The joint survival super learner: A super learner for right-censored data

Anders Munch & Thomas A. Gerds

July 1, 2025

Abstract

Risk prediction models are widely used to guide real-world decision-making in areas such as healthcare and economics, and they also play a key role in estimating nuisance parameters in semiparametric inference. The super learner is a machine learning framework that combines a library of prediction algorithms into a meta-learner using cross-validated loss. In the context of right-censored data, careful consideration must be given to both the choice of loss function and the estimation of expected loss. Moreover, estimators such as inverse probability of censoring weighting (IPCW) require accurate modeling and estimates of the censoring distribution.

We propose a novel approach to super learning for survival analysis that jointly evaluates candidate learners for both the event time distribution and the censoring distribution. Our method imposes no restrictions on the algorithms included in the library, accommodates competing risks, and does not rely on a single pre-specified estimator of the censoring distribution. We establish theoretical guarantees for our proposed method, including a finite-sample oracle inequality. In a simulation study, our super learner was able to better account for different censoring mechanisms than existing methods. We demonstrate the practical utility of our method using prostate cancer data.

Keywords: *Competing risks, cross-validation, loss based estimation, right-censored data, super learner*

1 Introduction

Accurately predicting risk from time-to-event data is a central challenge in diverse research fields including epidemiology, economics and weather forecasting with applications in clinical decision making and policy interventions. For instance, in prostate cancer management, clinicians often need to estimate a patient's risk of disease progression and mortality over time to make informed decisions about treatment strategies such as active surveillance versus immediate intervention. Reliable time-to-event risk predictions can help tailor care to individual patients, avoid overtreatment, and allocate healthcare resources more effectively. Super learning [van der Laan et al., 2007], also known as ensemble learning or stacked regression [Wolpert, 1992, Breiman, 1996], provides a powerful approach to this problem by combining multiple candidate prediction models to reduce the risk of bias incurred by a single potentially mispecified model. In survival analysis, a super learner may combine a stack of Cox regression models with a stack of random survival forests [Gerds and Kattan, 2021, Section 8.4]. Such a strategy has recently produced KDpredict (`https://kdpredict.com/`)

a model which jointly predicts the risks of kidney failure and all-cause mortality at multiple time horizons based on different sets of covariates [Liu et al., 2024]. To evaluate the prediction performance of the learners the super learner behind KDpredict uses inverse probability of censoring weighting (IPCW) where the censoring distribution is estimated under the restrictive assumption that the censoring distribution does not depend on the covariates. This is a potential source of bias which is difficult to overcome with the currently available methods.

In this paper we propose the joint survival super learner, a new super learner designed to handle the specific challenges of ensemble learning with right-censored data. The joint survival super learner jointly learns prediction models for the event-time and censoring distributions. The joint survival super learner is based on the simple idea of an artificial competing risk model where censoring is included as a state of its own. Candidate learners of the event time hazard and censoring hazard functions are then assessed based on how well they predict the state occupation probabilities of the artificial competing risk model across time based on baseline covariate information. Our estimation framework allows for competing risks, avoids restrictive assumptions on the censoring distribution, and is also fully flexible with respect to the choice of learners. The latter is in contrast to other proposals which restrict the library of learners to specific model classes [Polley and van der Laan, 2011, Golmakani and Polley, 2020], see Section 3.

To analyse the theoretical properties of the joint survival super learner we focus on the discrete super learner which picks the model in the library with best estimated performance [van der Laan et al., 2007]. We provide theoretical guarantees for the performance of the joint survival super learner and in particular show that the discrete joint survival super learner is consistent under natural conditions and prove a finite-sample oracle inequality. We demonstrate how to construct a library for the joint survival super learner using common survival models, and how to obtain risk predictions from the resulting ensemble.

The rest of the paper is organized as follows. We introduce our notation and framework in Section 2. Section 3 introduces loss-based super learning and presents existing super learners for right-censored data. In Section 4 we define the joint survival super learner, while Section 5 provides theoretical guarantees. Section 6 reports the results of numerical experiments, and Section 7 illustrates the method on prostate cancer data. We conclude with a discussion in Section 8. Proofs are collected in Appendix A. Code and an implementation of the joint survival super learner are available at `https://github.com/amnudn/joint-survival-super-learner`.

