10-31-2016

Censoring Unbiased Regression Trees and Ensembles

Jon Arni Steingrimsson
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Liqun Diao
Department of Statistics and Actuarial Science, University of Waterloo

Robert L. Strawderman
Department of Biostatistics and Computational Biology, University of Rochester, robert.strawderman@urmc.rochester.edu

Censoring Unbiased Regression Trees and Ensembles

JON ARNI STEINGRIMSSON
Department of Biostatistics
Johns Hopkins University, Baltimore MD, USA
jsteing5@jhu.edu

LIQUN DIAO *
Department of Statistics and Actuarial Science
University of Waterloo, Waterloo ON, Canada
l2diao@uwaterloo.ca

ROBERT L. STRAWDERMAN †
Department of Biostatistics and Computational Biology,
University of Rochester, Rochester NY, USA
robert_strawderman@urmc.rochester.edu

Abstract. This paper proposes a novel approach to building regression trees and ensemble learning in survival analysis. By first extending the theory of censoring unbiased transformations, we construct observed data estimators of full data loss functions in cases where responses can be right-censored. This theory is used to construct two specific classes of methods for building regression trees and regression ensembles that respectively make use of Buckley-James and doubly robust estimating equations for a given full data risk function. For the particular case of squared error loss, we further show how to implement these algorithms using existing software (e.g., CART, random forests) by making use of a related form of response imputation. Comparisons of these methods to existing ensemble procedures for predicting survival probabilities are provided in both simulated settings and through applications to four datasets. It is shown that these new methods either improve upon, or remain competitive with, existing implementations of random survival forests, conditional inference forests, and recursively imputed survival trees.

*Co-First Author
†To whom correspondence should be addressed.

1 Introduction

Recursive partitioning methods for regression problems provide a useful nonparametric alternative to parametric and semiparametric methods. Methods based on the Classification and Regression Trees (CART; Breiman et al., 1984) algorithm are the most popular recursive partitioning procedures in use today. One of the most attractive features of CART is its focus on building a simple, interpretable tree-structured prediction model. In the original formulation of this algorithm, the resulting hierarchically structure predictor is determined by maximizing within-node homogeneity using principles of loss minimization. However, a common criticism of CART is that the final predictor can suffer from instability, a phenomenon that usually occurs in settings where a small change in the loss can induce a large change in the form of the predictor (Breiman, 1996).

Bagging is a general method for variance reduction that averages several prediction models derived from bootstrapping the original data. Bagging has been shown to work well with models with low bias and high variance (e.g., fully grown CART trees); see Breiman (1996). Random Forests (RF) attempts to further improve prediction accuracy (i.e., in the regression setting under squared error loss) by de-correlating the individual trees through random feature selection at each split-point (Breiman, 2001). The combination of bagging with random feature selection used by the RF algorithm has proved to be a very effective tool for increasing prediction accuracy.

A survival tree (survival forest) is built when a suitably modified version of the CART (RF) algorithm is applied to data involving an outcome that can be right-censored. For building single trees, several variations of CART have been proposed and can be divided into two general categories: one category focused on maximizing within-node homogeneity (e.g., Gordon and Olshen, 1985; Davis and Anderson, 1989; LeBlanc and Crowley, 1992; Keleş and Segal, 2002; Molinaro et al., 2004; Steingrimsson et al., 2016) and the other category focused on maximizing between-node heterogeneity (e.g., Segal, 1988; Leblanc and Crowley, 1993). Ishwaran et al. (2008) proposed the Random Survival Forests (RSF) algorithm, modifying the RF algorithm for survival data through building the individual trees in the forest via maximizing the between-node log-rank statistic. More recently, Zhu and Kosorok (2012) proposed the recursively imputed survival trees (RIST) algorithm. Similarly to the RSF algorithm, RIST makes splitting decisions by maximizing a log-rank test statistic; however, it differs from RSF in several important ways. In particular, in place of

bagging, the RIST algorithm generates an ensemble of predictors by recursively imputing censored observations; second, it makes use of extremely randomized trees, replacing RF's search for optimal split points with a splitting value that "is chosen fully at random" (i.e., decisions to split are made on the basis of K randomly selected pairs of covariates and possible split points).

With the exception of Molinaro et al. (2004) and Steingrimsson et al. (2016), the afore-cited methods for censored data all use splitting rules specifically constructed to deal with the presence of censored outcome data. These various methods share a common feature, namely that none reduce to a loss-based method of analysis that might ordinarily be used if censoring were absent. For single trees, Molinaro et al. (2004) closed this gap between tree-based regression methods used for censored and uncensored data by applying inverse probability censoring weighted (IPCW) theory to construct an IPCW-weighted loss function that (i) reduces to the "full data" loss function that would be used by CART in the absence of right-censored outcome data; and, (ii) is an unbiased estimator of the desired full data risk in the presence of right-censored outcome data. Based on similar principles, Hothorn et al. (2006a) proposed a RF-type algorithm that selects subjects into the bootstrap sample using the non-parametric IPCW bootstrap. Steingrimsson et al. (2016) proposed to use doubly robust survival trees, generalizing the methods of Molinaro et al. (2004) by using augmentation to construct "doubly robust" loss functions that use more of the available information, thereby improving the efficiency and stability of the tree building process. However, they did not generalize these ideas to the problem of constructing an appropriate ensemble procedure.

The main focus of this paper is on developing a new class of ensemble algorithms for building survival forests that use unpruned survival trees as the base learner. It is first shown that the IPCW and the "doubly robust" survival tree methods considered in Molinaro et al. (2004) and Steingrimsson et al. (2016) are special cases of a new and more general class of tree building procedures that can be derived from CART by making use of the theory of censoring unbiased transformations (CUTs; e.g., Fan and Gijbels, 1996). The theory of CUTs, a type of mean imputation procedure, has long been of importance to survival analysis; important examples include Buckley and James (1979), Koul et al. (1981), Leurgans (1987), Fan and Gijbels (1994), and Rubin and van der Laan (2007). The use of a CUT for the desired loss function (i.e., a loss function that would be used in the absence of censoring) ensures that the resulting regression tree algorithm directly generalizes its uncensored counterpart. On the basis of these results, we then propose a new class of ensemble

algorithms for a right-censored outcome. The proposed methods extend traditional ensemble procedures by permitting the use of exchangeably weighted bootstrap sampling schemes (Præstgaard and Wellner, 1993); in the case of the nonparametric bootstrap and squared error loss, the proposed algorithm also reduces to the RF algorithm when censoring is absent. For the particular case of squared error loss, we additionally show how to make use of existing CART and RF procedures for uncensored outcomes to implement these new algorithms when censoring is present.

The remainder of this paper is organized as follows. Section 2.1 defines notation and data structures. Section 2.2 extends the existing theory on CUTs in a substantial way. Section 2.3 uses these results to construct observed (i.e., censored) data loss functions that are unbiased estimators of the expected loss (i.e., risk) that is obtained when censoring is absent. Section 3 gives an overview of how the results in Section 2.3 can be used to derive two new classes of methods for building regression trees and regression ensembles with a right-censored outcome. An important feature of these new algorithms is that each represents a direct generalization of a loss-based algorithm that would be used if censoring were absent. Section 4 provides a detailed development of these algorithms in the important case of squared error loss and shows, in particular, how response imputation can be used to implement each one using existing software for CART and RF (i.e., for uncensored outcomes). Simulation studies and applications to several datasets are respectively presented in Section 5 and 6. Section 7 contains a discussion and some general remarks on topics for future research. A Supplementary Web Appendix contains proofs, additional developments and further results.

2 Notation and Important Preliminaries

2.1 Data Structures

Let the full data available on a given subject be $(Z, W')'$, where $Z = h(T)$, $T > 0$ denotes a survival time, $W \in \mathcal{S}$ is a bounded p-dimensional vector of covariates, and $h(\cdot)$ is a specified continuous, monotone increasing function (e.g., $h(u) = u$ or $h(u) = \log u$) mapping $\mathbf{R}^+$ to $\mathbf{R}^* \subseteq \mathbf{R}$.

Let $S_0(t|w) = P(T > t|W = w)$ denote the conditional survivor function for T given $W = w$; it is assumed that T is continuous and that $\vartheta_{S_0} = \inf\{t : S_0(t|w) = 0\}$ is independent of $w \in \mathcal{S}$.

The observed (i.e., right-censored) data on a given subject will be denoted $O = (\widetilde{Z}, \Delta, W')'$, where $\tilde{Z} = h(\tilde{T})$, $\widetilde{T} = \min(T, C)$ for a censoring time C, and $\Delta = I(T \leq C)$ indicates whether T or C was observed. Let $G_0(t|w) = P(C > t|W = w)$ be the conditional survivor function for C given $W = w$; it is assumed that C is conditionally independent of T given W, that C is continuous and that $\vartheta_{G_0} = \inf\{t : G_0(t|w) = 0\}$ is independent of $w \in \mathcal{S}$. Finally, define $\mathcal{F} = \{(Z_i, W_i')', i = 1 \ldots n\}$ to be the full data available on an independent and identically distributed sample of data; similarly, let $\mathcal{O} = \{(\tilde{Z}_i, \Delta_i, W_i')', i = 1 \ldots n\}$ denote the corresponding observed data.

2.2 Censoring Unbiased Transformations: Review and Generalization

Let $\phi(r, w), (r, w) \in \mathbf{R}^* \times \mathcal{S}$ be a known scalar function that is continuous for $r \in \mathbf{R}^*$ except possibly at a finite number of points; in addition, assume $|\phi(r, w)| < \infty$ whenever $\max\{|r|, \|w\|\} < \infty$, and suppose that $E[\phi(Z, W)|W = w]$ exists for each $w \in \mathcal{S}$. Let $Y = \phi(Z, W)$ and suppose $Y^*(O)$ is a function of the observed data. Then, $Y^*(O)$ is said to be a censoring unbiased transformation (CUT) for Y if $E[Y^*(O)|W = w] = E[Y|W = w]$ for every $w \in \mathcal{S}$; see Fan and Gijbels (1996) and also Rubin and van der Laan (2007).

Let $G(t|w)$ and $S(t|w)$ be functions on $\mathbf{R}^+ \times \mathcal{S}$. For every $w \in \mathcal{S}$, we assume throughout this section that $G(0|w) = S(0|w) = 1$ and that $G(u|w) \geq 0$ and $S(u|w) \geq 0$ are continuous, non-increasing functions for $u \geq 0$ (e.g., proper survivor functions). We will additionally have repeated need for the integral

$$\Lambda_G(t|w) = -\int_0^t \frac{dG(u|w)}{G(u|w)}; \tag{1}$$

note that $\Lambda_G(t|w)$ is just the cumulative hazard function corresponding to $G(\cdot|\cdot)$ in the case where $G(\cdot|\cdot)$ is a proper survivor function.

The Buckley-James mapping (Buckley and James, 1979) is one of the earliest such examples of a CUT. In the current context, we define the relevant Buckley-James transformation as

$$Y_b^*(O; S) = \Delta\phi(\tilde{Z}, W) + (1 - \Delta)m_\phi(\tilde{T}, W; S) \tag{2}$$

where

$$m_\phi(t, w; S) = \frac{\int_t^\infty \phi(h(u), w)dF(u|w)}{S(t|w)} \tag{3}$$

is continuous and $F(u|w) = 1 - S(u|w)$ for any $u \geq 0$. The original Buckley-James transformation is recovered upon setting $\phi(h(u), w) = u$. When $S(\cdot|\cdot)$ is a proper survivor function and t is such that (3) exists, we may interpret $m_\phi(t, w; S)$ as being equal to $E_S[\phi(Z, W)|T > t, W = w]$, the expectation being calculated assuming $S(\cdot|\cdot)$ is the conditional survivor function for T. The transformation (2) has the desirable property of reducing to $Y = \phi(Z, W)$ when $\Delta = 1$ (i.e., in the absence of censoring). Provided (3) exists for each $t \geq 0$, it is also easily proved that $Y_b^*(O; S)$ is a CUT when $S(t|w) = S_0(t|w)$ for all $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$ and that $Y_b^*(O; S_0)$ is the best predictor of Y in the sense that it minimizes $E[(Y^*(O) - Y)^2|W]$ among all possible CUTs $Y^*(O)$ (e.g., Fan and Gijbels, 1996). Because $S_0(\cdot|\cdot)$ is unknown, any plug-in estimator $\hat{S}(\cdot|\cdot)$ must be consistent for $S_0(\cdot|\cdot)$ in order for $Y_b^*(O; \hat{S})$ to behave as a CUT in large samples.

Motivated by the need to correctly specify $S(\cdot|\cdot)$ in (2), Rubin and van der Laan (2007, Eqn. 7) proposed a "doubly robust" CUT in the case where $Y = T$; we discuss the specific meaning of this property later in this section. In this paper, we introduce the following generalization:

$$Y_d^*(O; G, S) = \frac{\Delta\phi(\tilde{Z}, W)}{G(\tilde{T}|W)} + \frac{1 - \Delta}{G(\widetilde{T}|W)} m_\phi(\widetilde{T}, W; S) - \int_0^{\widetilde{T}} \frac{m_\phi(u, W; S)}{G(u|W)} d\Lambda_G(u|W), \tag{4}$$

where $\Lambda_G(t|w)$ is given by (1). The selection $\phi(h(u), w) = u$ in (4) reproduces the doubly robust CUT studied in Rubin and van der Laan (2007, Eqn. 7). Section 2.3 introduces an important class of examples where $\phi(h(u), w)$ depends on both $h(u)$ and w.

It can be seen that (i) (4) reduces to (2) if one defines $G(t|w) = 1$ for all $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$; and, (ii) (4) reduces to the IPCW estimate $\Delta\phi(\tilde{Z}, W)/G(\tilde{T}|W)$ if $m_\phi(t, w; S) = 0$ for all $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$. Under certain conditions, it will be shown that the transformation (4) can be considered doubly robust in the sense that one obtains a CUT for $Y = \phi(Z, W)$ if either $S(t|w) = S_0(t|w)$ or $G(t|w) = G_0(t|w)$ for all $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$. Moreover, if both of these functions are correctly specified, then $Y_d^*(O; G_0, S_0)$ can be shown to minimize the variance among all transformations of the form $Y_d^*(O; G_0, S)$. These results are summarized in Theorem 2.1 and proved in the Supplementary Web Appendix (Section S.4). The transformation (4) also reduces to $Y = \phi(Z, W)$ regardless of $S(\cdot|\cdot)$ when $\Delta = 1$ provided that $G(t|W) = 1$ for $t \leq \tilde{T}$ (i.e., with probability one); that is, when there is no possibility of censoring on $[0, \tilde{T}]$. As written, (4) depends on the generic survivor functions $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$; the double robustness property implies that the estimated transformation $Y_d^*(O; \hat{G}, \hat{S})$

will behave as a CUT in large samples if at least one, but not necessarily both, of the plug-in estimators $\hat{G}(\cdot|\cdot)$ and $\hat{S}(\cdot|\cdot)$ are respectively consistent for $G_0(\cdot|\cdot)$ and $S_0(\cdot|\cdot)$.

Theorem 2.1. *Let $S(\cdot|\cdot)$ and $G(\cdot|\cdot)$ be any two functions on $(t,w) \in \boldsymbol{R}^+ \times \mathcal{S}$ satisfying the regularity conditions given in Appendix S.4. Then, the transformations $Y_d^*(O;G,S_0)$, $Y_d^*(O;G_0,S)$ and $Y_d^*(O;G_0,S_0)$ are each CUTs for Y; furthermore, $Var(Y_d^*(O;G_0,S)|W) \geq Var(Y_d^*(O;G_0,S_0)|W)$.*

Theorem 2.1 shows $Var(Y_d^*(O;G_0,S)|W) \geq Var(Y_d^*(O;G_0,S_0)|W)$ for any suitable proper survivor function. One may also ask whether $Var(Y_d^*(O;G,S_0)|W) \geq Var(Y_d^*(O;G_0,S_0)|W)$ holds for all suitable choices of $G(\cdot|\cdot)$. However, a general result in this direction is not available even for the interesting case where $G(\cdot|\cdot) = 1$ (i.e., for (2)). The inability to establish such a domination result reflects more general open questions surrounding the development of efficiency properties for doubly robust estimators under misspecification of the missing data mechanism; see Rotnitzky and Vansteelandt (2014, Sec. 9.6) for further discussion.

2.3 Using CUTs to Derive Unbiased Estimates of Risk with Censored Data

We remind the reader of the notation and assumptions introduced in Sections 2.1 and 2.2. Define $\psi : \mathcal{S} \to \mathbf{R}$ to be a real-valued function of W, where $\psi \in \Psi$. Let $L(Z,\psi(W))$ denote a loss function that depends on the full data (Z,W); we can then define the corresponding risk, or expected loss, as $\mathcal{R}(\psi) = E\left[L(Z,\psi(W))\right]$. The $\psi \in \Psi$ that minimizes this risk function, say ψ_0, defines a target parameter of interest that is ideally uniquely specified. For example, given the squared error (i.e., L_2) loss function $L(Z,\psi(W)) = (Z-\psi(W))^2$, denoted hereafter by $L_2(Z,\psi(W))$, the associated risk is $\mathcal{R}(\psi) = E[(Z-\psi(W))^2]$ and is minimized at the target parameter $\psi_0(W) = E[Z|W]$. Thus, for example, selecting $h(s) = \log s$ yields $Z = \log T$ and leads to a full data loss function with corresponding risk minimized at $\psi_0(W) = E[\log T|W]$. Alternatively, for a given $t > 0$, selecting $h(s) = I(s > t)$ yields $Z = I(T > t)$ and leads to a full data loss function with corresponding risk minimized at $\psi_{0t}(W) = S_0(t|W)$. This latter formulation is directly related to the so-called Brier score (cf., Brier, 1950).

In the case where Z can be right-censored, one cannot simply replace Z by $\tilde{Z}$ in $L(\tilde{Z},W)$; for example, $E[L(\tilde{Z},W)] \neq \mathcal{R}(\psi)$ in general. However, as shown in Molinaro et al. (2004), it is still possible to construct an observed data loss function that has the same risk $R(\psi)$. Specifically,

assuming $\mathcal{R}(\psi)$ exists and that $P(G(T|W) \geq \epsilon) = 1$ for some $\epsilon > 0$, the inverse probability of censoring weighted (IPCW) loss function

$$L_{ipcw}(O, \psi; G) = \frac{\Delta L(\tilde{Z}, \psi(W))}{G(\tilde{T}|W)} = \frac{\Delta L(Z, \psi(W))}{G(T|W)}$$

satisfies $E[L_{ipcw}(O, \psi; G_0)] = E[L(Z, \psi(W))] = \mathcal{R}(\psi)$. That is, $L_{ipcw}(O, \psi; G_0)$ is an unbiased estimator of the desired risk $\mathcal{R}(\psi)$ when $G(t|w) = G_0(t|w)$ for all $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$. In fact, given any $\psi(\cdot)$, $E[L_{ipcw}(O, \psi; G_0)|W] = E[L(Z, \psi(W))|W]$ under the same regularity conditions. Consequently, $L_{ipcw}(O, \psi; G_0)$ is a CUT for $\phi(Z, W) = L(Z, \psi(W))$; see Section 2.2.

