using the methods described in Section 3.1, we consider the methods *BJ-RF* and *DR-RF* based on Algorithm M_1 using optimally tuned parameters as described in the Simulation results, with $R = 500$ and $B = 1$ (i.e., 500 bootstrap samples). For comparison, we also report results obtained using `rfsrc` using the same approaches to setting the indicated parameters.

To summarize the results in a meaningful way, we created partial dependence plots (PDP) [e.g., [31] to characterize the influence of Age and KPS on the CIF. For reference, the middle 50% of patients in this dataset are aged 54–67, and 76% of patients have KPS scores of 90 or above. The PDPs for the 50th percentile time point for the outfield failure and death outcomes are summarized in Figure 5. For outfield failure, the CIFs do not demonstrate substantial changes across the levels of KPS, though a slight uptick in risk is observed for the healthiest patients when using both *BJ-RF* and *DR-RF*; however, there is a decreasing risk with increasing age, which may possibly be due to patients dying before experiencing outfield failure. For death, all methods suggest a decreased risk of death for healthier patients, and an increasing risk of death as patients age, particularly for the oldest patients. Similar trends are observed when looking at the CIF values calculated at other time points; see Figure A.21–A.24 in Section A.2 of the Supplementary Material. Other noteworthy features from these plots include (i) the trends seen in the PDPs for all methods are generally comparable; and, (ii) the CIF estimates produced by `rfsrc` are always smaller than those obtained using the proposed methods for these data, even in the case where the data are not censored.

In Figure 6, the difference in the PDPs for KPS and Age that are obtained using censored and uncensored data are compared. Specifically, for *BJ-RF* and *DR-RF*, we calculate the difference compared to the uncensored estimator obtained using the methods described in Section 3.1; for `rfsrc`, we compute the estimators obtained using the censored and uncensored outcomes and calculate the difference. In general it can be seen that censoring has a minimal effect on the PDP estimates for all methods, though there is evidence of a somewhat more pronounced difference for outfield failure at the lowest and highest ages, particularly for `rfsrc`. Importantly, however, these results may reflect the comparatively small number of patients at these ages (i.e., only 5% of the patients are under 45, and only 5% are older than 74). Overall, *BJ-RF* tends to exhibit the smallest changes when comparing results for censored and uncensored data.