2 Notation and framework

In a competing risk framework [Andersen et al., 2012], let T be a time to event variable, $D \in \{1, 2\}$ the cause of the event, and $X \in \mathcal{X}$ a vector of baseline covariates taking values in a bounded subset $\mathcal{X} \subset \mathbb{R}^p$, $p \in \mathbb{N}$. Let $\tau < \infty$ be a fixed prediction horizon. We use $\mathcal{Q}$ to denote the collection of all probability measures on $[0, \tau] \times \{1, 2\} \times \mathcal{X}$ such that $(T, D, X) \sim Q$ for some unknown $Q \in \mathcal{Q}$. For $j \in \{1, 2\}$, the cause-specific conditional cumulative hazard functions $\Lambda_j \colon [0, \tau] \times \mathcal{X} \to \mathbb{R}_+$ are defined as

$$\Lambda_j(t \mid x) = \int_0^t \frac{Q(T \in \mathrm{d}s, D = j \mid X = x)}{Q(T \geq s \mid X = x)}.$$

For ease of presentation we assume throughout that the map $t \mapsto \Lambda_j(t \mid x)$ is continuous for all x and j, however, all technical arguments extend naturally to the general case [Andersen

et al., 2012]. The event-free survival function conditional on covariates is

$$S(t \mid x) = \exp\{-\Lambda_1(t \mid x) - \Lambda_2(t \mid x)\}. \tag{1}$$

Let $\mathcal{M}_\tau$ denote the space of all conditional cumulative hazard functions on $[0,\tau] \times \mathcal{X}$. Any distribution $Q \in \mathcal{Q}$ can be characterized by

$$\begin{aligned} Q(\mathrm{d}t, j, \mathrm{d}x) = &\{S(t- \mid x)\Lambda_1(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=1\}} \\ &\{S(t- \mid x)\Lambda_2(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=2\}}, \end{aligned}$$

where $\Lambda_j \in \mathcal{M}_\tau$ for $j = 1, 2$ and H is the marginal distribution of the covariates.

We consider the right-censored setting in which we observe $O = (\tilde{T}, \tilde{D}, X)$, where $\tilde{T} = \min(T, C)$ for a right-censoring time C, $\Delta = \mathbb{1}\{T \leq C\}$, and $\tilde{D} = \Delta D$. Let $\mathcal{P}$ denote a set of probability measures on the sample space $\mathcal{O} = [0,\tau] \times \{0,1,2\} \times \mathcal{X}$ such that $O \sim P$ for some unknown $P \in \mathcal{P}$. We assume that the event times and the censoring times are conditionally independent given covariates, $T \perp\!\!\!\perp C \mid X$. This implies that any distribution $P \in \mathcal{P}$ is characterized by a distribution $Q \in \mathcal{Q}$ and a conditional cumulative hazard function for C given X [c.f., Begun et al., 1983, Gill et al., 1997]. We use $\Gamma \in \mathcal{M}_\tau$ to denote the cumulative hazard function of the conditional censoring distribution given covariates. For ease of presentation we assume that $t \mapsto \Gamma(t \mid x)$ is continuous for all x. We let $(t,x) \mapsto G(t \mid x) = \exp\{-\Gamma(t \mid x)\}$ denote the survival function of the conditional censoring distribution. The distribution P is characterized by

$$\begin{aligned} P(\mathrm{d}t, j, \mathrm{d}x) = &\{G(t- \mid x)S(t- \mid x)\Lambda_1(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=1\}} \\ &\{G(t- \mid x)S(t- \mid x)\Lambda_2(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=2\}} \\ &\{G(t- \mid x)S(t- \mid x)\Gamma(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=0\}} \\ = &\{G(t- \mid x)Q(\mathrm{d}t, j, \mathrm{d}x)\}^{\mathbb{1}\{j\neq 0\}} \\ &\{G(t- \mid x)S(t- \mid x)\Gamma(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=0\}}. \end{aligned} \tag{2}$$