By applying the theory for augmented estimators in missing data problems as developed in Tsiatis (2007, Ch. 9 & 10), Steingrimsson et al. (2016) derived a doubly robust estimator for $R(\psi)$. Specifically, the observed data estimator used in Steingrimsson et al. (2016) is given by

$$L_d(O, \psi; G, S) = \frac{\Delta L(\tilde{Z}, \psi(W))}{G(\tilde{T}|W)} + \frac{1 - \Delta}{G(\widetilde{T}|W)} m_L(\widetilde{T}, W; S) - \int_0^{\widetilde{T}} \frac{m_L(u, W; S)}{G(u|W)} d\Lambda_G(u|W), \qquad (5)$$

where $m_L(r, w; S) = E_S[L(h(r), \psi(w))|T > r, W = w]$ for any $r \geq 0$ and $w \in \mathcal{S}$ and the expectation is calculated assuming $T|W = w$ has survivor function $S(\cdot|w)$. Under certain regularity conditions (e.g., boundedness), the double robustness property stems from the fact that the marginal expectation of (5) is $R(\psi)$ if either $G(\cdot|\cdot) = G_0(\cdot|\cdot)$ or $S(\cdot|\cdot) = S_0(\cdot|\cdot)$. This estimator for $R(\psi)$, hereafter referred to as the doubly robust loss function, is also easily seen to be a CUT for $L(Z, \psi(W))$ that is of the form (4). Specifically, for a fixed $\psi(\cdot)$, (5) is obtained directly from (4) upon setting $\phi(Z, W) = L(Z, \psi(W))$, where $m_L(\cdot, w; S)$ is defined as in (3) and depends on $\psi(\cdot)$. Under regularity conditions that permit the application of Theorem 2.1, the observed data estimator (5) satisfies $E[L_d(O, \psi; G_0, S_0)|W] = E[L_d(O, \psi; G_0, S)|W] = E[L_d(O, \psi; G, S_0)|W] = E[L(Z, \psi(W))|W]$. Following Section 2.2, the observed data loss function

$$L_b(O, \psi; S) = \Delta L(\tilde{Z}, \psi(W)) + (1 - \Delta) m_L(\widetilde{T}, W; S) \qquad (6)$$

is obtained as a special case of (5) upon setting $G(t|w) = 1$ for all $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$. This observed data estimator, hereafter referred to as the Buckley-James loss function, is a CUT of the form (2)

for $L(Z, \psi(W))$ when $S(t|w) = S_0(t|w)$ for all $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$.

By Theorem 2.1, for any fixed $\psi(\cdot)$ and under sufficient regularity conditions, $L_d(O, \psi; G_0, S_0)$ has a smaller conditional variance for estimating $\mathcal{R}(\psi)$ in comparison to $L_{ipcw}(O, \psi; G_0)$. However, it is not possible to determine in general whether the conditional variance of $L_b(O, \psi; S_0)$ exceeds that of $L_d(O, \psi; G_0, S_0)$. Estimators derived under these various loss functions can be expected to exhibit different efficiencies. It is further expected that estimators derived under $L_d(O, \psi; G_0, S_0)$ will be more efficient than those derived under $L_{ipcw}(O, \psi; G_0)$. However, it is in general difficult to claim that using one of these loss functions will always lead to better estimators than another, particularly when the unknown functions $G_0(\cdot|\cdot)$ and/or $S_0(\cdot|\cdot)$ are replaced by suitable estimators.

3 Censoring Unbiased Regression Trees and Ensembles

Regression procedures typically rely on the specification of a loss function that quantifies performance. This includes, but is not limited to, algorithms like CART and RF, where the loss function plays a key role in all aspects of the model fitting process. The use of a loss function implies focusing on a prediction model that minimizes a corresponding measure of risk. Below, we show how the developments of Sections 2.2 and 2.3 can be used to devise new regression tree and ensemble methods for right-censored outcomes.

3.1 Regression Tree Algorithms for General Loss Functions

The basic CART algorithm relies on three key steps: (i) Use reduction in loss to split the covariate space until some predetermined criteria are met; (ii) Use the training error plus a penalty proportional to the number of terminal nodes of the tree to create a sequence of candidate trees; (iii) Use cross-validation to select the "best" model from the sequence of candidate trees. Each of these steps relies on the loss function specified in step (i), which in turn defines the relevant risk function and thus the parameter of interest. The CART algorithm is generic in this regard and does not rely on any specific choice of loss function. However, with a continuous outcome Z, the default choice of loss function is almost always $L_2(Z, \psi(W))$, with $\psi(W)$ being a piecewise constant prediction function on $\mathcal{S}$ that CART estimates using a recursive partitioning procedure.

A direct extension of CART to the case where Z can be right-censored is obtained by replac-

ing the uncomputable full data loss $L(Z, \psi(W))$ with a reasonable observed data estimator for $R(\psi) = E[L(Z, \psi(W))]$. For example, as suggested in Molinaro et al. (2004) and for a suitable estimator $\hat{G}(\cdot|\cdot)$ of $G_0(\cdot|\cdot)$, one can replace the loss function $L(Z, \psi(W))$ with $L_{ipcw}(O, \psi; \hat{G})$ throughout the CART algorithm; without further modification, this leads to a new method for building regression trees with right-censored outcomes. Following this same idea and using suitable estimators $\hat{G}(\cdot|\cdot)$ and $\hat{S}(\cdot|\cdot)$ for both $G_0(\cdot|\cdot)$ and $S_0(\cdot|\cdot)$, Steingrimsson et al. (2016) showed how the CART algorithm can be modified to use $L_d(O, \psi; \hat{G}, \hat{S})$ in place of $L(Z, \psi(W))$ to build "doubly robust" regression trees in this same setting. Considering the developments in Section 2.3, it can be seen that both of these procedures are derived by replacing the unobservable full data loss function with a corresponding CUT. Although not specifically considered in the literature to date, the Buckley-James CUT $L_b(O, \psi; \hat{S})$ could also be used in place of either $L_{ipcw}(O, \psi; \hat{G})$ or $L_d(O, \psi; \hat{G}, \hat{S})$. Each choice generates a new survival tree algorithm. We will use CURT to describe any generalization of the CART algorithm constructed using a CUT for a given full data loss $L(Z, \psi(W))$. It is reasonable to expect any CUT for $L(Z, \psi(W))$ to reduce to $L(Z, \psi(W))$ in the absence of censoring; in this case, a CURT algorithm that uses a CUT for $L(Z, \psi(W))$ also reduces to the corresponding CART algorithm that uses $L(Z, \psi(W))$ when censoring is absent.

3.2 Regression Ensemble Algorithms for General Loss Functions

Prediction accuracy is usually improved by averaging multiple bootstrapped trees; see, for example, Hastie et al. (2009). Breiman (1996) proposed bagging, which averages fully grown CART trees (i.e., generated by step (i) of the CART algorithm as described in Section 3.1) using many independent nonparametric bootstrap samples. These boostrapped trees, though conditionally independent of each other, are marginally correlated. Breiman (2001) proposes to reduce this correlation by additionally making use of random feature selection; specfically, the RF algorithm modifies the tree growing procedure so that only $mtry \leq p$ randomly selected covariates are considered for splitting at any given stage. The use of bootstrapping and/or random feature selection does not modify the basic loss-based decision making process that lies at the core of the original RF algorithm. Therefore, at least in principle, it is easy to extend the RF algorithm to the case of a right-censored outcome and/or more general bootstrap schemes (e.g., the exchangeably weighted bootstrap; see

Præstgaard and Wellner, 1993) using the same type of loss-substitution principle described in Section 3.1. An extensive search of the literature revealed no examples of RF algorithms that use bootstrap procedures other than the nonparametric bootstrap.

Specifically, consider $L_d(O, \psi; G, S)$ in (5) calculated using some appropriate full data loss $L(Z, \psi(W))$. Let $\mathcal{O} = (O_1, \ldots, O_n)$ be the observed data and define $\omega_1, \ldots, \omega_n$ to be a set of exchangeable, non-negative random variables such that $E[\omega_i] = 1$ and $\text{Var}(\omega_i) = \sigma^2 < \infty$ for $i = 1, \ldots, n$, $\sum_{i=1}^n \omega_i = n$, and $\omega_1, \ldots, \omega_n$ are completely independent of the observed data. Define the weighted doubly robust loss function

$$L_{d,\omega}(\mathcal{O}, \psi; G, S) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \omega_i L_d(O_i, \psi; G, S) \tag{7}$$

for some (typically estimated) choice of $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$. For reasons that will soon be evident, we assume that $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$ do not depend on $\omega_1, \ldots, \omega_n$. The loss function (7) reduces to the empirical observed data loss function $n^{-1} \sum_{i=1}^n L_d(O_i, \psi; G, S)$ if $P(\omega_1 = \cdots = \omega_n = 1) = 1$; more generally, $E[L_{d,\omega}(\mathcal{O}, \psi; G, S)|\mathcal{O}] = L_d(\mathcal{O}, \psi; G, S)$, implying that the weighted and empirical observed data loss functions have the same marginal expectation. Substituting (7) in for the unobserved empirical full data loss function throughout the CART algorithm leads to a general class of case-weighted CURT algorithms that can be used to build ensemble predictors. A class of ensemble algorithms suitable for right-censored outcomes and general sets of bootstrap weights is summarized in Algorithm 1. The base learners used in Algorithm 1 are modified versions of fully grown CURTs as described in Section 3.1 that incorporate random feature selection.

Algorithm 1 Censoring Unbiased Regression Ensembles (CURE)

1: Generate M independent sets of exchangeable bootstrap weights $\omega_1, \ldots, \omega_n$.
2: For each set of bootstrap weights, build a fully grown CURT tree using the loss function (7) where, at each stage of splitting, *mtry* covariates are randomly selected for candidate splits. At each stage, the variable and split that gives the largest reduction in the loss (7) is used. Splitting continues until some pre-specified criterion is met.
3: For each tree in the forest, calculate an estimator at each terminal node and average over the bootstrap samples to get the final ensemble predictor.

The use of nonparametric bootstrap sampling in Steps (i) & (ii) is equivalent to the multinomial sampling scheme $(\omega_1, \ldots, \omega_n) \sim \text{Multinomial}(n, (n^{-1}, \ldots, n^{-1}))$ and places positive weights on

approximately 63% of the observations in any given bootstrap sample. The exchangeably weighted bootstrap avoids generating additional ties in the data when $P(\cup_i\{\omega_i = 0\}) = 0$; each observation then appears in every bootstrap sample with some positive weight attached to its contribution.

In the absence of censoring, (7) respectively reduces to either the weighted empirical full data loss $n^{-1}\sum_{i=1}^{n}\omega_i L(Z_i, \psi(W_i))$ or, when $P(\omega_1 = \cdots = \omega_n = 1) = 1$, to the unweighted empirical full data loss $n^{-1}\sum_{i=1}^{n} L(Z_i, \psi(W_i))$. Thus, in the case where $L(Z_i, \psi(W_i)) = L_2(Z_i, \psi(W_i))$, it follows that Algorithm 1 reduces to Breiman's original RF algorithm when there is no censoring and the nonparametric bootstrap is used to construct the forest.

Typically, the choice of loss function governs the estimand of interest. For example, using $L(Z_i, \psi(W_i)) = L_2(Z_i, \psi(W_i))$, the nominal focus of estimation is $E[Z|W]$. However, the structure of Algorithm 1 is flexible enough to permit other approaches since the estimator used in Step 3 need not be the same as that induced by the loss function used to construct each tree in Step 2.

4 Implementations of CURT and CURE

Perhaps the most widely used implementation of the CART algorithm is rpart (Therneau, 2014), a package available as part of the R software platform. The rpart package permits the use of $L_{ipcw}(O, \psi; \hat{G})$ for $L(Z, \psi(W)) = L_2(Z, \psi(W))$ through the incorporation of case weights (e.g., Molinaro et al., 2004). The use of case weights that are exactly zero restricts consideration of the set of covariate values as possible split points to those available on uncensored observations only. This can be handled differently by taking advantage of the ability to incorporate user-written splitting and evaluation functions directly into rpart (Therneau et al., 2014). Making use of this feature, Steingrimsson et al. (2016) proposed a special case of the CURT algorithm in Section 3.1 using $L_d(O, \psi; \hat{G}, \hat{S})$. Consistent with expectations outlined earlier, the extensive simulation study in Steingrimsson et al. (2016) demonstrates important gains in performance when using CART implemented with $L_d(O, \psi; \hat{G}, \hat{S})$ in place of $L_{ipcw}(O, \psi; \hat{G})$ for $Z = \log T$ and $L(Z, \psi(W)) = L_2(Z, \psi(W))$. Excellent performance is also seen in comparison with the algorithm of LeBlanc and Crowley (1992) as currently implemented in rpart. We refer the reader to Steingrimsson et al. (2016) for additional details on implementation, including methods used to construct the estimators $\hat{G}(\cdot|\cdot)$ and $\hat{S}(\cdot|\cdot)$. There is currently no implementation of CART that uses the Buckley-James-type loss

function $L_b(O, \psi; \hat{S})$; however, the flexibility of rpart easily permits this extension.

Like CURT, the CURE algorithm (see Algorithm 1) is applicable to CUTs for general full data loss functions. However, unlike CURT, implementation for general loss functions (e.g., $L_d(O, \psi; \hat{G}, \hat{S})$ or $L_b(O, \psi; \hat{G}, \hat{S})$) is not as straightforward due to limitations in the prevaling RF software packages. For the setting where $L(Z, \psi(W)) = L_2(Z, \psi(W))$, Section 4.1 shows how response imputation can be used to implement CURT given some implementation of CART for squared error loss (e.g., rpart). These developments also provide the necessary framework for implementing the CURE algorithm using any implementation of RF for squared error loss that employs CART trees with random feature selection as the base learner, such as the R functions `randomForest` (Liaw and Wiener, 2002) and `rfsrc` (Ishwaran and Kogalur, 2015); see Section 4.2. It is further shown there how one can generalize CURE for squared error loss to more general weighted boostrap schemes provided one has available an implementation of the RF algorithm for squared error loss that is able to incorporate case weights into the loss function calculation.

4.1 CURT With Squared Error Loss (CURT-L_2)

In this section, we show how an existing implementation of CART for $L(Z, \psi(W) = L_2(Z, \psi(W))$ can be used to implement CURT using the CUT $L_b(O, \psi; \hat{S})$ or $L_d(O, \psi; \hat{G}, \hat{S})$ for $L_2(Z, \psi(W))$. Because $L_b(O, \psi; S) = L_d(O, \psi; 1, S)$ for any choice of $S(\cdot|\cdot)$ (see Section 2.2) it suffices to demonstrate this equivalence for $L_d(O, \psi; G, S)$ with general choices of $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$. For reasons to be explained later, the results to be developed below do not extend to $L_{ipcw}(O, \psi; G, S)$, despite the fact that it can also be recovered as a special case of $L_d(O, \psi; G, S)$.

Define $L_{2,d}(O, \psi; G, S)$ as (5) calculated using $L(Z, \psi(W)) = L_2(Z, \psi(W))$. Given $\mathcal{O} = (O_1, \ldots, O_n)$, the corresponding empirical loss is $L_{2,d}(\mathcal{O}, \psi; G, S) = n^{-1} \sum_{i=1}^{n} L_{2,d}(O_i, \psi; G, S)$. The CURT-$L_2$ algorithm implemented in Steingrimsson et al. (2016) substitutes in $L_{2,d}(\mathcal{O}, \psi; G, S)$ for the empirical full data loss $n^{-1} \sum_i L_2(Z_i, \psi(W_i))$ throughout the CART algorithm and is implemented using rpart. As shown below, this same algorithm can also be implemented using an unmodified form of the CART algorithm that employs a related CUT of the form (4) for the response variable.

We begin by establishing an equivalent representation for $L_{2,d}(\mathcal{O}, \psi; G, S)$. Let

$$A_{ki}(G) = \frac{\Delta_i \widetilde{Z}_i^k}{G(\widetilde{T}_i|W_i)},\ B_{ki}(G,S) = \frac{(1-\Delta_i) m_k(\widetilde{T}_i, W_i; S)}{G(\widetilde{T}_i|W_i)},\ C_{ki}(G,S) = \int_0^{\widetilde{T}_i} \frac{m_k(u, W_i; S) d\Lambda_G(u|W_i)}{G(u|W_i)},$$

for $k = 0, 1, 2$, where

$$m_k(t, w; S) = \frac{\int_t^\infty [h(u)]^k dF(u|w)}{S(t|w)},\ k = 1, 2 \tag{8}$$

and we have defined $m_0(t, w; S) = 1$ for each $(t, w) \in \mathbf{R}^+ \times \mathcal{S}$. Writing $L_2(Z_i, \psi(W_i)) = Z_i^2 - 2Z_i\psi(W_i) + \psi^2(W_i)$, straightforward algebra gives

$$L_{2,d}(\mathcal{O}, \psi; G, S) \ = \ \frac{1}{n}\sum_{i=1}^n \left[Q^{(1)}(O_i; G, S) - 2\hat{Z}(O_i; G, S)\psi(W_i) + K(O_i; G)\psi^2(W_i) \right]. \tag{9}$$

In this expression, $K(O_i; G) = A_{0i}(G) + B_{0i}(G) - C_{0i}(G)$, $\hat{Z}(O_i; G, S) = A_{1i}(G) + B_{1i}(G, S) - C_{1i}(G, S)$, and $Q^{(1)}(O_i; G, S) = A_{2i}(G) + B_{2i}(G, S) - C_{2i}(G, S)$ for every i. That $K(O_i; G)$ does not depend on $S(\cdot|\cdot)$ follows from (i) the definition of $A_{0i}(G)$; and, (ii) the assumption $m_0(t, w; S) = 1$, which implies $B_{0i}(G, S)$ and $C_{0i}(G, S)$ are each independent of $S(\cdot|\cdot)$ for every i. In fact, we have

$$K(O_i; G) = \frac{\Delta_i}{G(\widetilde{T}_i|W_i)} + \frac{(1-\Delta_i)}{G(\widetilde{T}_i|W_i)} - \int_0^{\widetilde{T}_i} \frac{d\Lambda_G(u|W_i)}{G(u|W_i)} = 1$$

for every i under very weak conditions on $G(\cdot|\cdot)$; this follows immediately from Theorem 4.1 below upon making the identifications $D = \Delta_i$, $\tilde{t} = \tilde{T}_i$, and $w = W_i$.

Theorem 4.1. *For each $w \in \mathcal{S}$, assume $G(0|w) = 1$, $G(u|w) \geq 0$, and that $G(u|w)$ is a right-continuous, non-increasing function for $u \geq 0$ with at most a finite number of discontinuities on any finite interval. Fix $w \in \mathcal{S}$, let $\tilde{t} > 0$ be finite, suppose $G(\tilde{t}|w) > 0$, and let D be any indicator variable taking on the value 0 or 1. Then,*

$$\frac{D}{G(\tilde{t}|w)} + \frac{(1-D)}{G(\tilde{t}|w)} - \int_0^{\tilde{t}} \frac{d\Lambda_G(u|w)}{G(u|w)} = 1.$$

The proof of Theorem 4.1 may be found in the Supplementary Web Appendix (Section S.5).

The observed data quantity $\hat{Z}(O_i; G, S)$ is an example of (4). Hence, if either $G(\cdot|\cdot)$ or $S(\cdot|\cdot)$

is correctly specified, then $\hat{Z}(O_i; G, S)$ is an unbiased estimate of $\psi(W_i) = E[Z|W = W_i]$. Define the modified loss function $L^*_{2,d}(O, \psi; G, S) = n^{-1} \sum_{i=1}^n (\hat{Z}(O_i; G, S) - \psi(W_i))^2$; then, expanding the square in $L^*_{2,d}(O, \psi; G, S)$ and simplifying the resulting expression gives

$$L^*_{2,d}(\mathcal{O}, \psi; G, S) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \left[Q^{(2)}(O_i; G, S) - 2\hat{Z}(O_i; G, S)\psi(W_i) + \psi^2(W_i) \right], \tag{10}$$

where $Q^{(2)}(O_i, G, S) = [\hat{Z}(O_i; G, S)]^2$ for each i. As shown above, for any $G(\cdot|\cdot)$ satisfying the regularity conditions of Theorem 4.1, we have $K(O_i; G) = 1$ for every i in (9); as a result, (9) and (10) are identical up to a term that does not involve $\psi(\cdot)$.