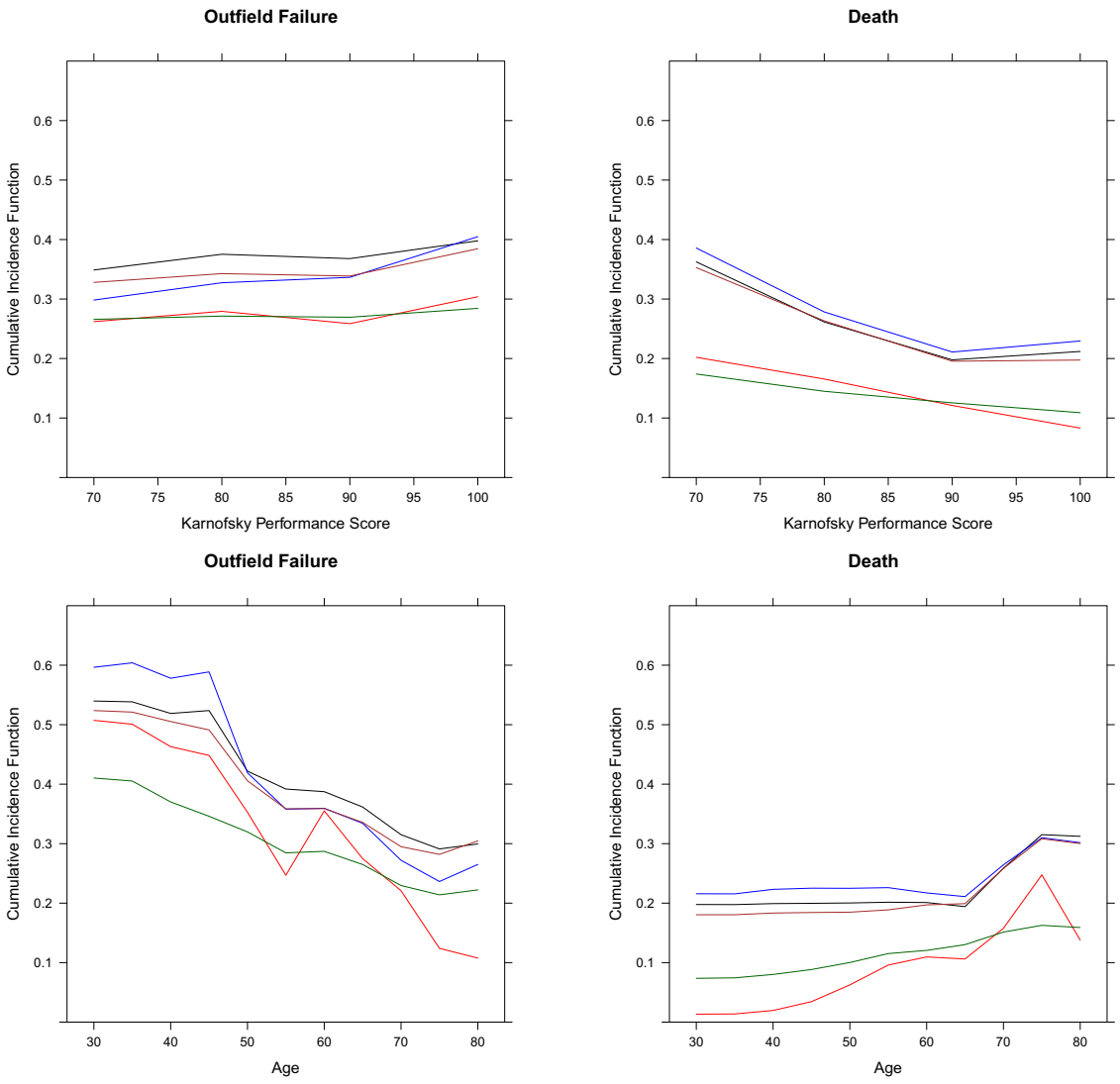

Figure 5: Partial dependence plots with respect to KPS (top) and age (bottom) from forests with outfield failure (left) and death (right) using artificially censored data and uncensored data for 50th percentile time. The five lines are as follows: black = BJ; blue = DR; red = `rfsrc`(censored); brown = uncensored data estimate using proposed methods in Section 3.1; dark green = uncensored data estimate using `rfsrc`.

We also compute similar measures for the categorical variables and report those in Tables A.1 and A.2 of the Supplementary Material. Similarly to Age and KPS, the proposed methods tend to estimate CIFs that are larger than those estimated by `rfsrc` for these data. In all cases, the estimated CIF values demonstrate monotonicity in time; this is easier to see in Tables A.1 and A.2 than it is in the figures generated for KPS and Age. In addition, we again observe that the impact of censoring is relatively small when comparing the results for censored and uncensored data within each method for estimating the CIF.

The difference in the estimates obtained between `rfsrc` and the proposed methods persist whether or not there is censoring present. In the absence of censoring, both `rfsrc` and the proposed methods use a bootstrap ensemble of trees; moreover, for each tree in the ensemble, the CIF is estimated in each terminal node using the corresponding average cause-specific number of events in that node. Hence, the differences observed here in the case of uncensored data, and consequently also in the case of censored data, appear to stem from the different splitting rules used to build the trees that make up each ensemble.