Hence, we may write $\mathcal{P} = \{P_{Q,\Gamma} : Q \in \mathcal{Q}, \Gamma \in \mathcal{G}\}$ for some $\mathcal{G} \subset \mathcal{M}_\tau$. We also have H-almost everywhere

$$P(\tilde{T} > t \mid X = x) = S(t \mid x)G(t \mid x) = \exp\{-\Lambda_1(t \mid x) - \Lambda_2(t \mid x) - \Gamma(t \mid x)\}.$$

We assume that there exists $\kappa < \infty$ such that $\Lambda_j(\tau- \mid x) < \kappa$, for $j \in \{1,2\}$, and $\Gamma(\tau- \mid x) < \kappa$ for almost all $x \in \mathcal{X}$. This implies that $G(\tau- \mid x)$ is bounded away from zero for almost all $x \in \mathcal{X}$. Under these assumptions, the conditional cumulative hazard functions Λ_j and Γ can be identified from P by

$$\Lambda_j(t \mid x) = \int_0^t \frac{P(\tilde{T} \in \mathrm{d}s, \tilde{D} = j \mid X = x)}{P(\tilde{T} \geq s \mid X = x)}, \tag{3}$$

$$\Gamma(t \mid x) = \int_0^t \frac{P(\tilde{T} \in \mathrm{d}s, \tilde{D} = 0 \mid X = x)}{P(\tilde{T} \geq s \mid X = x)}. \tag{4}$$

Thus, we can consider Λ_j and Γ as operators which map from $\mathcal{P}$ to $\mathcal{M}_\tau$.

3 Loss-based super learning

Loss-based super learning requires a library of candidate models, or *learners*, a cross-validation algorithm, and a loss function for evaluating prediction performance on hold-out

samples. In this section we provide an abstract definition of loss-based super learning. As input to a super learner we need a data set $\mathcal{D}_n = \{O_i\}_{i=1}^n$ of i.i.d. observations from $P \in \mathcal{P}$ and a collection of candidate models or learners $\mathcal{A}$. Let Θ be the parameter space, which in our case is a class of functions representing different models. Each learner $a \in \mathcal{A}$ is a map $a \colon \mathcal{O}^n \to \Theta$ which takes a data set as input and returns an estimate $a(\mathcal{D}_n)$ of θ. Let $L \colon \Theta \times \mathcal{O} \to \mathbb{R}_+$ be a loss function, representing the performance of the model $\theta \in \Theta$ at the observation $O \in \mathcal{O}$, where lower values mean better performance.

The expected loss of a learner is estimated by splitting the data set $\mathcal{D}_n$ into K disjoint approximately equally sized subsets $\mathcal{D}_n^1, \mathcal{D}_n^2, \ldots, \mathcal{D}_n^K$ and then calculating the cross-validated loss

$$\hat{R}_n(a; L) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{D}_n^k|} \sum_{O_i \in \mathcal{D}_n^k} L\big(a(\mathcal{D}_n^{-k}), O_i\big), \quad \text{with} \quad \mathcal{D}_n^{-k} = \mathcal{D}_n \setminus \mathcal{D}_n^k. \tag{5}$$

The subset $\mathcal{D}_n^{-k}$ is referred to as the k'th training sample, while $\mathcal{D}_n^k$ is referred to as the k'th test or hold-out sample. The discrete super learner is defined as

$$\hat{a}_n = \operatorname*{argmin}_{a \in \mathcal{A}} \hat{R}_n(a; L),$$

and depends on the library of learners and on the specific partition of the data into the folds $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^K$.