The arguments above show that each of $L_{2,d}(\mathcal{O}, \psi; G, S)$ and $L^*_{2,d}(\mathcal{O}, \psi; G, S)$ takes the form $n^{-1} \sum_i L_2(O_i, \psi; G, S, Q)$, where $L_2(O_i, \psi; G, S, Q) = \psi(W_i)^2 + H(O_i; G, S)\psi(W_i) + Q(O_i; G, S)$ and the losses differ only in the specification of $Q(O_i; G, S)$. Theorem 4.2, given below and proved in the Supplementary Web Appendix (Section S.6), demonstrates that the decisions made by the CART algorithm on the basis of $L_2(O_i, \psi; G, S, Q), i = 1, \ldots, n$ do not depend on $Q(O_i; G, S), i = 1, \ldots, n$.

Theorem 4.2. *For each $i = 1, \ldots, n$, define the loss function $L_2(O_i, \psi; G, S, Q) = \psi(W_i)^2 + H(O_i; G, S)\psi(W_i) + Q(O_i; G, S)$ and assume $\max\{|H(O_i; G, S)|, |Q(O_i; G, S)|\} < \infty$. Then, the CART algorithm that uses the loss function $L_2(O, \psi; G, S, Q)$ does not depend on $Q(O; G, S)$.*

In practical terms, Theorem 4.2 implies that one can implement CURT-L_2 with (9) as described in Section 4 by applying any (full data) CART algorithm for squared error loss to the imputed dataset $\{\hat{Z}(O_i; G, S), W_i; i = 1, \ldots, n\}$. This works because all decisions made by the algorithm depend on either changes in loss or loss minimization, neither of which is affected by terms in the loss function that are independent of $\psi(\cdot)$. It should be noted that these results also do not depend on the specific nature of Z (i.e., except that it is univariate). In practice, use of either (9) and (10) does require that the empirical positivity condition

$$G(\tilde{T}_i|W_i) \geq \epsilon > 0 \tag{11}$$

holds and that (8) exists for $k = 1, 2$ for every $(\tilde{T}_i, W_i), i = 1, \ldots, n$.

It was remarked at the beginning of this section that the equivalences just established do not

extend to the case where $L_{2,ipcw}(\mathcal{O}, \psi; G)$ is used in place of $L_{2,d}(\mathcal{O}, \psi; G, S)$. The equivalence results for $L_{2,d}(\mathcal{O}, \psi; G, S)$ and $L_{2,b}(\mathcal{O}, \psi; G, S)$ rely heavily on the fact that $m_0(t, w; S) = 1$ for every (t, w) and hence that $K(O_i; G) = 1$ for every i. These identities fail in the case of $L_{2,ipcw}(\mathcal{O}, \psi; G)$ because this loss function can only be treated as a special case of $L_{2,d}(\mathcal{O}, \psi; G, S)$ in the event that $m_0(t, w; S) = 0$ for every (t, w). Under this assumption, the loss function (9) is still appropriate; however, $K(O_i; G) = \Delta_i / G(\tilde{T}_i | W_i) \neq 1$ for any i in general, showing that (9) and (10) are no longer equivalent up to terms that do not depend on $\psi(W)$.

4.2 CURE With Squared Error Loss (CURE$-L_2$)

Algorithm 1 of Section 3.2 implemented using nonparametric bootstrap sampling in combination with one of $L_{2,ipcw}(O, \psi; \hat{G})$, $L_{2,d}(O, \psi; \hat{G}, \hat{S})$, or $L_{2,b}(O, \psi; \hat{G}, \hat{S})$ generalizes Breiman's original RF algorithm to the case of a right-censored outcome. As an example of a more general bootstrap weighting scheme satisfying the conditions needed to use Algorithm 1, let $D_1, \ldots, D_n$ be i.i.d. positive random variables with finite variance that are completely independent of the observed data and define the weights $\omega_i = D_i / \sum_{j=1}^{n} D_j$ for $i = 1, \ldots, n$. In contrast to the nonparametric bootstrap, this i.i.d. weighted bootstrap (Præstgaard and Wellner, 1993) puts a positive weight on every observation in every bootstrap sample. The Bayesian bootstrap (Rubin, 1981) is a special case that is obtained when $D_1, \ldots, D_n$ are standard exponential; in this case $(\omega_1, \ldots, \omega_n)$ follow a uniform Dirichlet distribution, having the same expected value and correlation as the nonparametric bootstrap weights but a variance that is smaller by a factor of $n/(n+1)$.

In Algorithm 1, the loss function only comes into consideration in Step 2, where it governs the process of growing the unpruned regression trees used to create the ensemble predictor. For the case of $L_{2,ipcw}(O, \psi; \hat{G})$, Hothorn et al. (2006a, Sec. 3.1, p. 359) proposed a special case of Algorithm 1 that used a multinomial bootstrap with sampling weights $\hat{p}_i = w_i / \sum_{i=1}^{n} w_i, i = 1, \ldots, n$ where $w_i = \Delta_i [\hat{G}(\tilde{T}_i | W_i)]^{-1}$. This ensemble algorithm resamples only uncensored observations and uses fully grown CART trees combined with random feature selection to estimate $E[Z|W]$. Implementation of this algorithm is possible using `rfsrc` (Ishwaran and Kogalur, 2015) because this R function accepts general multinomial sampling weights. The `randomForest` function (Liaw and Wiener, 2002) could also be used here as long as the indicated multinomial bootstrap scheme

carried out externally to this R function (i.e., `randomForest` is used to build each CART tree using random feature selection, but is not directly used to build the entire forest).

Hereafter, we will use CURE$-L_2$ to denote any implementation of Algorithm 1 that uses the loss function $L_{2,d}(\mathcal{O}, \psi; \hat{G}, \hat{S})$. For $L_{2,b}(\mathcal{O}, \psi; \hat{S}) = L_{2,d}(\mathcal{O}, \psi; 1, \hat{S})$ and $L_{2,d}(\mathcal{O}, \psi; \hat{G}, \hat{S})$, Theorem 4.2 implies that CURE$-L_2$ can be implemented with any multinomial bootstrap scheme (e.g. nonparametric bootstrap) by applying any standard implementation of the RF algorithm (e.g., `randomForest` or `rfsrc`) to the imputed dataset $\{\hat{Z}(O_i; \hat{G}, \hat{S}), W_i; i = 1, \ldots, n\}$. The use of random feature selection in growing the CURT does not affect the applicability of Theorem 4.2 in justifying this useful equivalence. The use of imputation as stated works because a multinomial bootstrap allows for sampling weights that are exactly zero; hence, one only needs to modify the input dataset corresponding to each bootstrap sample and not the process used for building the unpruned trees that make up the forest. Implicit here is the assumption that $\hat{G}(\cdot|\cdot)$ and $\hat{S}(\cdot|\cdot)$ are held fixed and not recalculated for each boostrap sample.

Unlike the multinomial boostrap, a general exchangeably weighted bootstrap with strictly positive bootstrap weights cannot be implemented using a simple resampling scheme. However, the corresponding version of CURE$-L_2$ can still be implemented given an implementation of RF for squared error loss that allows case weights in calculating the loss function. Specifically, consider

$$L_{2,d,w}(\mathcal{O}, \psi; G, S) = \frac{1}{n} \sum_{i=1}^{n} \omega_i \left[Q^{(1)}(O_i; G, S) - 2\hat{Z}(O_i; G, S)\psi(W_i) + \psi^2(W_i) \right], \quad (12)$$

the case-weighted version of (9). Because $\omega_1, \ldots, \omega_n$ are generated independently of the data and each has unit mean, (9) and (12) have the same expectation. Now, consider the comparably weighted version of loss function (10), that is,

$$L^*_{2,d}(\mathcal{O}, \psi; G, S) = \frac{1}{n} \sum_{i=1}^{n} \omega_i \left[Q^{(2)}(O_i; G, S) - 2\hat{Z}(O_i; G, S)\psi(W_i) + \psi^2(W_i) \right], \quad (13)$$

where $Q^{(2)}(O_i, G, S) = [\hat{Z}(O_i; G, S)]^2$ for each i. It follows that (10) and (13) also have the same expectation and, in addition, that the weighted losses (13) and (12) are equal up to terms that do not involve $\psi(\cdot)$. Observe that these results hold as stated even when $G(\cdot|\cdot)$ and/or $S(\cdot|\cdot)$ are estimated, provided that neither depends on $\omega_1, \ldots, \omega_n$. An easy generalization of the arguments in Section 4.1

now shows that CURE$-L_2$ can be implemented using the imputed dataset $\{\hat{Z}(O_i;\hat{G},\hat{S}),W_i;i=1,\ldots,n\}$ with case weights $\omega_1,\ldots,\omega_n$. We are not currently aware of an implementation of the RF algorithm that accepts case weights in the manner required above. For the simulation study of Section 5 and data analysis of Section 6, we have therefore extended the randomForest package to permit case weights in the calculation of the loss function (and associated estimators) and we use this modified RF algorithm to implement CURE$-L_2$ for the iid-weighted bootstrap for several different possible choices of the loss functions $L_{2,b}(O,\psi;\hat{S})$ and $L_{2,d}(O,\psi;\hat{G},\hat{S})$.

5 Simulations

In this section, we use simulation to compare the performance of several CURE$-L_2$ algorithms to several available implementations of survival forests. The following subsections describe the simulation settings used (Section 5.1) and the choices made for implementing the CURE$-L_2$ algorithm (Sections 5.2 and 5.3). Section 5.4 summarizes the results; further results are also provided the Supplementary Web Appendix, where we also revisit the simulation study conducted in Steingrimsson et al. (2016) and compare the performance of the CURT$-L_2$ algorithm using the Buckley-James (see (6)) and doubly robust (see (5)) loss functions, with both being implemented using the imputation approach described in Section 4.1.

5.1 Simulation Parameters

The simulation settings used here are very similar to Settings $1-4$ in Zhu and Kosorok (2012). The four settings considered are respectively described below:

1. Each simulated dataset is created using 300 independent observations where the covariate vector $(W_1,\ldots,W_{25})$ is multivariate normal with mean zero and covariance matrix with element (i,j) equal to $0.9^{|i-j|}$. Survival times are simulated from an exponential distribution with mean $\mu = e^{0.1\sum_{i=11}^{20} W_i}$ (i.e., a proportional hazards model) and the censoring distribution is exponential with mean chosen to get approximately 30% censoring.

2. Each simulated dataset is created using 200 independent observations where the covariate vector $(W_1,\ldots,W_{25})$ consists of 25 i.i.d. uniform random variables on the interval $[0,1]$. The

survival times follow an exponential distribution with mean $\mu = \sin(W_1\pi) + 2|W_2 - 0.5| + W_3^3$. Censoring is uniform on $[0, 6]$ which results in approximately 24% censoring. Here, the proportional hazards assumption is mildly violated.

3. Each simulated dataset is created using 300 independent observations where the covariates $(W_1, \ldots, W_{25})$ are multivariate normal with mean zero and covariance matrix with element (i, j) given by $0.75^{|i-j|}$. Survival times are gamma distributed with shape parameter

$$\mu = 0.5 + 0.3 \left| \sum_{i=11}^{15} W_i \right|$$

and scale parameter 2. Censoring times are uniform on $[0, 15]$ which results in approximately 20% censoring. Here, the proportional hazards assumption is strongly violated.

4. Each simulated dataset is created using 300 independent observations where the covariates $(W_1, \ldots, W_{25})$ are multivariate normal with mean zero and covariance matrix where element (i, j) is given by $0.75^{|i-j|}$. Survival times are simulated according to a log-normal distribution with mean

$$\mu = 0.1 \left| \sum_{i=1}^{5} W_i \right| + 0.1 \left| \sum_{i=21}^{25} W_i \right|.$$

Censoring times are log-normal with mean $\mu + 0.5$ and scale parameter one, and the censoring rate is approximately 32%. Here, the underlying censoring distribution depends on covariates.

5.2 Squared Error Loss Functions

With time-to-event data, a survival probability of the form $P(T > t|W)$ is most often of interest. The output from any CURT or CURE algorithm can be post-processed to generate estimators for $P(T > t|W)$ with right-censored outcomes. For example, regardless of the underlying loss function used, one can simply use a Kaplan-Meier estimator applied to the observations falling into each terminal node of each CURT estimate that is used to build the ensemble; see Step 3 of Algorithm 1. Other related approaches (e.g., parametric survival models) can also be used to process these terminal nodes. Alternatively, the loss function used by the CURE$-L_2$ algorithm can be chosen to focus specifically on estimation of $P(T > t|W)$ for some given $t > 0$; see, for example, the Brier

loss function of Section 2.3. In the next two subsections, we describe examples of each approach to be used in this simulation study.

5.2.1 Ensembles from Loss Functions that Estimate $E[\log T|W]$

With uncensored data and a continuous outcome, the most common loss function used in connection with both the CART and RF algorithms is the L_2 loss. Taking $Z = \log T$, the relevant full data loss function is $(\log T - \psi(W))^2$ and the nominal focus of estimation becomes $\psi_0(W) = E[\log T|W]$ whether CURT-L_2 or CURE$-L_2$ is used. Equation (9) gives the corresponding doubly robust L_2 loss $L_{2,d}(\mathcal{O}, \psi; G, S)$ for suitable choices of $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$; the Buckley-James L_2 loss is given by $L_{2,d}(\mathcal{O}, \psi; 1, S)$. Further details on the calculation of $L_{2,d}(\mathcal{O}, \psi; G, S)$ for $Z = \log T$ may be found in Steingrimsson et al. (2016). In the results summarized in Section 5.4, we use *L2* to denote the CURE$-L_2$ algorithm that (i) employs $L_{2,d}(\mathcal{O}, \psi; \hat{G}, \hat{S})$ to construct a tree using random feature selection for each bootstrapped dataset ; and, (ii) estimates $P(T > t|W)$ by using Kaplan-Meier estimators in each terminal node (see Step 3 of Algorithm 1). Here, the nonparametric bootstrap is used to generate the ensemble. We use *L2 BJ* to denotes the same procedure, but where $L_{2,b}(\mathcal{O}, \psi; \hat{S})$ replaces $L_{2,d}(\mathcal{O}, \psi; \hat{G}, \hat{S})$. As noted earlier in Section 4.2, it is also possible to implement such a procedure using the IPCW loss function $L_{2,ipcw}(\mathcal{O}, \psi; \hat{G})$. However, the results of Steingrimsson et al. (2016) demonstrate that trees built using this loss function tend to exhibit substantially larger errors than do $L_{2,d}(\mathcal{O}, \psi; \hat{G}, \hat{S})$; consequently, we do not consider the use of this loss function further in connection with ensemble estimators.

5.2.2 Ensembles from Loss Functions that Estimate $P(T > t|W)$

Let $L_{t;2}(T, \psi_t(W)) = (I(T > t) - \psi_t(W))^2$ denote the (full data) Brier loss function; see Section 2.3. Minimizing this squared error loss function induces a simple estimator for $S_0(t|W) = P(T > t|W)$. The IPCW Brier loss function

$$L_{t;2,ipcw}(O, \psi_t; G) = \frac{\Delta(I(\tilde{T} > t) - \psi_t(W))^2}{G(\tilde{T}|W)}$$

is also a CUT for $L_{t;2}(T, \psi_t(W))$ when $G(\cdot|\cdot) = G_0(\cdot|\cdot)$, and leads to one possible observed data estimator for $S_0(t|W) = P(T > t|W)$ when integrated into a tree or ensemble procedure.

Graf et al. (1999) proposed a "time-dependent" Brier loss function with right-censored outcome data that also requires the specification (or estimation) of $G_0(\cdot|\cdot)$. Calculations similar to Lostritto et al. (2012) show that this loss function is constructed from terms of the form

$$L_{t;2,ipcw}(O(t),\psi_t;G) = \frac{\Delta(t)(I(\tilde{T}(t)>t)-\psi_t(W))^2}{G(\tilde{T}(t)|W)} \equiv \frac{\Delta(t)(I(\tilde{T}>t)-\psi_t(W))^2}{G(\tilde{T}(t)|W)},$$

where $O(t)=\{\tilde{T}(t),\Delta(t),W\}$, $\tilde{T}(t)=\min(T,t,C)$ and $\Delta(t)=I(\tilde{T}(t)\leq C)$ and the stated equivalence follows from the fact that $L_{t;2}(T(t),\psi_t(W))=L_{t;2}(T,\psi_t(W))$. Similarly to $L_{t;2,ipcw}(O,\psi_t;G)$, straightforward calculations show that $L_{t;2,ipcw}(O(t),\psi_t;G)$ is a CUT for $L_{t;2}(T,\psi_t(W))$ when $G(\cdot|\cdot)=G_0(\cdot|\cdot)$. The value of $I(T>t)$ can be unambiguously determined when $\Delta(t)=1$, which occurs if either $\Delta=1$ or if $\Delta=0$ and $\tilde{T}>t$. As a result, $L_{t;2,ipcw}(O(t),\psi_t;G)$ uses more of the available information in the data for estimating $S_0(t|W)$ when compared to $L_{t;2,ipcw}(O,\psi_t;G)$.

Mathematically, the equivalence $L_{t;2}(T,\psi_t(W))=L_{t;2}(T(t),\psi_t(W))$ combined with the construction of $L_{t;2,ipcw}(O(t),\psi;G)$ suggests applying (5) to the observed data structure $O(t)$ with $L(Z,\psi(W))=L_{t;2}(T(t),\psi(W))$; simplifying the resulting expression, we obtain

$$\frac{\Delta(t)(I(T>t)-\psi_t(W))^2}{G(\tilde{T}(t)|W)}+\frac{(1-\Delta(t))m_{t;2}(\tilde{T}(t),W;S)}{G(\tilde{T}(t)|W)}-\int_0^{\tilde{T}(t)}\frac{m_{t;2}(u,W;S)}{G(u|W)}d\Lambda_G(u|W) \quad (14)$$

where $m_{t;2}(u,w;S)=E_S[(I(T>t)-\psi_t(W))^2|T>u,W=w]$ for any proper survival function $S(\cdot|\cdot)$. The leading term in (14) is $L_{t;2,ipcw}(O(t),\psi_t;G)$ and it can be shown that (14) and $L_{t;2}(T,\psi_t(W))$ have equal conditional (i.e., given W) expectations if either $G(\cdot|\cdot)=G_0(\cdot|\cdot)$ or $S(\cdot|\cdot)=S_0(\cdot|\cdot)$; hence, it is a doubly robust CUT for $L_{t;2}(T,\psi_t(W))$. Similarly, the direct application of (6) to $O(t)$ combined with the equivalence $L_{t;2}(T,\psi_t(W))=L_{t;2}(T(t),\psi_t(W))$ gives

$$\Delta(t)(I(T>t)-\psi_t(W))^2+(1-\Delta(t))m_{t;2}(\tilde{T}(t),W;S) \quad (15)$$

as a CUT for $L_{t;2}(T,\psi_t(W))$ when $S(\cdot|\cdot)=S_0(\cdot|\cdot)$.

We respectively refer to (14) and (15) as the doubly robust and Buckley-James Brier loss functions, denoted respectively by $L_{t;2,dr}(O(t),\psi_t;G,S)$ and $L_{t;2,b}(O(t),\psi_t;S)$. Because (14) and (15) are each CUTs for the Brier loss function (i.e., squared error loss) and focus directly on

estimating $S_0(t|W)$, the results of Section 4.1 and Theorem 4.2 can be used to justify implementing the corresponding CURT-L_2 algorithm by applying CART with squared error loss to the imputed dataset $\{\hat{Z}(O_i(t); G, S), W_i; i = 1, \ldots, n\}$ where $\hat{Z}(O_i(t); G, S) = A_{1i}(G) + B_{1i}(G, S) - C_{1i}(G, S)$ is computed by replacing $(\tilde{Z}_i, \Delta_i)$ with $(I(\tilde{T}_i > t), \Delta_i(t))$. Section 4.2 gives the corresponding recipe for implementing CURE$-L_2$ with either (14) or (15). In the results summarized in Section 5.4, *Brier* and *Brier BJ* denote the CURE$-L_2$ algorithms respectively implemented using (14) and (15). In both cases, the terminal node estimators used to construct the ensemble predictor are naturally induced by the choice of loss function.