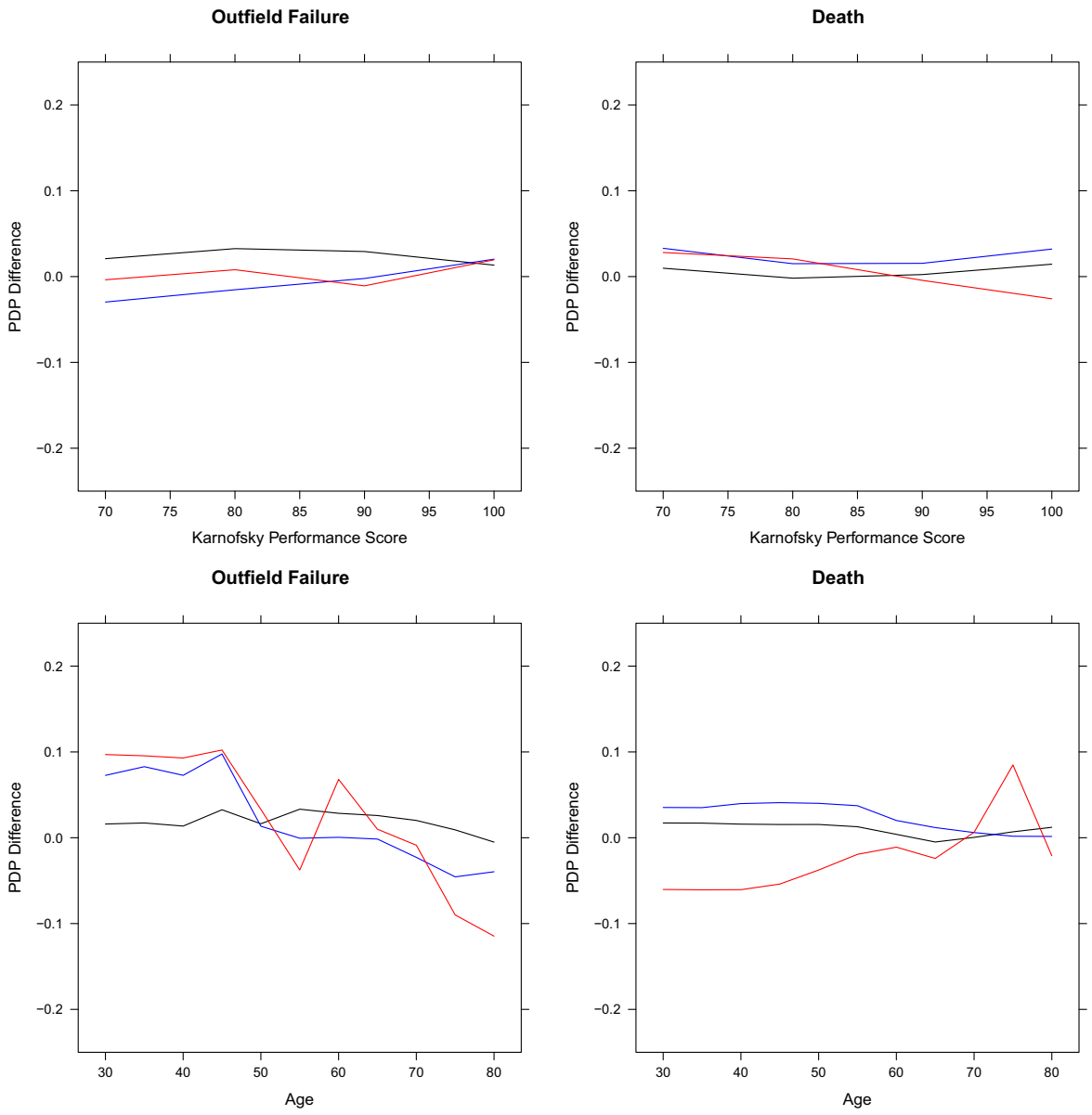

Figure 6: Plots of the difference in partial dependence CIF estimates with respect to KPS (top) and age (bottom) between uncensored and censored data with outfield failure (left) and death (right) for 50th percentile time (black = BJ; blue = DR; red = `rfsrc`).

5.2 Predictive performance and variable importance

In order to quantify the predictive performance of the ensemble models fit to the data, hence prognostic potential, we used the methods of [19, Thm. 4.1] to estimate the expected quadratic loss of prediction (i.e., expected Brier score) using a weighted mean squared error calculation. This risk estimate is calculated using 5-fold cross-validation to reduce the potential finite sample bias resulting from using the same data to both estimate the CIF and assess prediction performance. The results are reported in Supplementary Table A.3 for the artificially censored outcomes. The table shows (i) negligible differences across methods; and, (ii) prediction error increases with time, as expected.