5.3 Estimating $S(\cdot|\cdot)$ and $G(\cdot|\cdot)$

The CURE$-L_2$ methods *L2*, *L2 BJ*, *Brier*, and *Brier BJ* each require specifying estimators $\hat{S}(t|w)$ and/or $\hat{G}(t|w)$ for $S_0(r|w) = P(T > r|W = w)$ and $G_0(r|w) = P(C > r|W = w)$. These required estimators are calculated prior to running these learning algorithms and not re-calculated for each set of bootstrap weights.

Many methods are available for estimating a conditional survivor function. Preserving the double robustness property suggests avoidance of IPCW estimators. In building survival trees using a special case of the CURT algorithm, Steingrimsson et al. (2016) considered estimators for $S(\cdot|\cdot)$ respectively derived from Cox regression, survival regression tree models, random survival forests, and parametric accelerated failure time (AFT) models when computing the augmented loss function. Although the performance of the doubly robust methods differed noticeably from IPCW, the method for estimating $S(\cdot|\cdot)$ otherwise made little difference among doubly robust methods in the chosen performance measures. Consequently, we use $m_1(u, w; \hat{S})$ in (8) with $\hat{S}(t|W_i), i = 1, \ldots, n$ estimated using the random survival forests (RSF) procedure as proposed by Ishwaran et al. (2008) and implemented in `rfsrc`.

In Settings 1$-$3, the censoring distribution is independent of covariates. Because the dependency of the censoring distribution on covariates can be checked empirically, the censoring distribution is estimated using the product-limit estimator. To ensure that the estimated censoring probabilities remain bounded away from zero, a sample-dependent truncation time $\hat{\vartheta}$ is set such that the proportion of observed times exceeding $\hat{\vartheta}$ is 10%; "Method 2" truncation as described in Steingrimsson

et al. (2016) is then used. In short, times $\tilde{T}_i$ exceeding $\hat{\vartheta}$ are designated as failures and $\hat{G}(\tilde{T}_i|W_i)$ and $\hat{G}(u|W_i)$ are respectively replaced by $\hat{G}(\hat{\vartheta} \wedge \tilde{T}_i|W_i)$ and $\hat{G}(\hat{\vartheta} \wedge u|W_i)$ in calculating $\hat{Z}(O_i; \hat{G}, \hat{S})$ above, but survival times are not otherwise modified in the remainder of the calculations. As shown in Steingrimsson et al. (2016), this typically performs better than the standard approach to truncation (i.e., truncating all follow-up times that exceed $\hat{\vartheta}$ and treating each as uncensored). The doubly robust Brier loss of Section 5.2.2 at time t automatically induces this kind of truncation with $\hat{\vartheta} = t$; hence as long as $\hat{\vartheta}$ is larger than the time-point used to calculate the Brier loss, "Method 2" truncation has no discernible effect.

In Setting 4, the censoring distribution depends on the covariates and the censoring distribution is modeled using `rfsrc` to reduce the possibility of misspecification. The truncation methods described in Steingrimsson et al. (2016) only protect against large failure times; extreme covariate values can still lead to small censoring probabilities and hence thus large inverse probability of censoring weights. A commonly accepted method of truncation that protects against both extreme failure time and covariate values at the expense of introducing some bias is to truncate the weights directly. Specifically, by truncating the estimated censoring probabilities using $\max(\hat{G}(u|W), 0.05)$, each inverse probability of censoring weight will be no more than 20, limiting the influence of each observation on the loss function. Because $\hat{G}(\tilde{T}(t)|W) \geq \hat{G}(\tilde{T}|W)$, this kind of truncation scheme affects the doubly robust Brier loss less compared to the L_2 loss.

5.4 Simulation Results

Settings $1-4$ are used to compare the performance of the CURE$-L_2$ algorithms *L2*, *L2 BJ*, *Brier*, and *Brier BJ* to other implementations of survival forests. We focus on estimation of survival probabilities of the form $P(T > t|W)$ for a given fixed time-point t. In this section we only focus on CURE$-L_2$ algorithms using the nonparametric bootstrap as it has the advantage of being easily implemented using existing RF software with uncensored outcomes. In Supplementary Web Appendix S.1.2 we present comparisons of these algorithms in Settings $1-4$ to CURE$-L_2$ algorithms fit using the Bayesian bootstrap. The Bayesian bootstrap CURE$-L_2$ procedure is fit by extending the capabilities of `randomForest` (Liaw and Wiener, 2002) to handle arbitrary nonnegative case weights in calculating the loss function. In general, Figures S-3-S-5 in Supplementary Web

Appendix S.1.2 demonstrate comparable performance between the nonparametric and Bayesian bootstraps in all settings at all quantiles for all combinations of loss functions and CUTs. One interesting trend that does emerge is that the nonparametric bootstrap tends to perform the same as, or slightly better than, the Bayesian bootstrap when used in connection with *Brier* and *Brier BJ*. However, the comparative performance in the case of *L2* and *L2 BJ* depends more on both the setting and time point, with no similar trend that emerges.

We will compare the results of *L2*, *L2 BJ*, *Brier*, and *Brier BJ* to three currently available ensemble algorithms for survivor function prediction: the default method for censored data in the `party` package (Hothorn et al., 2010); the default method for censored data in the `randomForestSRC` package (Ishwaran and Kogalur, 2015); and, recursively imputed survival trees (RIST; Zhu and Kosorok, 2012). The default method in the `party` package constructs a survival ensemble where conditional inference trees based on the two sample log-rank statistic are used in place of CART trees as the base learner (Hothorn et al., 2006b). The default method in the `randomForestSRC` package implements RSF as proposed in Ishwaran et al. (2008) and relies on the logrank statistic for splitting decisions. The RIST code is currently available from `https://sites.google.com/site/teazrq/software`.

All of these algorithms require specifying several tuning parameters. The tuning parameters for the RIST algorithm are chosen as in the example code provided by the authors with the exception that the length of the study parameter is chosen larger than the largest survival time. This includes using two fold recursively imputed survival trees with 50 trees in each fold and $mtry = \lceil \sqrt{p} \rceil$. For all other methods $mtry = \lceil \sqrt{p} \rceil$ and the number of trees is set to 1000. All other tuning parameters are selected as the default in the corresponding `R` functions.

Each survival forest procedure predicts $P(T > t|W)$ on an independent test set consisting of 1000 observations simulated from the full data distribution with t respectively chosen as the 25, 50 and 75th quantile of the marginal failure time distribution. For all four simulation settings the mean squared estimation error is calculated as $0.001 \times \sum_{i=1}^{1000} (\hat{S}(t|W_i) - S_0(t|W_i))^2$, where $\hat{S}(t|W)$ is the prediction from the algorithm and $S_0(t|W)$ is the true conditional survival curve. Boxplots from 1000 simulations for t equal to the median of the marginal survival distribution for the four different simulation settings are shown in Figure 1. The corresponding plots for t equal to the 25 and 75th quantile of the marginal survival distributions are given in Figures S-1 and S-2 in Supplementary Web Appendix S.1. In all plots, *L2*, *L2 BJ*, *Brier*, and *Brier BJ* denote the methods described

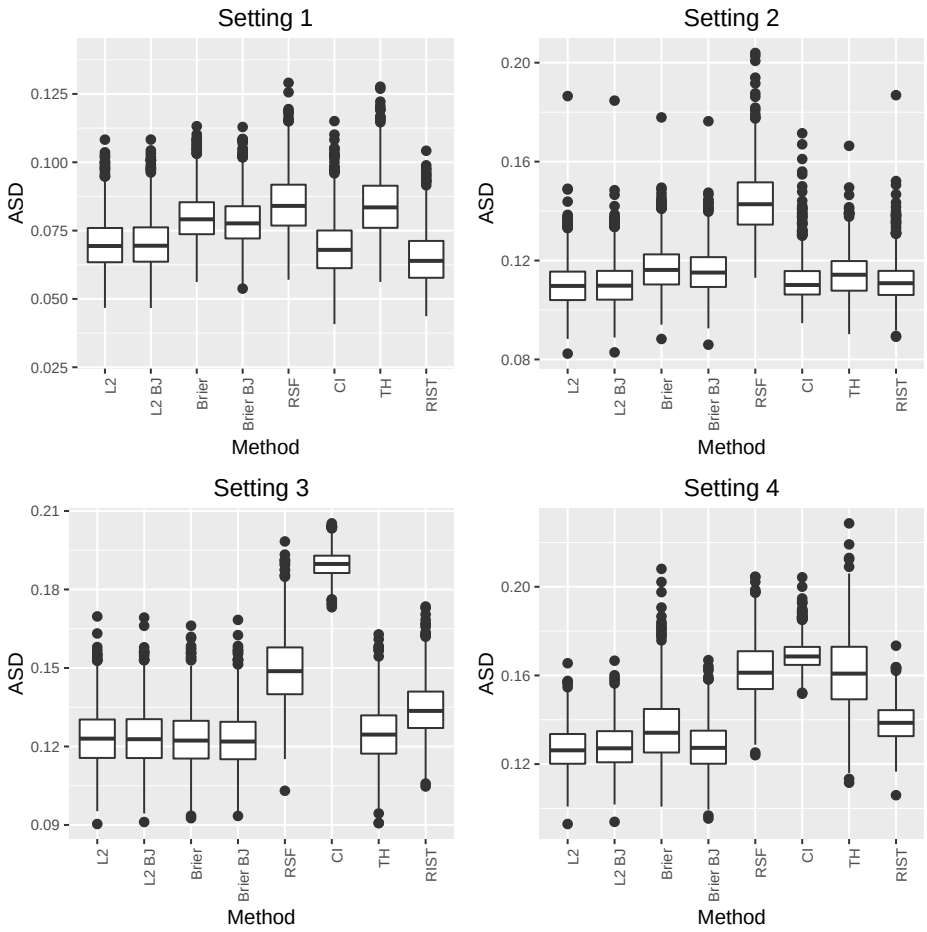


Figure 1: Boxplots of MSE estimated at the $50th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` functions. *RIST* is the recursively imputed survival trees algorithm.

previously, *RSF* is the default method for `rfsrc`, *CI* is the default method in the `party` package (i.e., the conditional inference forest), and *RIST* refers to the recursively imputed survival trees.

The main results from Figure 1 are summarized below.

- Overall, the CURE$-L_2$ algorithms *L2*, *L2 BJ*, *Brier* and *Brier BJ* perform similarly in all simulation settings, with *L2* and *L2 BJ* showing the best overall performance. The *RIST* algorithm is also a strong performer, doing the best in Setting 1 and remaining competitive in all others. Using currently available software, the CURE$-L_2$ algorithms run considerably faster when compared to *RIST*, even when accounting for the calculations needed to compute

the augmentation terms needed for the Buckley-James and doubly robust loss functions.

- Settings $1-3$ are used to illustrate the performance under different degrees of misspecification of the proportional hazard assumption (correctly specified, mildly misspecified, and a more severe misspecification). Figure 1 shows that as the severity of the misspecification increases the relative performance of the methods not utilizing log-rank based splitting statistics (i.e., the CURE$-L_2$ algorithms) becomes better compared to the algorithms where splitting decisions utilize such statistics (i.e., *RSF, CI, RIST*).

The use of the L_2 loss function for predicting $E(\log T|W)$ can be viewed as making use of information across time, whereas the Brier loss for predicting $P(T > t|W)$ arguably makes more limited use of the available data. This may help to explain why the CURE$-L_2$ algorithms tend to have somewhat lower MSE.

6 Applications to Four Public-Use Datasets

In this section we evaluate the performance of the CURE$-L_2$ algorithms on four datasets:

1. *TRACE Study Group Data:* This dataset consists of 1878 subjects that were randomly sampled from 6600 patients and is included in the `R` package `timereg`. The event of interest is death from acute myocardial infarction. Subjects that died from other causes or were alive when they left the study were considered censored. Two observations with an undefined censoring status were removed from the dataset. Information on gender, age, diabetes status, if clinical heart pump failure (CHF) was present, if the patient had ventricular fibrillation, and a measure of the heart pumping effect was also collected. As in Steingrimsson et al. (2016), who analyzed the dataset using doubly robust survival trees (i.e., an example of CURT), we focus on the subset of patients surviving past 30 days. The final dataset analyzed consists of 1689 patients with a 53.8% censoring rate.

2. *Worcester Heart Attack Study Data:* This dataset consists of 500 patients followed after hospital admission for acute myocardial infarction (AMI). Data were collected during thirteen 1-year periods beginning in 1975 and extending through 2001 on all AMI patients admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. The event

of interest is overall survival and the censoring rate is 57%. There are 14 covariates, which are listed in Table S-2 in Supplementary Web Appendix S.3. These publicly available data are available from `www.umass.edu/statdata/statdata/data/` and this dataset consists of data from the years 1997, 1999, and 2001. The data has been analyzed for illustrative purposes in Hosmer et al. (2008).

3. *Netherlands Breast Cancer Study Data:* This dataset consists of 144 lymph node positive breast cancer patients and is included in the `R` package `penalized`. The event of interest is time to distant metastasis; subjects who were alive at the end of study, died from causes other than breast cancer, had recurrence of local or regional disease, or developed a second primary cancer were considered censored. The clinical factors measured are: number of affected lymph nodes, age, diameter of the tumor, estrogen receptor status, and grade of the tumor. Additionally, the dataset includes gene expression information for 70 genes that are used to build the prognostic model in both Van't Veer et al. (2002) and Van De Vijver et al. (2002). The censoring rate is 67%.

4. *R-Chop Study Data:* This dataset consists of 233 patients with diffuse large B-cell lymphoma undergoing R-Chop treatment and followed until death. The dataset is publicly available in the `R` package `bujar`. The covariate vector consists of microarray data, with 3833 probe sets preselected from a set of 54675 probe sets as described in Wang and Wang (2010). The censoring rate is 74%.

We first compare the prediction performance of the CURE$-L_2$ algorithms to the default methods in the `randomForestSRC` and `party` package. The *RIST* method is only included in the comparison for the third dataset because its current software implementation is unable to handle categorical predictors. All algorithms are used to predict $P(T > t|W)$, where t is set equal to 3 years; respectively, the corresponding marginal survival probabilities (i.e., estimated using a Kaplan Meier curve) are 0.73, 0.61, 0.83, and 0.73 for the TRACE, Worcester, Netherlands, and R-Chop datasets. The estimator for $S_0(\cdot|\cdot)$ used in calculating the augmentation terms in the CURE$-L_2$ algorithms is obtained using the random survival forest procedure; the doubly robust methods employ a Kaplan-Meier estimator for $G_0(\cdot|\cdot)$ and use Method 2 truncation as described in Section 5.3. Prediction performance is evaluated using a cross-validated version of the censored data Brier score

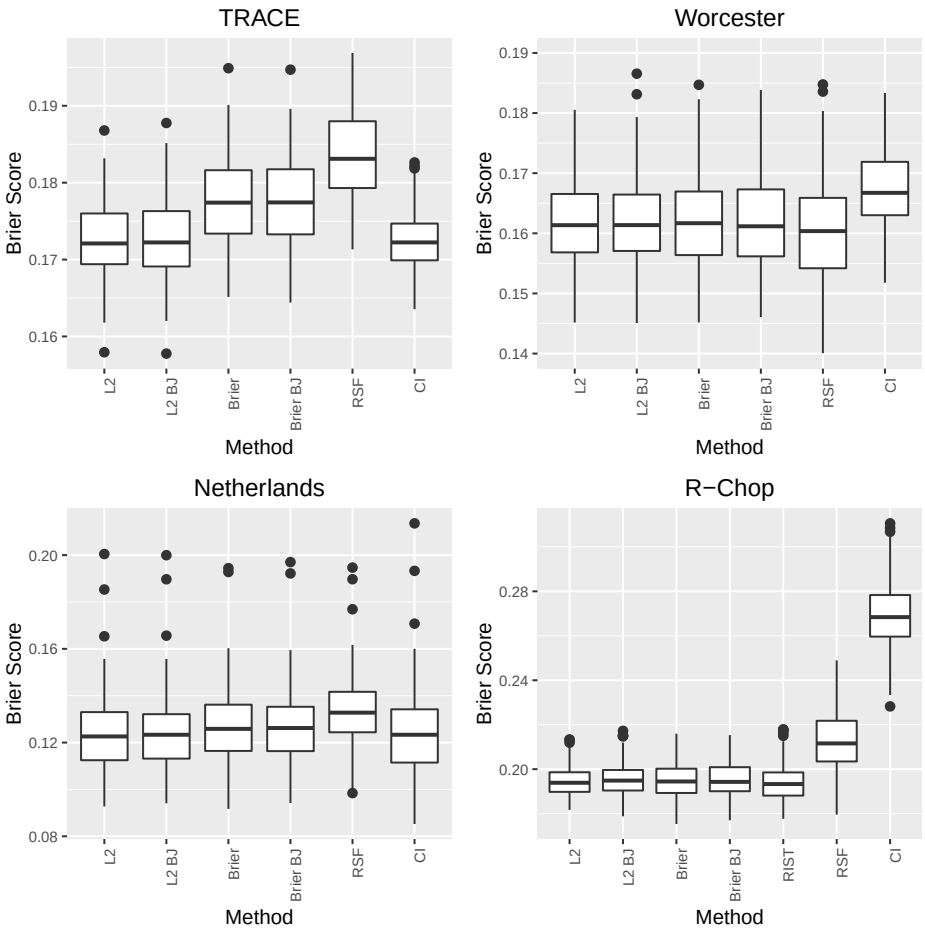


Figure 2: Censored data Brier Score at $t = 3$ years for the four datasets described in Section 6; lower values indicate better prediction accuracy. *L2*, *L2 BJ*, *Brier*, and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. *RSF* and *CI* are the default methods for the `rfsrc` and `cforest` `R` functions. *RIST* is the recursively imputed survival trees algorithm.

of Graf et al. (1999, Sec. 6); this MSE-type measure is calculated using two fold cross-validation procedure that approximately balances censoring rates in the training and test sets. Figure 2 shows boxplots of the censored data Brier score for 200 different splits into test and training sets for the four datasets; lower values indicate better performance. The results show that *L2* and *L2 BJ* have the overall best performance, performing similarly to or better than all other methods for all four datasets. The *Brier* and *Brier BJ* algorithms also show good overall performance.

Variable importance measures (VIMPs) are commonly used to evaluate the importance of each variable in the predictions generated by an ensemble algorithm. With the nonparametric bootstrap,

every bootstrap sample excludes a subset of subjects (i.e., the out-of-bag, or OOB, data). The VIMP measure for a covariate j is often calculated as the increase in L_2 prediction error compared to that for the original forest, where the increase is calculated under a setting in which the relationship between covariate j and the response is destroyed; see, for example, Breiman (2001) and Ishwaran et al. (2008). The method of Breiman (2001) involves permuting the observed values of covariate j in each OOB sample before evaluating the increase in prediction error; see Section S.2 for further details on the calculation of this OOB prediction error measure for the case of L_2 loss and the corresponding VIMP. Theorem S.2.1 in Supplementary Web Appendix S.2 shows that calculating the VIMP of Breiman (2001) using the imputed dataset $\{\hat{Z}(O_i; G, S), W_i; i = 1, \ldots, n\}$ is identical to the version that would be calculated if the (unobserved full data) L_2 loss were replaced by the CUT for this loss function that corresponds to $\hat{Z}(O_i; G, S),\ i = 1, \ldots, n$.