Measures of variable importance for the artificially censored outcomes have also been computed. Currently, `MultivariateRandomForest` only reports the average number of occurrences of each variable across

all splits in all bootstrap trees. Because we have not seen this measure used elsewhere, we have elected to compute the default Breiman–Cutler measure of variable importance that is supplied by `rfsrc` [4]. As noted earlier, the multivariate regression loss functions used in [4] use a dynamic form of normalization that fails to maintain the desired equivalence between (12) and (15). Hence, for the proposed methods, we use `rfsrc` to implement these losses separately at each time point (i.e., using Algorithm M_0 in the case of DR); for comparison, we also compute the default measure reported by `rfsrc` as implemented for a competing risk outcome. This approach complements the composite univariate measure reported by `rfsrc` by allowing a user to investigate how a variable's importance may change across time. Tables A.4–A.7 summarize the results for the outfield failure and death events. Tables A.4 and A.6 report the computed variable importance measure; Tables A.5 and A.7 report the corresponding ranks of the variables within each method to facilitate comparisons across time and method. Tables A.5 and A.7 show, in particular, that while there is some variability in the ranked importance measures (i.e., across time as well as method), there is also a reasonable degree of agreement as far as the variables that rank as being most important. For example, taking the median rank for outfield field failure across time for the doubly robust and Buckley–James losses, the variables Age, Histology, and AJCC stage are ranked as most important; for `rfsrc` as implemented for a competing risk outcome, the variables Age, AJCC, and Karnofsky Performance score are ranked as most important. In the case of death, we instead respectively see that Age, Histology and Gender and Karnofsky Performance score, Histology and Age are ranked as most important.

6 Discussion

In our simulation studies, the proposed methods demonstrate similar or better performance (i.e., with respect to the chosen MSE measure) compared to the methods of [4], as implemented in the `randomForestSRC` package. Of interest is the ease with which the proposed methods can be implemented using existing software, including for CIF regression trees.

The proposed methods focus on building a tree or ensemble estimator for a single cause in the presence of other possible causes. When interest lies in multiple causes, the method can be applied to each cause in exactly the same manner. However, this is arguably inefficient, and one interesting direction for further research would be to extend both regression tree and ensemble procedures to the problem of simultaneous estimation of multiple CIFs. For example, when building a regression tree, one could easily adapt the approach taken earlier to accommodate multiple causes in addition to multiple time points; see [4] for a similar proposal. In the case of a regression tree, such an approach makes the restrictive assumption that the predictor space is to be partitioned in the same way for all causes. This restriction may ultimately be less worrisome when building a predictor derived from ensembles of trees, but it would be interesting to explore alternative ways in which one might proceed. A second area of possible extension would be to consider generalizations of the model-based recursive partitioning algorithm proposed in [32] to the setting of CIF estimation using a composite loss function, leading to alternative tree and/or ensemble estimators.

The reliance of the proposed methods on the need to estimate a "nuisance" parameter that coincides with the target of primary interest (i.e., the CIF) is an evident drawback of the proposed approach. Of course, this problem is inherent to using all augmented IPCW estimators. Currently, we use `rfsrc` for this purpose, and it is therefore perfectly reasonable to ask whether the proposed approach offers any advantages over the methods introduced in [4]. We believe the answer to this question is affirmative. For example, our approach provides the ability to estimate what happens at specific points in time. In addition, `rfsrc` uses a log-rank-type splitting process for competing risks that may be adversely influenced by informative censoring, particularly so in the early stages of splitting; see, for example [9, 10], and especially [33] for discussion and results in the case of right-censored survival data. Although the validity of the proposed methods also requires that $(T, M) \perp C \mid W$, the observed data loss functions used in Algorithms M_0 and M_1 are approximately unbiased both conditionally and unconditionally on W and should be less susceptible to similar biases. It would be interesting to study the performance of iterated versions of the proposed algorithms in which the CIF required

for computing the augmentation term is updated with each iteration of the proposed algorithm, possibly only being initialized with `rfsrc`.