Ishwaran et al. (2010) proposed an alternative VIMP measure based on the intuitively sensible idea that splits made early in the individual trees in the forest are more likely to be important predictors. In particular, if the depth of a given node in a given tree is defined as the number of splits that are made between that node and the root node, then one can determine the minimal observed depth for any given variable by calculating the depth for all nodes that split on that particular variable. This calculation can be done for each variable in each tree in the ensemble; the resulting minimal depth VIMP for each variable is then calculated as the average of the minimal observed depths for that variable over all trees. Variables with lower average minimal depth are considered more influential. Because the minimal depth VIMPs do not require the presence of an OOB sample, such measures can be calculated for CURE$-L_2$ algorithms using more general bootstrap schemes, such as the i.i.d.-weighted or Bayesian bootstrap.

Below, we illustrate the use of the minimal depth VIMP measure using the TRACE data. Table 1 shows these measures calculated using the four CURE$-L_2$ algorithms and the *RSF* method. One of the main findings in Jensen et al. (1997) was that the effect of ventricular fibrillation, an acute emergency condition, vanished when analyzing the data consisting of subjects surviving beyond 30 days. The results in Table 1 are arranged in decreasing importance as measured by the *RSF* VIMP values and support this conclusion as ventricular fibrillation has the highest minimal depth VIMP of all variables for all algorithms (i.e., judged as the least important). Age and CHF have the two lowest VIMP measures for all methods, a result consistent with those in Steingrimsson et al. (2016),

where all trees are observed to split on age and (with one exception) also on CHF. Diabetes was the only other variable split on in the analysis in Steingrimsson et al. (2016) and it is seen to be the third most influential variable in three of the four CURE$-L_2$ algorithms; the *BJ L2* algorithm switches the order of diabetes and gender, factors that are strongly associated with each other. The corresponding results for the OOB prediction error VIMPs are presented in Supplementary Web Appendix S.3 and lead to the same conclusions.

	L2	L2 BJ	Brier	Brier BJ	RSF
Age	0.90	1.13	0.89	0.92	0.82
CHF	1.02	0.96	1.07	0.98	1.11
Diabetes	1.67	2.04	1.51	1.51	1.45
Gender	1.99	1.15	1.90	1.98	2.02
VF	2.14	2.17	2.27	2.27	2.43

Table 1: Minimal depth variable importance measures for the TRACE data; lower values indicate more influential variables. *Brier* and *L2* refer to the loss function used. *BJ* refers to the Buckley-James transformation. *RSF* is the default method in the `randomForestSRC` package. CHF stands for clinical heart pump failure and VF stands for ventricular fibrillation.

Further results for the remaining three studies may be found in Supplementary Web Appendix S.3. For example, we show that earlier results on the importance of certain predictors in the Worcester study are supported by the VIMP measures; see Table S-2. In addition, Figure S-8 compares the prediction performance of the CURE$-L_2$ algorithms fit using only clinical factors and using both clinical factors and gene expression measurements for the Netherlands study. The results demonstrate that a substantial improvement in prediction accuracy is achieved when the gene expression data are included, supporting one of the main conclusions in Van't Veer et al. (2002). Natural killer cell counts have been shown to be important predictors for survival in diffuse large B-cell lymphoma patients. As further discussed in Supplementary Web Appendix S.3, both the CURE$-L_2$ algorithms fit to the R-Chop data identify a probe set that is a known natural killer cell receptor as being the most influential probe set.

7 Discussion

This paper makes several contributions to the literature. We extend the theory of censoring unbiased transformations in a substantial way and establish some useful efficiency results. This theory

is applied to the problem of risk estimation, leading to the so-called class of censoring unbiased loss functions. These results are subsequently used to extend the CART and RF algorithms to the case of right-censored outcome data by replacing the full data loss by doubly robust and Buckley-James observed data loss functions. For the special case of the L_2 loss function, we show that a certain form of response imputation can be used to implement these new algorithms using standard software for uncensored responses. The proposed methods are shown to perform well compared to several existing ensemble methods both in simulations and when predicting risk using four different public-use datasets.

The use of the L_2 loss function for predicting $E(\log T|W)$ can be viewed as making use of information across time, whereas the Brier loss for predicting $P(T > t|W)$ arguably makes more limited use of the available data. This may help to explain why the CURE$-L_2$ algorithms based on the former tend to have somewhat lower MSE than those based on the latter. The use of a composite Brier loss function incorporating information for estimating $P(T > t|W)$ using several different choices for t may improve performance further. However, it is unclear whether one can use imputation methods like those introduced earlier in combination with existing software to implement such methods; we intend to explore this in future work. Other potentially interesting future research directions include: extensions to more complex data-structures such as multivariate outcomes, competing risks, missing covariate data and more complex sampling schemes (i.e. case-cohort or nested case-control designs); studying the performance of iterated versions of this algorithm, where the conditional expectations required for computing the doubly robust and Buckley James loss functions are updated using the latest ensemble predictor (or possibly updated dynamically in batches); and, deriving asymptotic properties of the CURE algorithm (or certain special cases, such as CURE$-L_2$), such as extending the consistency results in Scornet et al. (2015) or developing methods to calculate asymptotically valid confidence intervals for the predictions from the CURE algorithm (e.g. Mentch and Hooker, 2016).

The theory justifying the use of censoring unbiased loss functions is not restricted to the CART algorithm or to ensemble methods that use CART trees as building blocks. For example, it is possible to use the results in this paper in connection with other recursive partitioning methods (e.g., the partDSA algorithm; see Lostritto et al., 2012), which builds a predictor by recursively partitioning the covariate space using both 'and' and 'or' statements. Implementation using im-

puted response data as done here in the case of L_2 loss remains possible more generally in cases where model building decisions do not depend on the absolute level of loss (e.g., relative change, loss minimization, etcetera). Finally, we note that the doubly robust Brier loss function (14) and the Buckley-James Brier loss function (15) are potentially of interest in settings that extend outside the scope of this paper. For example, similarly to Graf et al. (1999), one may find these methods useful in validating prognostic models (Gerds et al., 2008; Kim, 2009; Collins and Altman, 2010).

References

Breiman, L. "Bagging Predictors." *Machine Learning*, 24(2):123–140 (1996).

—. "Random forests." *Machine learning*, 45(1):5–32 (2001).

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group: The Wadsworth Statistics/Probability Series (1984).

Brier, G. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 78:1–3 (1950).

Brooks, A. G., Posch, P. E., Scorzelli, C. J., Borrego, F., and Coligan, J. E. "NKG2A Complexed with CD94 Defines a Novel Inhibitory Natural Killer Cell Receptor." *The Journal of Experimental Medicine*, 185(4):795–800 (1997).

Buckley, J. and James, I. R. "Linear regression with censored data." *Biometrika*, 66:429–436 (1979).

Collins, G. S. and Altman, D. G. "An Independent and External Validation of QRISK2 Cardiovascular Disease Risk Score: a Prospective Open Cohort Study." *BMJ*, 340:c2442 (2010).

Davis, R. B. and Anderson, J. R. "Exponential Survival Trees." *Statistics in Medicine*, 8(8):947–961 (1989).

Fan, J. and Gijbels, I. "Censored Regression: Local Linear Approximations and their Applications." *Journal of the American Statistical Association*, 89(426):560–570 (1994).

—. *Local Polynomial Modeling and Its Applications*. Chapman and Hall (1996).

Fitzgibbons, T. P., Hardy, O. T., Lessard, D., Gore, J. M., Yarzebski, J., and Goldberg, R. J. "Body Mass Index, Treatment Practices, and Mortality in Patients with Acute Heart Failure." *Coronary Artery Disease*, 20(8):536 (2009).

Gerds, T. A., Cai, T., and Schumacher, M. "The Performance of Risk Prediction Models." *Biometrical Journal*, 50(4):457–479 (2008).

Goldberg, R. J., Gore, J. M., Gurwitz, J. H., Alpert, J. S., Brady, P., Strohsnitter, W., Chen, Z., and Dalen, J. E. "The Impact of Age on the Incidence and Prognosis of Initial Acute Myocardial Infarction: the Worcester Heart Attack Study." *American Heart Journal*, 117(3):543–549 (1989).

Gordon, L. and Olshen, R. "Tree-structured Survival Analysis." *Cancer Treatment Reports*, 69(10):1065–1069 (1985).

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. "Assessment and Comparison of Prognostic Classification Schemes for Survival Data." *Statistics in Medicine*, 18(17-18):2529–2545 (1999).

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learnin*. New York: Springer (2009).

Hosmer, D., Lemeshow, S., and May, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data, Second Edition*. New York: John Wiley and Sons (2008).

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. "Survival Ensembles." *Biostatistics*, 7(3):355–373 (2006a).

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. "Party: A Laboratory for Recursive Partytioning." (2010).

Hothorn, T., Hornik, K., and Zeileis, A. "Unbiased recursive partitioning: A conditional inference framework." *Journal of Computational and Graphical statistics*, 15(3):651–674 (2006b).

Ishwaran, H. and Kogalur, U. *Random Forests for Survival, Regression and Classification (RF-SRC)* (2015). R package version 1.6.0.
URL `http://cran.r-project.org/web/packages/randomForestSRC/`

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. "Random Survival Forests." *The Annals of Applied Statistics*, 841–860 (2008).

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. "High-dimensional Variable Selection for Survival Data." *Journal of the American Statistical Association*, 105(489):205–217 (2010).

Jensen, G., Torp-Pedersen, C., Hildebrandt, P., Kober, L., Nielsen, F., Melchior, T., Joen, T., and Andersen, P. "Does in-hospital Ventricular Fibrillation Affect Prognosis after Myocardial Infarction?" *European Heart Journal*, 18(6):919–924 (1997).

Keleş, S. and Segal, M. R. "Residual-based Tree-structured Survival Analysis." *Statistics in Medicine*, 21(2):313–326 (2002).

Kim, J. H. "Estimating Classification Error Rate: Repeated Cross-validation, Repeated Hold-out and Bootstrap." *Computational Statistics & Data Analysis*, 53(11):3735–3745 (2009).

Koul, H., Susarla, V., and Van Ryzin, J. "Regression analysis with randomly right-censored data." *The Annals of Statistics*, 9(6):1276–1288 (1981).

Last, G. and Brandt, A. *Marked Point Processes on the Real Line : the Dynamic Approach*. Probability and its applications. Berlin, New York: Springer-Verlag (1995).

LeBlanc, M. and Crowley, J. "Relative Risk Trees for Censored Survival Data." *Biometrics*, 411–425 (1992).

Leblanc, M. and Crowley, J. "Survival Trees by Goodness of Split." *Journal of the American Statistical Association*, 88(422):457–467 (1993).

Leurgans, S. "Linear Models, Random Censoring and Synthetic Data." *Biometrika*, 74(2):301–309 (1987).

Liaw, A. and Wiener, M. "Classification and Regression by randomForest." *R News*, 2(3):18–22 (2002).
URL `http://CRAN.R-project.org/doc/Rnews/`

Lostritto, K., Strawderman, R. L., and Molinaro, A. M. "A Partitioning Deletion/Substitution/Addition Algorithm for Creating Survival Risk Groups." *Biometrics*, 68(4):1146–1156 (2012).

Mentch, L. and Hooker, G. "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests." *Journal of Machine Learning Research*, 17(26):1–41 (2016).

Molinaro, A. M., Dudoit, S., and van der Laan, M. J. "Tree-based Multivariate Regression and Density Estimation with Right-censored Data." *Journal of Multivariate Analysis*, 90(1):154–177

(2004).

Nicod, P., Gilpin, E., Dittrich, H., Polikar, R., Henning, H., and Ross, J. "Long-term Outcome in Patients with Inferior Myocardial Infarction and Complete Atrioventricular Block." *Journal of the American College of Cardiology*, 12(3):589–594 (1988).

Plonquet, A., Haioun, C., Jais, J., Debard, A., Salles, G., Bene, M., Feugier, P., Rabian, C., Casasnovas, O., Labalette, M., et al. "Peripheral Blood Natural Killer Cell Count is Associated with Clinical Outcome in Patients with aaIPI 2–3 Diffuse Large B-cell Lymphoma." *Annals of Oncology*, 18(7):1209–1215 (2007).

Præstgaard, J. and Wellner, J. A. "Exchangeably Weighted Bootstraps of the General Empirical Process." *Annals of Probability*, 2053–2086 (1993).

Rose, S. and van der Laan, M. J. "Why TMLE?" In *Targeted Learning*, 101–118. Springer (2011).

Rotnitzky, A. and Vansteelandt, S. "Double-robust methods." In Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G. (eds.), *Handbook of Missing Data Methodology*, Handbooks of Modern Statistical Methods, 185–212. CRC Press (2014).

Rubin, D. and van der Laan, M. J. "A doubly Robust Censoring Unbiased Transformation." *The International Journal of Biostatistics*, 3(1) (2007).

Rubin, D. B. "The Bayesian Bootstrap." *The Annals of Statistics*, 9(1):130–134 (1981).

Scornet, E., Biau, G., and Vert, J.-P. "Consistency of Random Forests." *The Annals of Statistics*, 43(4):1716–1741 (2015).

Segal, M. R. "Regression Trees for Censored Data." *Biometrics*, 35–47 (1988).

Steingrimsson, J., Diao, L., Molinaro, A. M., and Strawderman, R. L. "Doubly Robust Survival Trees." *Statistics in medicine*, 35(17-18):3595–3612 (2016).

Suzukawa, A. "Unbiased Estimation of Functionals under Random Censorship." *Journal of the Japan Statistical Society*, 34(2):153–172 (2004).

Therneau, T. "User Written Splitting Functions for RPART." (2014).

Therneau, T., Atkinson, B., and Ripley, B. *rpart: Recursive Partitioning and Regression Trees* (2014). R package version 4.1-8.
URL `http://CRAN.R-project.org/package=rpart`

Tsiatis, A. *Semiparametric Theory and Missing Data*. Springer Science & Business Media (2007).

Van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. "A Gene-expression Signature as a Predictor of Survival in Breast Cancer." *New England Journal of Medicine*, 347(25):1999–2009 (2002).

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature*, 415(6871):530–536 (2002).

Wang, Z. and Wang, C. "Buckley-James Boosting for Survival Analysis with High-dimensional Biomarker Data." *Statistical Applications in Genetics and Molecular Biology*, 9(1) (2010).

Zhu, R. and Kosorok, M. R. "Recursively Imputed Survival Trees." *Journal of the American Statistical Association*, 107(497):331–340 (2012).

Supplementary Web Appendix

References to figures, tables, theorems and equations preceded by "S-" are internal to this supplement; all other references refer to the main paper.

S.1 Additional Simulation Results

In this section we present additional simulation results supplementing the results given in Section 5 in the main document.

S.1.1 Results for 75th and 25th quantile

Figures S-1 and S-2 show results estimating $P(T > t|W)$ with t chosen as the 25th and 75th quantile of the marginal failure time distribution in simulation settings $1-4$. Details of the simulation setup can be found in Section 5.1. The trends for the CURE$-L_2$ algorithm are similar to the ones seen in Figure 1. The CURE$-L_2$ algorithms *L2* and *L2 BJ* outperform *RSF* in all settings. Compared to *CI*, these methods are significantly better in Settings 3 and 4 and perform similarly in Settings 1 and 2. For *RIST*, performance is similar in all settings and at all quantiles. There is somewhat greater variation in the performance comparisons for the single time point methods `Brier` and `Brier BJ` methods, though generally speaking these methods remain competitive to the others (each of which uses information across time). The respective performance of the Buckley-James and doubly robust CUTs is similar in all settings, though there are notable improvements using the Buckley-James CUT (*Brier BJ* vs. *Brier*) at the 75th quantile in Setting 4. It is well known that estimators based on IPCW weights, such as doubly robust estimators, have the disadvantage of not being guaranteed to respect the natural range of the target parameter (Rose and van der Laan, 2011). Using the Brier loss function, the target parameter is a probability and is therefore constrained to fall in $[0, 1]$; when the terminal node estimators used in CURT trees that comprise the *Brier* predictions are truncated to fall in that interval the performance of the modified *Brier* algorithm is again comparable to *Brier BJ*.

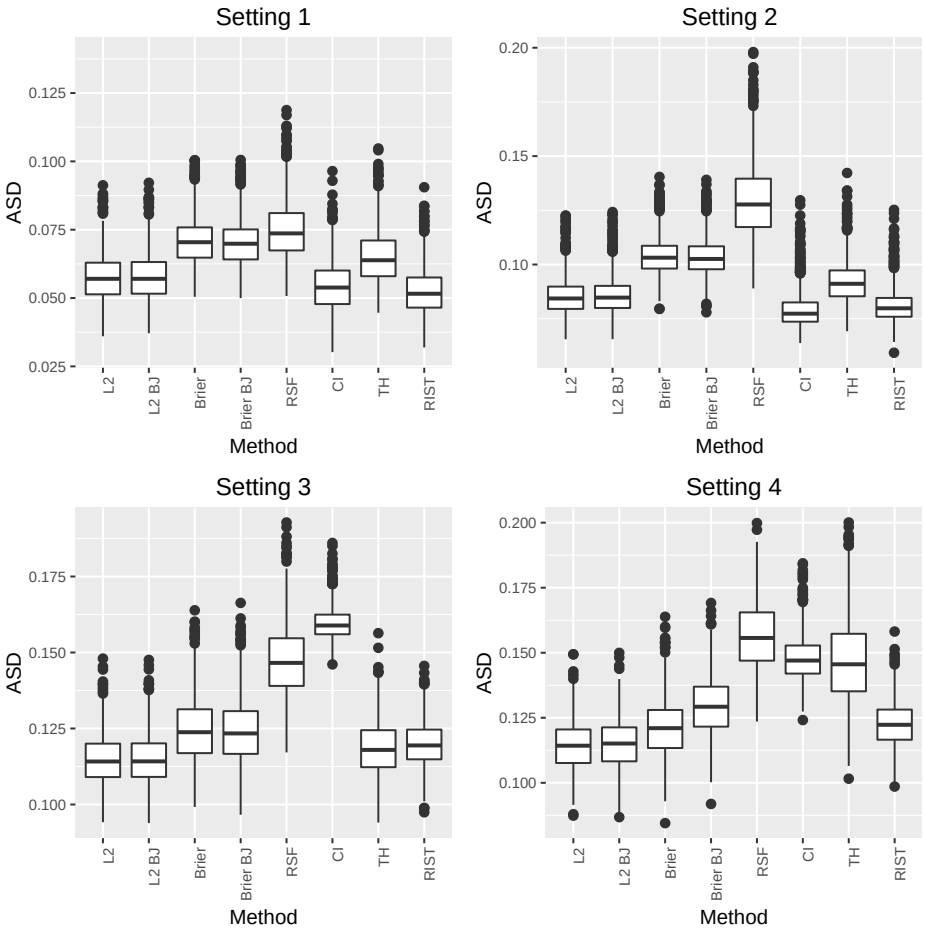


Figure S-1: Boxplots of MSE estimated at the $25th$ quantile of the marginal failure time distribution for the four simulation settings of Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` package. *RIST* is the recursively imputed survival trees.

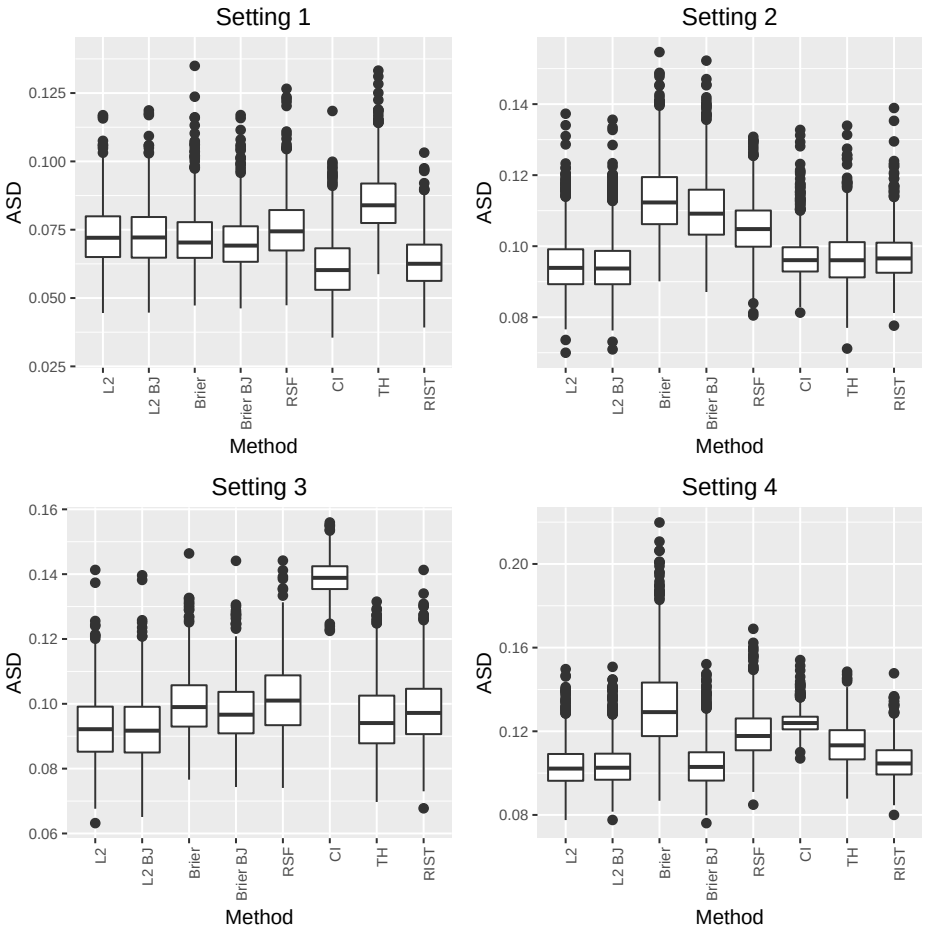


Figure S-2: Boxplots of MSE estimated at the $75th$ quantile of the marginal failure time distribution for the four simulation settings of Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. *RSF* and *CI* are the default methods for `rfsrc` and `cforest` package. *RIST* is the recursively imputed survival trees.

S.1.2 Comparison of nonparametric and Bayesian bootstrap for ensembles

In this section we compare the performance of the CURE$-L_2$ algorithm implemented using the Bayesian and the non-parametric bootstrap. Both bootstraps are implemented using the `R` function `randomForest` in the `randomForest` package (Liaw and Wiener, 2002) with the Bayesian bootstrap requiring extending the capabilities of the function to allow for arbitrary bootstrap weights. The simulation settings used are the same as used in the main document; see Section 5.1 for further details. The results for each *CURE*-L_2 algorithm are given in Figures S-3 - S-5.

From Figures S-3 - S-5 we see that the CURE$-L_2$ algorithm is not very sensitive to the choice of bootstrap weights. For the Brier loss function, the nonparametric bootstrap does as well or slightly better than the Bayesian bootstrap in all settings and at all quantiles. For the L_2 loss, the relative performance of the two bootstrap procedures depends on the simulation setting and quantile considered.

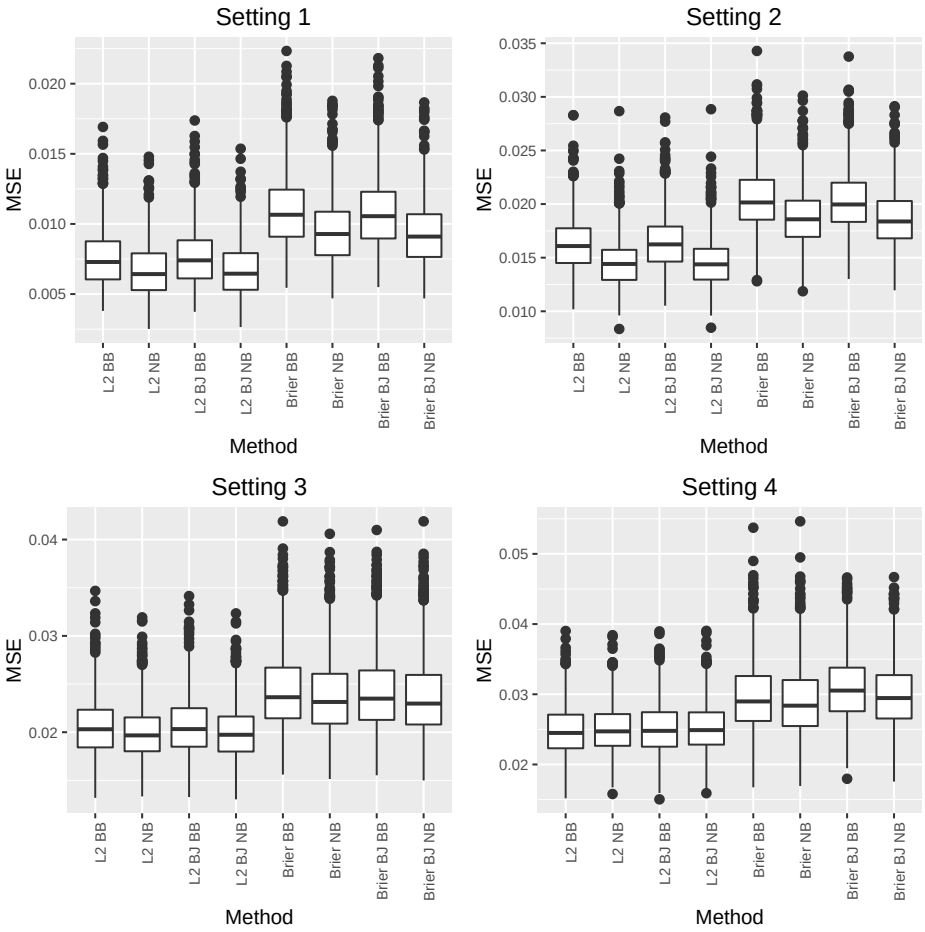


Figure S-3: Boxplots of MSE at the $25th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. NB and BB respectively indicate use of the non-parametric and Bayesian bootstrap weights.

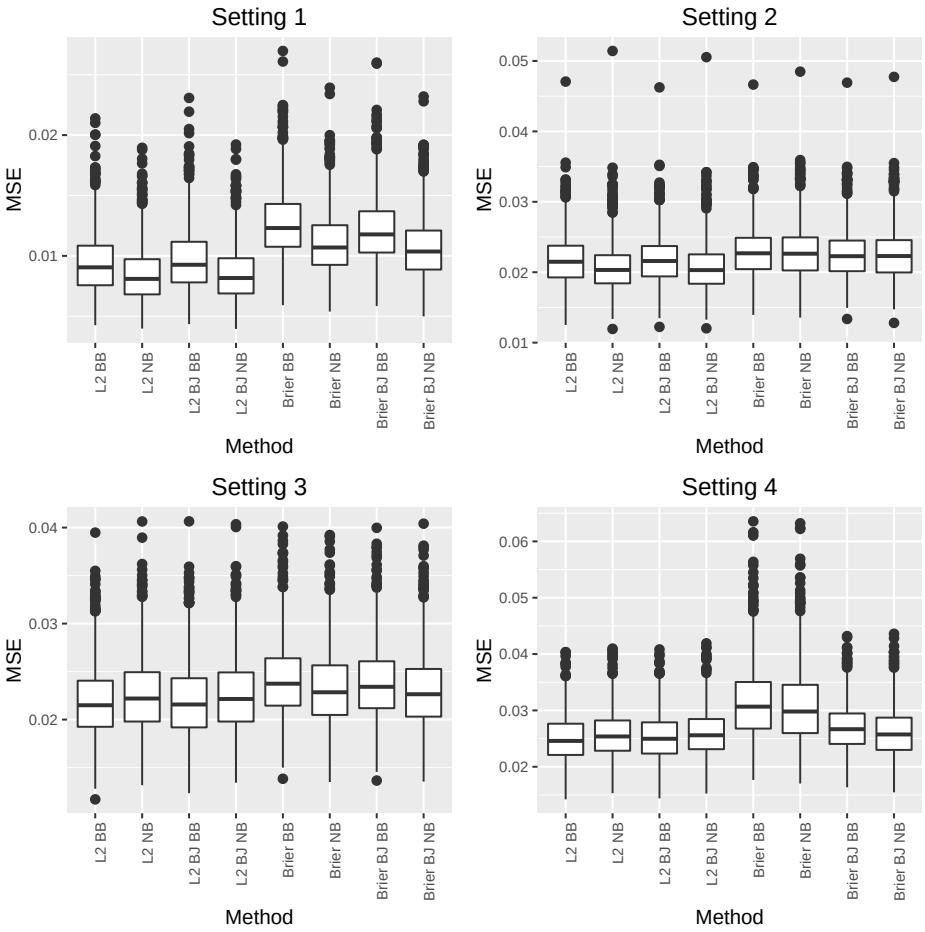


Figure S-4: Boxplots of MSE at the $50th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. NB and BB respectively indicate use of the non-parametric and Bayesian bootstrap weights.

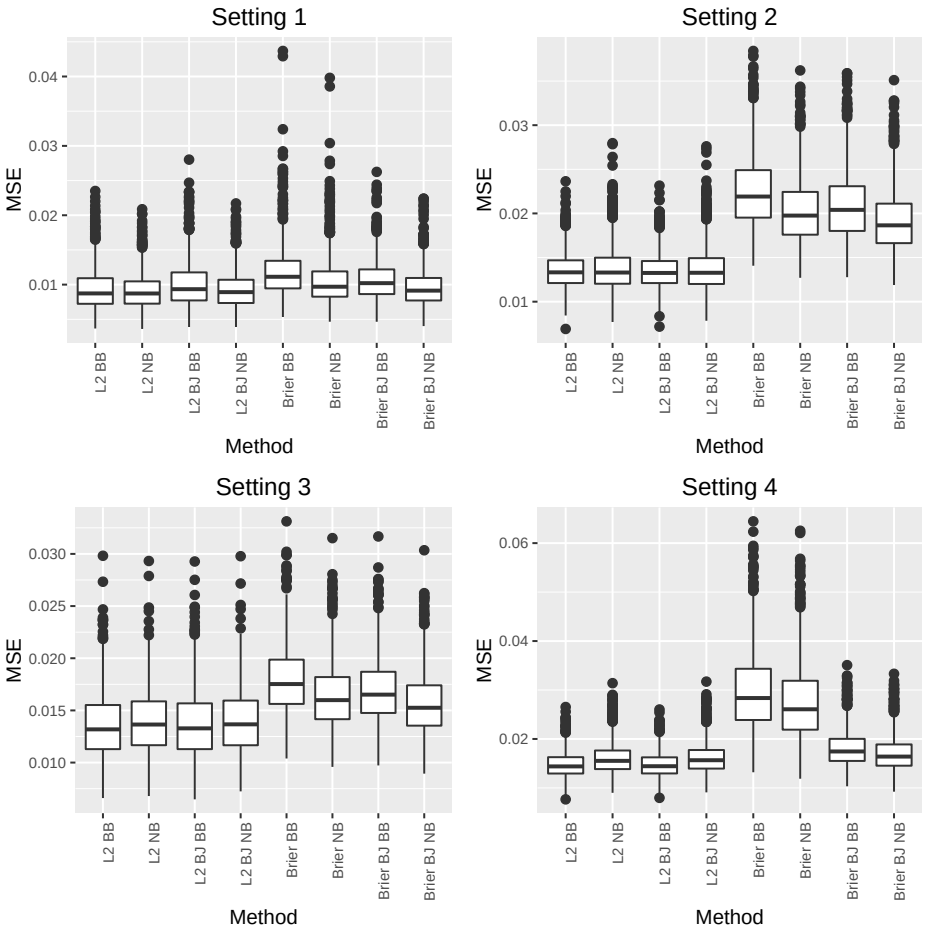


Figure S-5: Boxplots of MSE at the $75th$ quantile of the marginal failure time distribution for the four simulation settings described in Section 5.1. *L2, L2 BJ, Brier* and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT. NB and BB respectively indicate use of the non-parametric and Bayesian bootstrap weights.

S.1.3 Revisiting the simulation study in Steingrimsson et al. (2016)

In this section, we revisit the simulation studies for survival trees conducted in Steingrimsson et al. (2016) and compare the performance of the CURT$-L_2$ algorithm using the Buckley-James (see (6)) and doubly robust (see (5)) loss functions, with both being implemented using the imputation approach described in Section 4.1. The following two subsections revisit the simulation settings used in Steingrimsson et al. (2016) (Section S.1.3) and summarize the results (Section S.1.3).

A.1 Simulation Settings

Steingrimsson et al. (2016) considered two simulation settings. Both settings contain a training set of 250 independent subjects from the observed data distribution (subject to right censoring) and a test set of 2000 independent observations from the full data distribution (with no censoring). We simulate 1000 independent training and test set combinations. We briefly review the two settings considered below:

Simulation Setting 1: There are five covariates $W_1, \ldots, W_5$, each of which follows a discrete uniform distribution on the integers 1-100. The response $Z = \log(T)$ and survival times T are generated from an exponential distribution with a covariate-dependent mean parameter $\mu = aI(W_1 > 50 \mid W_2 > 75) + 0.5I(W_1 \leq 50 \ \& \ W_2 \leq 75)$. We consider "high" ($a = 5$), "medium" ($a = 2$) and "low" ($a = 1$) signal settings representing different degrees of separation in the survival curves. The censoring time C follows an exponential distribution with mean parameter μ_c, where μ_c is chosen to (approximately) achieve a 30% marginal censoring rate, in other words, $P(T \geq C; \mu_c) = 0.3$.

Simulation Setting 2: This simulation setting is similar to setting D in LeBlanc and Crowley (1992). It differs from Setting 1 in that the proportional hazard assumption does not hold. Assume that covariates $W_1, \ldots, W_5$ are independently uniformly distributed on the interval [0,1]. Survival times are generated from a distribution with survivor function $S(t|W) = [1 + t\exp(aI(W_1 \leq 0.5, W_2 > 0.5) + 0.367)]^{-1}$. The choices $a = 2, 1.5$ and 1 respectively correspond to "high", "medium" and "low" signal settings. The censoring times C follow a uniform distribution on $[0, b]$, where b is chosen to (approximately) achieve a 30% marginal censoring rate.

A.2 Simulation Results

The censoring distributions in both Settings 1 and 2 are independent of covariates; each is estimated using a Kaplan-Meier estimator. The conditional expectations required for computing the doubly robust and Buckley-James loss functions are respectively estimated using a parametric accelerated failure time (AFT) model with lognormal errors and also using random survival forests; see Section 3.2.2 of Steingrimsson et al. (2016) for details. The performance of the different survival trees for Settings 1 and 2 is respectively summarized in Figures S-6 and S-7 using the mean squared error of survival differences at the 25th, 50th and 75th quantile of the marginal failure time distribution (MSE25, MSE50 and MSE75). Each figure contains 9 plots and summarizes the results for MSE25, MSE50 and MSE75 under high, medium and low signal settings. The 6 boxplots in each plot respectively correspond to the method of LeBlanc and Crowley (1992) as implemented in `rpart` (EXP); the inverse probability censoring weighted L_2 loss (*IPCW*); the doubly robust L_2 loss calculated using the parametric AFT model (*DR-AFT*); the doubly robust L_2 loss calculated using random survival forest predictions (*DR-RF*); the Buckley-James L_2 loss calculated using the parametric AFT model (*BJ-AFT*); and, the Buckley-James L_2 loss calculated using random survival forest predictions (*BJ-RF*).

Figure S-6 shows that the performance of *EXP*, *IPCW* and doubly robust survival trees with conditional expectation estimated using either the AFT model or random survival forests are very

similar, a result entirely consistent with the results for Simulation 1 in Steingrimsson et al. (2016). The doubly robust trees perform better than the *IPCW* trees in the high and medium signal setting and show similar performance in the low signal setting; performs similarly or slightly better than EXP in high signal setting, however, as well or slightly worse in medium and low signal settings. The performance of Buckley-James trees is essentially the same as the doubly robust trees in nearly all signal settings, with the Buckley-James trees fit using the AFT model having slightly smaller MSE at the 75th quantile.

For Simulation Setting 2, Figure S-7 shows that the doubly robust and Buckley-James trees perform noticeably better than both the *IPCW* trees and *EXP* method in the high and medium setting; performance is comparable for all methods in the low signal setting. Each of *DR-AFT*, *DR-RF*, *BJ-AFT* and *BJ-RF* have comparable performance in high and low signal settings; *BJ-AFT* performs best, with *DR-RF* being second best, in the medium signal setting.

For completeness we also looked at the performance of all survival trees in terms of prediction error as we did in Steingrimsson et al. (2016) (results not shown here). The results for *EXP*, *IPCW*, *DR-AFT* and *DR-RF* are consistent with those results, the pattern of prediction error agreeing with that observed in MSE25, MSE50 and MSE75.

Figure S-6: Boxplots of mean squared error of survival differences at the 25th, 50th and 75th quantile of the marginal failure time distribution (MSE25, MSE50 and MSE75) using the default method in `rpart` (*EXP*), inverse probability censoring weighted loss (*IPCW*), doubly robust L_2 loss with the AFT model (*DR-AFT*), doubly robust L_2 loss with RSF (*DR-RF*), Buckley-James L_2 loss with the AFT model (*BJ-AFT*) and Buckley-James L_2 loss with RSF (*BJ-RF*), respectively for the high, medium and low signal settings in Setting 1.