Acknowledgment: We thank the NRG Oncology Statistics and Data Management Center for providing de-identified RTOG 9410 clinical trial data under a data use agreement.
Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: This work was partially supported by the National Institutes of Health (R01CA163687: AMM, RLS, YC; U10-CA180822: CH).
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

1. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999;94:496–509.
2. Dignam J, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. Clin Cancer Res 2012;18:2301–8.
3. Breiman L. Random forests. Mach Learn 2001;45:5–32.
4. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. Biostatistics 2014;15:757–73.
5. Mogensen UB, Gerds TA. A random forest approach for competing risks based on pseudo-values. Stat Med 2013;32:3102–14.
6. Ishwaran H, Kogalur U. Random Forests for Survival, Regression and Classification (RF-SRC). Version 2.4.1. R Foundation for Statistical Computing; 2016.
7. Gray RJ. A class of k-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat 1988;16:1141–54.
8. Aalen O, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand J Stat 1978;5:141–50.
9. Steingrimsson JA, Diao L, Molinaro AM, Strawderman RL. Doubly robust survival trees. Stat Med 2016;35:3595–612.
10. Steingrimsson JA, Diao L, Strawderman RL. Censoring unbiased regression trees and ensembles. J Am Stat Assoc 2019;114:370–83.
11. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3.
12. Scheike TH, Zhang M-J, Gerds TA. Predicting cumulative incidence probability by direct binomial regression. Biometrika 2008;95:205–20.
13. Molinaro AM, Dudoit S, Van der Laan MJ. Tree-based multivariate regression and density estimation with right-censored data. J Multivariate Anal 2004;90:154–77.
14. Lostritto K, Strawderman RL, Molinaro AM. A partitioning deletion/substitution/addition algorithm for creating survival risk groups. Biometrics 2012;68:1146–56.
15. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2:841–60.
16. Tsiatis AA. Semiparametric Theory and Missing Data. New York: Springer; 2007.
17. Strawderman RL. Estimating the mean of an increasing stochastic process at a censored stopping time. J Am Stat Assoc 2000;95:1192–208.
18. Buckley J, James I. Linear regression with censored data. Biometrika 1979;66:429–36.
19. Schoop R, Beyersmann J, Schumacher M, Binder H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biom J 2011;53:88–112.
20. Cruz-Uribe D, Neugebauer C. Sharp error bounds for the trapezoidal rule and simpson's rule. J Inequalities Pure Appl Math 2002;3:22.
21. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Monterey CA: Wadsworth and Brooks; 1984.
22. Rahman R. MultivariateRandomForest: models multivariate cases using random forests. R package version 1.1.5, (2017).
23. Segal M, Xiao Y. Multivariate random forests. Wiley Interdisciplinary Reviews: Data Min Knowl Discov 2011;1:80–7.
24. Cho Y, Molinaro AM, Hu C, Strawderman RL. Regression trees for cumulative incidence functions; 2020, arXiv (stat.ME; 2011.06706).
25. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 1993;80:557–72.
26. LeBlanc M, Crowley J. Relative risk trees for censored survival data. Biometrics 1992;48:411–25.

27. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the rpart routines; 2015.
28. Jeong J-H, Fine JP. Parametric regression on the cumulative incidence function. Biostatistics 2007;8:184–96.
29. Sun Y, Wang HJ, Gilbert PB. Quantile regression for competing risks data with missing cause of failure. Stat Sin 2012;22:703.
30. Curran WJ, Paulus R, Langer CJ, Komaki R, Lee JS, Hauser S, et al. Sequential vs concurrent chemoradiation for stage III non–small cell lung cancer: randomized phase III trial RTOG 9410. J Natl Cancer Inst 2011;103:1452–60.
31. Greenwell BM. pdp: an R package for constructing partial dependence plots. R J 2017;9:421.
32. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. J Comput Graph Stat 2008;17:492–514.
33. Cui Y, Zhu R, Zhou M, Kosorok M. Consistency of survival tree and forest models: splitting bias and correction; 2019, arXiv (math.ST; 1707.09631).

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/ijb-2021-0014).