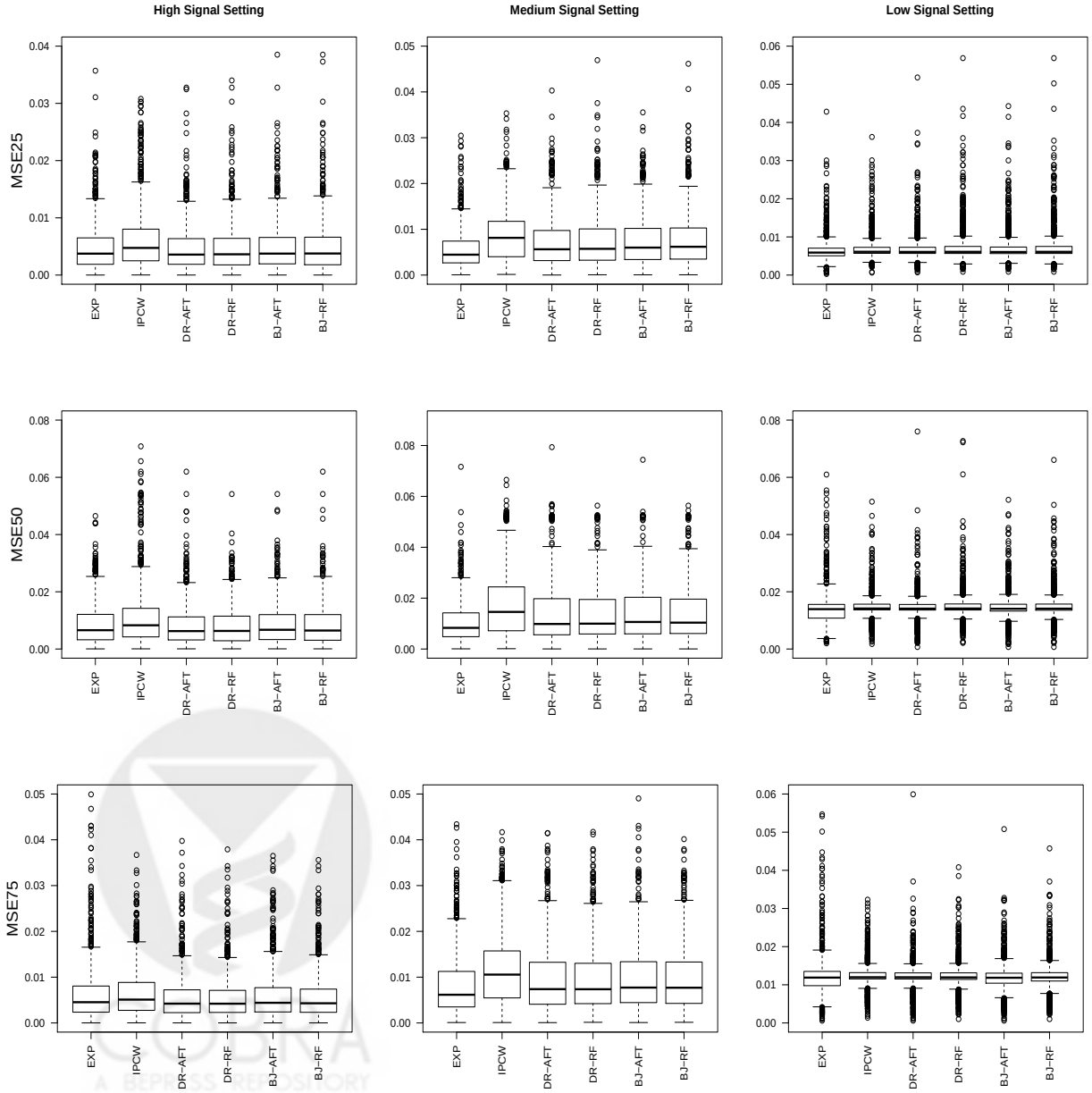
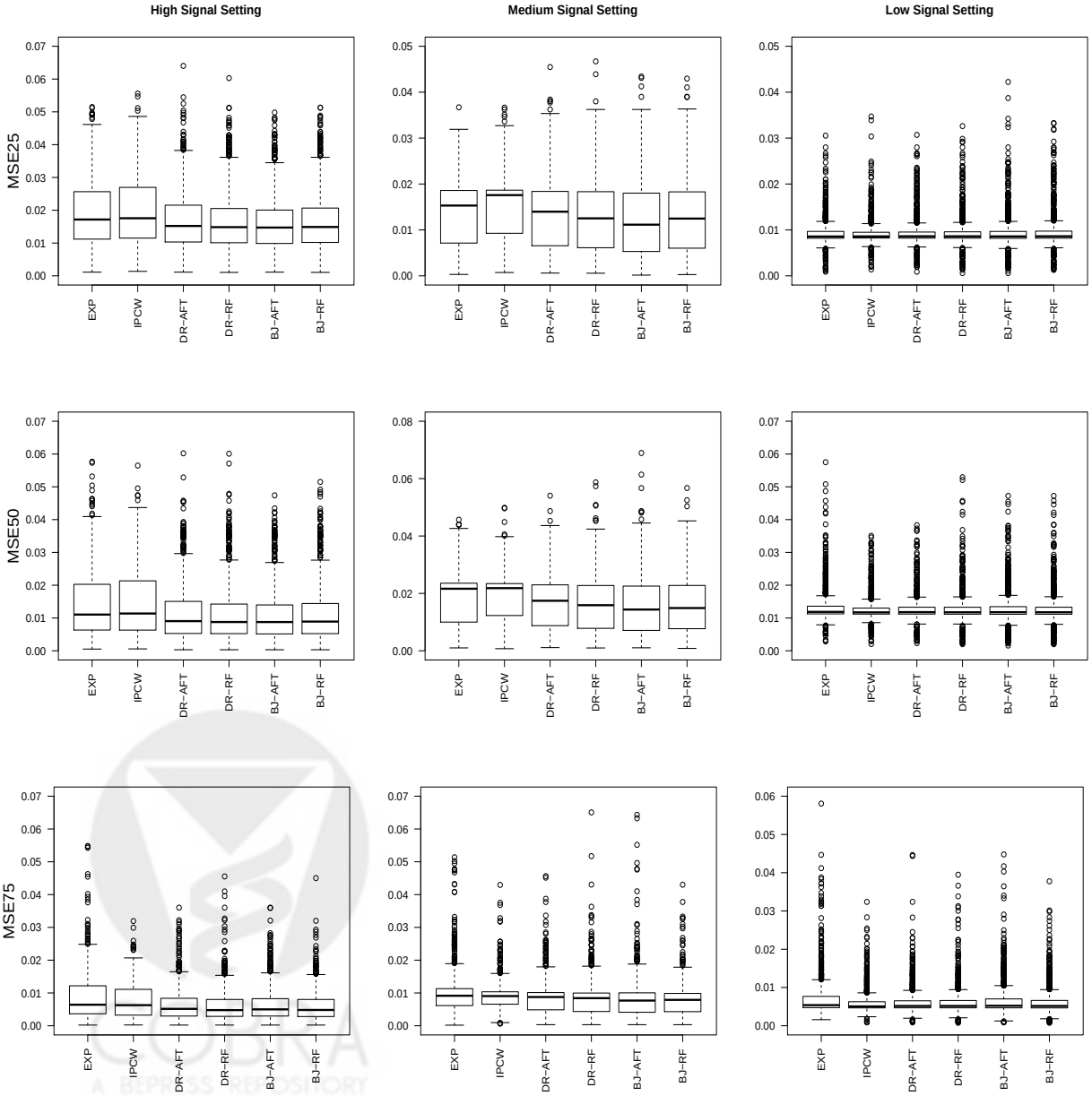


Figure S-7: Boxplots of mean squared error of survival differences at the 25th, 50th and 75th quantile of the marginal failure time distribution (MSE25, MSE50 and MSE75) using the default method in `rpart` (*EXP*), inverse probability censoring weighted loss (*IPCW*), doubly robust L_2 loss with the AFT model (*DR-AFT*), doubly robust L_2 loss with RSF (*DR-RF*), Buckley-James L_2 loss with the AFT model (*BJ-AFT*) and Buckley-James L_2 loss with RSF (*BJ-RF*), respectively for the high, medium and low signal settings in Setting 2.



S.2 Further Details on OOB-Based Variable Importance Measures

Consider an ensemble generated by the nonparametric bootstrap. Given a tree m from this ensemble, let $\hat{\psi}_m(W)$ be the corresponding prediction for a subject with covariate information W. Let B_m be the set of OOB data associated with the bootstrap sample used to create tree m. The L_2 OOB data prediction error for tree m is defined as

$$\frac{1}{|B_m|}\sum_{i=1}^{n} I(i \in B_m) L_2(Z_i, \hat{\psi}_m(W_i)), \tag{S-1}$$

where $|B_m|$ denotes the size of the OOB sample. For each $i \in B_m$ let $W_i^{(j)}$ be the covariate vector for subject i with the j-th component of the covariate permuted. Define the OOB L_2 loss prediction error using the resulting permuted OOB dataset as

$$\frac{1}{|B_m|}\sum_{i=1}^{n} I(i \in B_m) L_2(Z_i, \hat{\psi}_m(W_i^{(j)})). \tag{S-2}$$

The OOB prediction error VIMP proposed by Breiman (2001) is calculated as the difference between (S-2) and (S-1), averaged over all the trees in the ensemble. That is, for covariate j the OOB prediction error VIMP is defined as

$$\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{|B_m|}\sum_{i=1}^{n} I(i \in B_m)(L_2(Z_i, \hat{\psi}_m(W_i^{(j)})) - L_2(Z_i, \hat{\psi}_m(W_i)))\right). \tag{S-3}$$

This calculation assumes that $(Z_i, W_i), i \in B_m$ are fully observed. The corresponding VIMP using a CUT for the L_2 loss function can simply be defined as that which is obtained by replacing the (unobserved) L_2 loss in (S-3) with its corresponding CUT as given in (9).

As the OOB prediction error VIMP is defined as the difference between two loss functions the proof of the following theorem follows from exactly the same arguments as used to prove Theorem 4.2. For the notation used in the theorem we refer to Section 4 in main paper.

Theorem S.2.1. *For each $i = 1, \ldots, n$, define the loss function $L_2(O_i, \psi; G, S, Q) = \psi(W_i)^2 + H(O_i; G, S)\psi(W_i) + Q(O_i; G, S)$ and assume $\max\{|H(O_i; G, S)|, |Q(O_i; G, S)|\} < \infty$. The OOB prediction error VIMPs using the loss function $L_2(O, \psi; G, S, Q)$ do not depend on $Q(O; G, S)$.*

An important implication of Theorem S.2.1 is that the OOB data prediction error VIMPs using the L_2 loss in connection with doubly robust and Buckley-James CUTs can be implemented by running standard software calculating the OOB prediction error VIMPs for fully observed responses on the corresponding "imputed" dataset $\{(\hat{Z}(O_i; G, S), W_i); i = 1, \ldots, n\}$.

S.3 Additional Results from Data Analysis Section

In this section we present additional results for the data analyzed in Section 6 of the main paper.

Table S-1 shows the OOB prediction error for the TRACE data. In agreement with the results obtained from the minimal depth variable importance measures, Table S-1 shows that ventricular fibrillation is consistently the least important predictor across all methods and age and CHF are consistently the two most influential variables.

Table S-2 shows the minimal depth VIMPs for the Worcester heart attack study, again ordered by decreasing importance according to the *RSF* VIMPs. All algorithms show age as being the

	L2	L2 BJ	Brier	Brier BJ	RSF
Age	0.54	0.26	0.04	0.04	0.08
CHF	0.25	0.23	0.02	0.02	0.04
Diabetes	0.17	0.19	0.01	0.01	0.00
Gender	0.01	0.03	0.01	0.01	0.01
VF	-0.02	0.01	0.00	0.00	0.00

Table S-1: Out-of-bag prediction error variable importance measures for the TRACE data; higher values indicate more influential variables. *Brier* and *L2* refer to the loss function used. *BJ* refers to the Buckley-James transformation. *RSF* is the default method in the `randomForestSRC` package. CHF stands for clinical heart pump failure and VF stands for ventricular fibrillation.

most influential predictor, a result that agrees well with several studies showing the importance of age as a predictor for overall survival; see Goldberg et al. (1989) and references there within. BMI has the second lowest VIMP for three out of the five algorithms and the third lowest for the other two. BMI has been shown to be an important predictor for myocardial infarction (Fitzgibbons et al., 2009). Complete heart block is consistently the least important predictor; the results in Nicod et al. (1988, Figure 2) show no significant impact of complete heart block on the long-term prognosis of patients. Table S-3 shows the corresponding OOB prediction error VIMPs for these data, with similar conclusions between and across methods. Importantly, in contrast to minimal depth VIMPs, higher positive values indicate more influential variables and values close to zero (or negative) indicate that the variable is not important.

	L2	L2 BJ	Brier	Brier BJ	RSF
Age	1.13	1.16	1.08	1.05	1.74
BMI	1.65	1.73	1.72	1.71	1.95
Heart Rate	2.50	2.61	1.80	1.78	2.37
Diastolic Blood Pressure	2.53	2.37	2.49	2.49	2.47
Systolic Blood Pressure	2.40	2.38	2.78	2.84	2.57
Congestive Heart Complications	2.75	2.88	1.50	1.56	2.58
Cardiogenic Shock	3.57	3.25	7.57	7.02	3.58
Cohort Year	3.31	3.46	3.71	3.67	3.74
MI Type	6.35	6.28	6.14	5.96	4.43
MI Order	6.65	6.05	5.95	5.98	4.54
Atrial Fibrillation	6.62	6.89	5.56	6.34	4.62
Gender	6.63	6.46	5.51	5.66	4.86
History of Cardiovascular Disease	7.83	7.74	7.16	7.86	5.32
Complete Heart Block	12.42	12.05	12.22	10.67	7.49

Table S-2: Minimal depth variable importance measures for the Worcester Study; lower values indicate more influential variables. *Brier* and *L2* refer to the loss function used. *BJ* refers to the Buckley-James transformation. *RSF* is the default method in the `randomForestSRC` package.

The Netherlands and R-Chop datasets are comparatively high dimensional and tabular displays of variable importance are not especially informative. In the case of the Netherlands study (Van De Vijver et al., 2002), the genes included for analysis were already selected from a much larger pool, and one of the main conclusions in this study is that models that include these 70 gene expression profiles provide more information than models that do not rely on that information. In Figure

	L2	L2 BJ	Brier	Brier BJ	RSF
Age	0.952	0.948	0.071	0.072	0.090
BMI	0.140	0.199	0.013	0.011	0.012
Heart Rate	0.051	0.065	0.015	0.015	0.011
Diastolic Blood Pressure	0.156	0.135	0.003	0.003	0.005
Systolic Blood Pressure	0.095	0.098	0.001	0.002	0.005
Congestive Heart Complications	0.154	0.124	0.039	0.038	0.026
Cardiogenic Shock	0.138	0.127	0.002	0.003	0.008
Cohort Year	0.302	0.143	0.007	0.007	0.006
MI Type	0.011	0.018	0.000	0.000	-0.001
MI Order	0.005	0.002	0.001	0.002	0.002
Atrial Fibrillation	0.021	0.009	-0.001	-0.000	0.002
Gender	-0.008	-0.007	0.001	0.001	-0.001
History of Cardiovascular Disease	-0.018	-0.005	-0.000	-0.000	-0.001
Complete Heart Block	0.000	-0.002	-0.000	-0.000	-0.000

Table S-3: Out-of-bag prediction error variable importance measures for the Worcester Study; higher values indicate more influential variables. *Brier* and *L2* refer to the loss function used. *BJ* refers to the Buckley-James transformation. *RSF* is the default method in the `randomForestSRC` package.

S-8, we compare the prediction accuracy of the CURE$-L_2$ algorithms built using both clinical information and gene expression measurements to models built using only clinical information. Consistent with the findings in Van De Vijver et al. (2002), we see that adding gene expression information substantially improves the prediction power of the algorithms. In the case of the R-Chop data, which involves 3833 probe sets, the minimal depth VIMP measures for both the CURE$-L_2$ algorithms and the *RSF* algorithm identify the same probe set as being the most influential. This probe set corresponds to a killer cell lectin-like receptor NKG2A that is a known natural killer (NK) cell receptor; see Brooks et al. (1997). Plonquet et al. (2007) found NK cell counts to be an important predictor for clinical outcomes in diffuse large B-cell lymphoma. We also calculated the OOB prediction error VIMP measures and then evaluated the degree of overlap between the 25 most influential probe sets for both VIMPs. The number of probe sets that were in the top 25 most influential variables for both VIMP measures was 13, 10, 7, and 13 for the *DR* L_2, *BJ* L_2, *DR Brier* and *BJ Brier* algorithms, respectively.

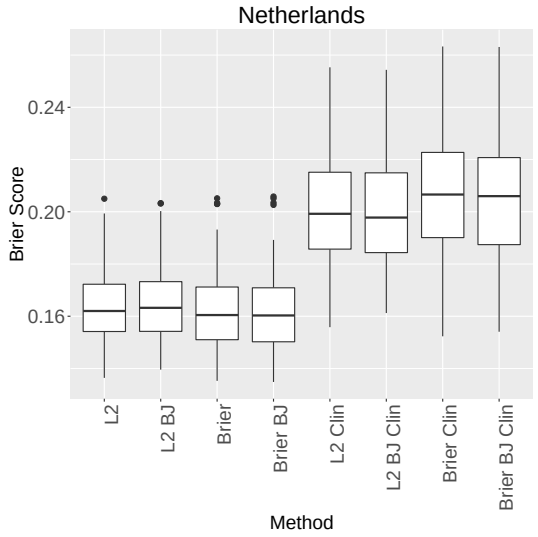


Figure S-8: Prediction error for four CURE$-L_2$ algorithms (with and without genetic information) on the Netherlands breast cancer study data. *L2*, *L2 BJ*, *Brier* and *Brier BJ* are the CURE$-L_2$ algorithms, with *BJ* referring to the use of the Buckley-James CUT, and *L2* and *Brier* referring to the choice of loss function. *Clin* refers to the model only being built using clinical factors.

S.4 Proof of Theorem 2.1

S.4.1 Regularity Conditions

The conditions of Section 2.1 specify that $S_0(t|w)$ and $G_0(t|w)$ are each continuous functions in $t \in \mathbf{R}^+$ for each $w \times \mathcal{S}$ and, in addition, that $\vartheta_{S_0} = \inf\{t : S_0(t|w) = 0\}$ and $\vartheta_{G_0} = \inf\{t : G_0(t|w) = 0\}$ are independent of $w \in \mathcal{S}$. The conditions of Section 2.2 imply that $\phi(h(u), w), (u, w) \in \mathbf{R}^+ \times \mathcal{S}$ is a known scalar function that is continuous in u except possibly at a finite number of points and bounded if $\max\{|r|, \|w\|\} < \infty$. Let $\mu(w) = E[\phi(Z, W)|W = w] = \int_0^\infty \phi(h(u), w) dF_0(u|w) < \infty$ for each $w \in \mathcal{S}$, where $F_0(u|w) = 1 - S_0(u|w)$. We assume that $S(t|w)$ and $G(t|w)$ are each right-continuous, non-increasing functions for $t \geq 0$ that satisfy $S(0|w) = G(0|w) = 1$, $S(t|w) \geq 0$ and $G(t|w) \geq 0$ for each $w \in \mathcal{S}$. Below, let $F(u|w) = 1 - S(u|w)$, $\bar{G}(u|w) = 1 - G(u|w)$, and $\bar{G}_0(u|w) = 1 - G_0(u|w)$.

The conditions imposed are weak enough to accomodate (2) as a special case of (4). For each $w \in \mathcal{S}$, we further assume

(C1) $I_1 = \int_0^\infty \phi(h(u), w) \frac{G_0(u|w)}{G(u-|w)} dF_0(u|w) < \infty$;

(C2) $M_1(r) = \int_0^r \frac{S_0(u|w)}{S(u|w)} \frac{d\bar{G}_0(u|w)}{G(u-|w)} < \infty$ and $M_2(r) = \int_0^r \frac{G_0(u|w)S_0(u|w)}{G(u|w)S(u|w)} \frac{d\bar{G}(u|w)}{G(u-|w)} < \infty$ for each $r > 0$;

(C3) $I_2 = \int_0^\infty \phi(h(u), w) \left[M_1(u-) - M_2(u-)\right] dF(u|w) < \infty$;

(C4) $\int_0^\infty \frac{[\phi(h(u),w)]^2}{G_0(u|w)} dF_0(u|w) < \infty$;

(C5) $\int_0^\infty \frac{m_\phi^2(u,w;S)}{G_0^2(u|w)} S_0(u|w) d\bar{G}_0(u|w) < \infty$;

(C6) $M_3(r) = \int_0^r \frac{|m_\phi(u,w;S)|}{G_0^2(u|w)} d\bar{G}_0(u|w) < \infty$ for each $r > 0$;

S.4.2 Proof that $Y_d^*(O;G,S_0)$, $Y_d^*(O;G_0,S)$ and $Y_d^*(O;G_0,S_0)$ are each CUTs for $Y=\phi(Z,W)$

Assume Conditions (C1)-(C3) hold. Then, calculations similar to those in Rubin and van der Laan (2007) show that $E[Y_d^*(O;G,S)|W=w]=I_1+I_2$. Consider now the following cases:

- Suppose only that $G(u|w)=G_0(u|w)$ for every (u,w). Then, because $G_0(u|w)$ is continuous, $G_0(u|w)/G_0(u-|w)=1$ everywhere and it follows that

$$I_1=\int_0^\infty \phi(h(u),w)dF_0(u|w)=\mu(w).$$

  In addition, for every $r\geq 0$, we obtain

$$M_1(r)-M_2(r)=\int_0^r \frac{S_0(u|w)}{S(u|w)}\frac{d\bar{G}_0(u|w)}{G_0(u|w)}-\int_0^r \frac{G_0(u|w)}{G_0(u|w)}\frac{S_0(u|w)}{S(u|w)}\frac{d\bar{G}_0(u|w)}{G_0(u|w)}=0$$

  and hence $I_2=0$. Consequently, $E[Y_d^*(O;G_0,S)|W=w]=I_1+I_2=\mu(w)$.

- Suppose only that $S(u|w)=S_0(u|w)$ for every (u,w). Then,

$$I_1+I_2=\int_0^\infty \phi(h(u),w)\frac{G_0(u|w)}{G(u-|w)}dF_0(u|w)+\int_0^\infty \phi(h(u),w)\left[M_1(u-)-M_2(u-)\right]dF_0(u|w)$$

  and we see that $E[Y_d^*(O;G,S_0)|W=w]=I_1+I_2=\mu(w)$ provided that

$$\frac{G_0(u|w)}{G(u-|w)}+[M_1(u-)-M_2(u-)]=1.$$

  Under the assumption $S(u|w)=S_0(u|w)$, the definitions of $M_i(\cdot), i=1,2$ and the fact that $G_0(u|w)$ is continuous implies that we need only show

$$\frac{G_0(u|w)}{G(u|w)}+\int_0^u \frac{d\bar{G}_0(r|w)}{G(r-|w)}-\int_0^u \frac{G_0(r|w)}{G(r|w)}\frac{d\bar{G}(r|w)}{G(r-|w)}=1 \tag{S-4}$$

  for every $u\geq 0$. Using integration by parts (e.g., Last and Brandt, 1995, Thm. A.4.6),

$$\frac{G_0(u|w)}{G(u|w)}=1+\int_0^u G_0(r|w)\left(\frac{-dG(r|w)}{G(r-|w)G(r|w)}\right)+\int_0^u \frac{dG_0(r|w)}{G(r-|w)};$$

  rearranging this expression, we see

$$\frac{G_0(u|w)}{G(u|w)}+\int_0^u \frac{d\bar{G}_0(r|w)}{G(r-|w)}-\int_0^u G_0(r|w)\left(\frac{d\bar{G}(r|w)}{G(r-|w)G(r|w)}\right)=1$$

  which is exactly (S-4). This proves $E[Y_d^*(O;G,S_0)|W=w]=\mu(w)$.

The result that $E[Y_d^*(O;G_0,S_0)|W=w]=\mu(w)$ clearly follows from either of the above arguments, completing this part of the proof.

S.4.3 Proof that $Var(Y_d^*(O;G_0,S)|W) \geq Var(Y_d^*(O;G_0,S_0)|W)$.

Let $G(u|w) = G_0(u|w)$ be continuous and consider the class of transformations

$$Y_s^*(O;G_0,\gamma) = \frac{\Delta\phi(\tilde{Z},W)}{G_0(\tilde{T}|W)} + (1-\Delta)\gamma(\widetilde{T},W) - \int_0^{\widetilde{T}} \gamma(u,W)d\Lambda_{G_0}(u|W), \tag{S-5}$$

where $\gamma(u,W)$ is some specified function. The class of transformations defined by (S-5) is essentially seen to be the same as that considered in Suzukawa (2004, Prop. 3; Eqn. 3.6), but generalized here to allow for covariates and not restricted to depend on $G_0(\cdot|\cdot)$ alone. Importantly, it is also easy to see that selecting $\gamma^*(u,W) = m(u,W;S)/G_0(u|W)$ in (S-5) gives $Y_s^*(O;G_0,\gamma^*) = Y_d^*(O;G_0,S)$. For continuous $G_0(u|w)$, the regularity conditions (C4)-(C6) generalize those in Suzukawa (2004) needed to prove Propositions 3, 5 and 6 in Suzukawa (2004). In particular, we have $E[Y_d^*(O;G_0,S)|W=w] = \mu(w)$ and, mimicking the arguments used to prove Propositions 5 and 6, that $Var[Y_d^*(O;G_0,S)|W=w] = H_1(w;G_0,S_0) + H_2(w;G_0,S_0,S)$, where

$$H_1(w;G_0,S_0) = \int_0^\infty \frac{[\phi(h(x),w)]^2}{G_0(x|w)}dF_0(x|w) - \int_0^\infty \frac{S_0(x|w)[m(x,w;S_0)]^2}{G_0(x|w)^2}d\bar{G}_0(x|w) - \mu^2(w)$$

and

$$H_2(w;G_0,S_0,S) = \int_0^\infty \frac{S_0(x|w)(m(x,w;S) - m(x,w;S_0))^2}{G_0(x|w)^2}d\bar{G}_0(x|w).$$

This proves $Var[Y_d^*(O;G_0,S)|W=w] \geq Var[Y_d^*(O;G_0,S_0)|W=w] = H_1(w,G_0,S_0)$, with strict inequality when $S(t|w)$ and $S_0(t|w)$ differ (hence $m(t,w;S)$ and $m(t,w;S_0)$ differ) for t in some interval with positive length.

S.5 Proof of Theorem 4.1

We require the following lemma.

Lemma S.5.1. *Let $x \geq 0$ be finite. Let d be any indicator variable taking on the values 0 and 1. Let $B(s)$ be a right-continuous, non-decreasing function for $s \geq 0$ with $B(0) = 0$. Define $\bar{B}(s) = 1 - B(s)$ and $H(s) = \int_0^s \bar{B}^{-1}(u-)dB(u)$ for any s such that $H(s)$ exists.*
Suppose $\bar{B}(x) > 0$. Then, $H(s)$ exists for $s \in [0,x]$ and

$$\frac{d}{\bar{B}(x)} + \frac{1-d}{\bar{B}(x)} - \int_0^x \frac{dH(u)}{\bar{B}(u)} = 1.$$

There is no specific relationship assumed between x, d and $B(\cdot)$, hence $H(\cdot)$. Using notation from both the theorem statement and Lemma S.5.1, we can make the following identifications: $\bar{B}(t) = G(t|w)$, $H(t) = \Lambda_G(t|w)$, $x = \tilde{t}$, and $d = D$. The conditions of Theorem ensure that the conditions of Lemma S.5.1 are satisfied; applying Lemma S.5.1 immediately gives the desired result:

$$\frac{D}{G(\tilde{t}|w)} + \frac{(1-D)}{G(\tilde{t}|w)} - \int_0^{\tilde{t}} \frac{d\Lambda_G(u|t)}{G(u|t)} = 1.$$

Proof of Lemma S.5.1. Because $\bar{B}(x) > 0$ and is non-increasing with $\bar{B}(0) = 1$, right-continuity

implies $\inf_{s\leq x}\bar{B}(s) > 0$. Hence, we may write (e.g., Last and Brandt, 1995, Cor A.4.8, p. 426)

$$d\left(\frac{1}{\bar{B}(u)}\right) = -\frac{d\bar{B}(u)}{\bar{B}(u-)\bar{B}(u)} = \frac{dB(u)}{\bar{B}(u-)\bar{B}(u)} \tag{S-6}$$

Let $K(\cdot)$ be any right-continuous function of bounded variation on $[0, x]$. Then, we may write (Last and Brandt, 1995, Thm. A.4.6)

$$\frac{K(x)}{\bar{B}(x)} - \frac{K(0)}{\bar{B}(0)} = \int_0^x K(u-)\, d\left(\frac{1}{\bar{B}(u)}\right) + \int_0^x \left(\frac{1}{\bar{B}(u)}\right) dK(u).$$

Using (S-6) and assuming $K(s) = 1$ for $s \geq 0$, we obtain the identity

$$\frac{1}{\bar{B}(x)} - 1 = \int_0^x d\left(\frac{1}{\bar{B}(u)}\right) + \int_0^x \left(\frac{1}{\bar{B}(u)}\right) d(1) = \int_0^x \frac{dB(u)}{\bar{B}(u-)\bar{B}(u)} = \int_0^x \frac{dH(u)}{\bar{B}(u)}. \tag{S-7}$$

Observe that we may also write

$$\frac{d}{\bar{B}(x)} = \frac{1}{\bar{B}(x)} - \frac{1-d}{\bar{B}(x)};$$

using (S-7), it then follows that

$$\frac{d}{\bar{B}(x)} + \frac{1-d}{\bar{B}(x)} - 1 = \int_0^x \frac{dH(u)}{\bar{B}(u)},$$

from which the required identity follows immediately. □

S.6 Proof of Theorem 4.2

Considering $G(\cdot|\cdot)$ and $S(\cdot|\cdot)$ as fixed functions, we will for simplicity rewrite $L_2(O, \psi(W); G, S, Q)$ as $L_2(O, \psi(W); Q) = \psi(W)^2 + H(O)\psi(W) + Q(O)$. Under the stated conditions, we can also assume without loss of generality that $L_2(O, \psi(W); Q) \geq 0$. The proof of this theorem will follow if one can show that all key decisions made by CART are invariant to the form of $Q(O)$. The availability of a sample $O_1, \ldots, O_n$ such that $H(O_i)$ and $Q(O_i)$ satisfy the conditions of the theorem for each $i = 1, \ldots, n$ is assumed. Throughout this proof, it is assumed at each stage of the algorithm that one is working with some finite partition $\{\tau_j, j = 1, \ldots, J\}$ of $\mathcal{S}$ and that $\psi(W) = \sum_{j=1}^{J} I\{W \in \tau_j\}\psi_j$ is the corresponding piecewise constant predictor. In this case, for any subset τ_j,

$$n^{-1}\sum_{i=1}^{n} I\{W_i \in \tau_j\} L_2(O_i, \psi(W_i); Q) = n^{-1}\sum_{i=1}^{n} I\{W_i \in \tau_j\}\left[\psi_j^2 + H(O_i)\psi_j + Q(O_i)\right]$$

is uniquely minimized at $\hat{\psi}_j = \sum_{i=1}^{n} -H(O_i)/(2n_j)$ for $n_j = \sum_{i=1}^{n} I\{W_i \in \tau_j\}$.

Now we show that all three steps of the CART algorithm used in connection with $L_2(O, \psi(W); Q)$ involve decisions that are invariant to the specification of $Q(O)$.

S.6.1 Growing the tree

The first stage of the tree building process is to grow a very large tree. Three elements are required to accomplish this step: (i) developing the candidate set of binary splits; (ii) specifying the node splitting rule; and, (iii) specifying the rule to stop splitting nodes. Only step (ii) depends on the

specification of the loss function; hence, we focus on this step below. The following lemma is critical.

Lemma S.6.1. *Suppose $L_2(O, \psi(W); Q) = \psi(W)^2 + H(O)\psi(W) + Q(O)$ is used to evaluate the loss. Let $R(\tau)$ denote the loss within a given subset $\tau \subset \mathcal{S}$; that is, $R(\tau) = \sum_{i=1}^{n} I\{W_i \in \tau\} L_2(O_i, \widehat{\psi}_\tau; Q)$, where $\widehat{\psi}_\tau$ minimizes the loss function using the data falling into τ. If τ is then split into $L \geq 2$ mutually exclusive subsets $\tau_1, \ldots, \tau_L$ and $\tau_1 \cup \tau_2 \cup \ldots \cup \tau_L = \tau$, the corresponding change in total loss is given by*

$$R(\tau) - \sum_{\ell=1}^{L} R(\tau_\ell) = \sum_{\ell=1}^{L} \sum_{i=1}^{n} I\{W_i \in \tau_\ell\} \left[(\widehat{\psi}_\tau^2 - \widehat{\psi}_{\tau_\ell}^2) + H(O_i)(\widehat{\psi}_\tau - \widehat{\psi}_{\tau_\ell}) \right],$$

where $\widehat{\psi}_{\tau_\ell}$ is the value which minimizes the loss function using the data from the ℓth subset.

Proof. We have

$$R(\tau) = \sum_{i=1}^{n} I\{W_i \in \tau\} \left[\widehat{\psi}_\tau^2 + H(O_i)\widehat{\psi}_\tau + Q(O_i) \right]$$

and

$$\sum_{\ell=1}^{L} R(\tau_\ell) = \sum_{\ell=1}^{L} \sum_{i=1}^{n} I\{W_i \in \tau_\ell\} \left[\widehat{\psi}_{\tau_\ell}^2 + H(O_i)\widehat{\psi}_{\tau_\ell} + Q(O_i) \right].$$

Subtracting the second from the first, algebra shows that the change in total loss reduces to

$$R(\tau) - \sum_{\ell=1}^{L} R(\tau_\ell) \quad = \quad \sum_{\ell=1}^{L} \sum_{i=1}^{n} I\{W_i \in \tau_\ell\} \left[(\widehat{\psi}_\tau^2 - \widehat{\psi}_{\tau_\ell}^2) + H(O_i)(\widehat{\psi}_\tau - \widehat{\psi}_{\tau_\ell}^2) \right].$$

□

In the process of growing a tree, CART considers at each step all possible candidate splits of a given parent node τ into left and right child nodes, say τ_L and τ_R, and then chooses the (covariate, split) combination that maximizes the decrease $R(\tau) - R(\tau_L) - R(\tau_R)$. This process continues until the stop-splitting rule used in (iii) takes effect, generating a maximally-sized tree $\mathcal{T}_{max}$. Lemma S.6.1 shows that the reduction in loss is independent of $Q(O_i), i = 1 \ldots n$ regardless of the stage of partitioning; hence, all splitting decisions made while growing the tree to its maximal size are invariant to the values of $Q(O_i), i = 1 \ldots n$.

S.6.2 Pruning

Once a maximally-sized tree $\mathcal{T}_{max}$ is obtained, the second stage of the CART algorithm involves generating a sequence of candidate trees from which a final tree can be selected. The indicated sequence of candidate trees is generated using minimal cost-complexity pruning (Breiman et al., 1984, Sec. 3.3, 8.5).

For a given tree $\mathcal{T}$, let $\widetilde{\mathcal{T}}$ and $N(\mathcal{T}) = \#(\widetilde{\mathcal{T}})$ respectively denote the set and number of terminal nodes. Define the loss of the tree $\mathcal{T}$ as total loss in all terminal nodes: $R(\mathcal{T}) = \sum_{\tau \in \widetilde{\mathcal{T}}} R(\tau)$. Finally, let the cost-complexity of a tree $\mathcal{T}$ be defined as $R_\alpha(\mathcal{T}) = R(\mathcal{T}) + \alpha N(\mathcal{T})$, where α is a non-negative real number called the complexity parameter.

Paraphrasing Breiman et al. (1984, Sec. 8.5), minimal cost complexity pruning generates a decreasing sequence of subtrees $\mathcal{T}_{max} \succ \mathcal{T}_1 \succ \mathcal{T}_2 \succ \cdots \succ \{\tau_1\}$ and an increasing sequence of

complexity parameters $\alpha_1 < \alpha_2 < \cdots$ such that $\mathcal{T}_k$ is the smallest subtree of $\mathcal{T}_{max}$ for $\alpha_k \leq \alpha < \alpha_{k+1}$ that minimizes $R_\alpha(\mathcal{T})$. Breiman et al. (1984, Sec. 3.3) provide a detailed description of the process by which the sequence of subtrees is generated. Briefly, beginning with the smallest subtree $\mathcal{T}_1$ of $\mathcal{T}_{max}$ such that $R(\mathcal{T}_1) = R(\mathcal{T}_{max})$, CART begins the pruning process by considering all nodes τ from the tree T_1 and computing

$$g_1(\tau) = \begin{cases} \frac{R(\tau) - R(\mathcal{T}_{1,\tau})}{N(\mathcal{T}_{1,\tau}) - 1}, & \tau \notin \widetilde{\mathcal{T}}_{1,\tau} \\ +\infty, & \tau \in \widetilde{\mathcal{T}}_{1,\tau} \end{cases}$$

where $\widetilde{\mathcal{T}}_{1,\tau}$ denotes the subtree of $\mathcal{T}_1$ with root node τ. The node(s) minimizing this function are pruned, yielding the next tree in the sequence $\mathcal{T}_2$. This process is repeated until the root node of $\mathcal{T}_1$ is reached.

Critically, the process for pruning any $\mathcal{T}_k$ and hence generating $\mathcal{T}_{k+1}$ depends on mimimizing

$$g_k(\tau) = \begin{cases} \frac{R(\tau) - R(\mathcal{T}_{k,\tau})}{N(\mathcal{T}_{k,\tau}) - 1}, & \tau \notin \widetilde{\mathcal{T}}_k \\ +\infty, & \tau \in \widetilde{\mathcal{T}}_k \end{cases}$$

for each $\tau \in \mathcal{T}_k$. Evidently, the function $g_k(\tau)$ depends on the loss function only through $R(\tau) - R(\mathcal{T}_{k,\tau})$; applying Lemma S.6.1 shows this quantity does not depend on $Q(O_i), i = 1 \ldots n$. As a result, the decision made to prune away any subtree and consequently the sequence of candidate trees generated by this process will be invariant to $Q(O_i), i = 1 \ldots n$.

S.6.3 Choosing the best candidate tree via cross-validation

Selection of the optimally sized tree from the sequence of candidate trees is done using V-fold cross validation. Specifically, suppose a given data set $\mathcal{O} = (O_1, \ldots, O_n)$ is divided into V mutually exclusive subsets $\mathcal{O}_1, \ldots, \mathcal{O}_V$. Suppose that the procedure of Section S.6.2 generated M trees with complexity parameters $\alpha_1, \ldots, \alpha_M$ using the loss function $L_2(O, \psi(W); Q) = \psi(W)^2 + H(O)\psi(W) + Q(O)$. Define $\gamma_1 = 0$, $\gamma_j = \sqrt{\alpha_j \alpha_{j+1}}, j = 2, \ldots, M-1$, and $\gamma_M = \infty$; see Breiman et al. (1984, Sec. 3.4 & 8.5.2) for discussion. For each $v = 1, \ldots, V$ let $\mathcal{T}_m(\mathcal{L}_{-v}), m = 1 \ldots M$ be a sequence of trees built using the learning set $\mathcal{L}_{-v} = \mathcal{O} - \mathcal{O}_v$ as follows: (i) growing a tree $\mathcal{T}_{max,v}$ as described in Section S.6.1; (ii) determining the associated sequence of pruned trees using minimal cost complexity pruning as described in Section S.6.2; and, (iii) identifying the sequence elements that correspond to using the complexity parameters $\gamma_1, \ldots, \gamma_M$. For $m = 1, \ldots, M$, let $\widehat{\psi}(W; \mathcal{T}_m(\mathcal{L}_{-v}))$ be the prediction for a subject with covariate information W that is obtained using the tree $\mathcal{T}_m(\mathcal{L}_{-v})$. Then, the cross-validation error associated with γ_m is $C_m/(nV)$ where

$$\begin{aligned} C_m &= \sum_{v=1}^{V} \sum_{i=1}^{n} I(i \in \mathcal{O}_v) L_2(O_i, \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v})); Q) \,, \\ &= \sum_{v=1}^{V} \sum_{i=1}^{n} I(i \in \mathcal{O}_v) \left[Q(O_i) + H(O_i) \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v})) + \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v}))^2 \right], \\ &= C^* + \sum_{v=1}^{V} \sum_{i=1}^{n} I(i \in \mathcal{O}_v) \left[H(O_i) \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v})) + \widehat{\psi}(W_i; \mathcal{T}_m(\mathcal{L}_{-v}))^2 \right], \end{aligned} \tag{S-8}$$

for $m = 1, \ldots, M$ and $C^* = \sum_{v=1}^{V} \sum_{i=1}^{n} I(i \in \mathcal{O}_v) Q(O_i)$. The optimal tree is now given by $\mathcal{T}_{m^*}(\mathcal{O})$, where $m^* = \text{argmin}_{m \in \{1,\ldots,M\}} C_m$ and $\mathcal{T}_m(\mathcal{O})$ is the m^{th} candidate tree built using the full dataset $\mathcal{O}$ (i.e., that corresponding to α_m). Evidently, the constant C^* plays no role in selecting the member of the sequence that minimizes $C_m, m = 1, \ldots, M$ and hence selection of the optimally sized tree is also invariant to $Q(O_i), i = 1 \ldots n$.