When designing a super learner for right-censored data, particular care must be taken in the choice of loss function and in the estimation of the expected loss. A commonly used loss function for right-censored is the partial log-likelihood loss [e.g., Li et al., 2016, Yao et al., 2017, Lee et al., 2018, Katzman et al., 2018, Gensheimer and Narasimhan, 2019, Lee et al., 2021, Kvamme and Borgan, 2021]. This loss function is also suggested for super learning with right-censored data by Polley and van der Laan [2011], where data is assumed to be observed in discrete time. However, the partial log-likelihood loss does not work well as a measure of performance in hold-out samples observed in continuous time. The reason is that the partial log-likelihood loss assigns an infinite value for any learner that predicts piece-wise constant cumulative hazard functions when there are event times in the test set which do not occur in the learning set. This problem occurs with prominent survival learners including the Kaplan-Meier estimator, random survival forests, and semi-parametric Cox regression models, and these learners cannot be included in the library of the super learner proposed by Polley and van der Laan [2011]. When a proportional hazards model is assumed, the baseline hazard function can be profiled out of the likelihood [Cox, 1972]. The cross-validated partial log-likelihood loss [Verweij and van Houwelingen, 1993] has therefore been suggested as a loss function for super learning by Golmakani and Polley [2020]. This choice of loss function restricts the library of learners to include only Cox proportional hazards models, and hence excludes many learners such as, e.g., random survival forests, additive hazards models, and accelerated failure time models.

Alternative approaches for super learning with right-censored data use an inverse probability of censoring weighted (IPCW) loss function [Graf et al., 1999, van der Laan and Dudoit, 2003, Molinaro et al., 2004, Keles et al., 2004, Hothorn et al., 2006, Gerds and Schumacher, 2006, Gonzalez Ginestet et al., 2021], censoring unbiased transformations [Fan and Gijbels, 1996, Steingrimsson et al., 2019], or pseudo-values [Andersen et al., 2003, Mogensen and Gerds, 2013, Sachs et al., 2019]. All these methods rely on an estimator of the censoring distribution, and their drawback is that this estimator has to be pre-specified. Recent work by Han et al. [2021] and Westling et al. [2021] circumvents the need to pre-specify a censoring model by iterating between estimation of the outcome and censoring models. However, this iterative procedure is in general not guaranteed to converge to the true data-generating mechanism [Munch, 2024, Appendix A.4].

4 The joint survival super learner

4.1 An artificial competing risk model

The main idea of the joint survival super learner is to jointly use learners for Λ_1, Λ_2, and Γ, and the relations in equation (2), to learn a feature of the observed data distribution P. The discrete joint survival super learner ranks a tuple of learners for the tuple of the cumulative hazard functions $(\Lambda_1, \Lambda_2, \Gamma)$ based on how well they jointly model the observed data. To formally introduce the joint survival super learner, we define the process

$$\eta(t) = \mathbb{1}\{\tilde{T} \leq t, \tilde{D} = 1\} + 2\,\mathbb{1}\{\tilde{T} \leq t, \tilde{D} = 2\} - \mathbb{1}\{\tilde{T} \leq t, \tilde{D} = 0\}, \quad \text{for} \quad t \in [0, \tau],$$

which takes values in $\{-1, 0, 1, 2\}$. The four values represent four mutually exclusive states. Specifically, value 0 represents the state where the individual is still event-free and uncensored, value 1 the state where the event of interest has occurred, value 2 the state where a competing risk has occurred, and value -1 the state where the observation is right-censored. The state occupation probabilities given baseline covariates X are given by the function

$$F(t, l, x) = P(\eta(t) = l \mid X = x), \tag{6}$$

for all $t \in [0, \tau]$, $l \in \{-1, 0, 1, 2\}$, and $x \in \mathcal{X}$.

The joint survival super learner is a super learner for the function-valued parameter $\theta(P) = F$ which is identified through equation (6). Under conditional independent censoring each tuple $(\Lambda_1, \Lambda_2, \Gamma, H)$ characterizes a distribution $P \in \mathcal{P}$, c.f. equation (2), which in turn determines (F, H). Hence, a learner for F can be constructed from learners for Λ_1, Λ_2, and Γ as follows:

$$\begin{aligned}
F(t, 0, x) &= P(\tilde{T} > t \mid X = x) = \exp\{-\Lambda_1(t \mid x) - \Lambda_2(t \mid x) - \Gamma(t \mid x)\}, \\
F(t, 1, x) &= P(\tilde{T} \leq t, \tilde{D} = 1 \mid X = x) = \int_0^t F(s-, 0, x)\Lambda_1(\mathrm{d}s \mid x), \\
F(t, 2, x) &= P(\tilde{T} \leq t, \tilde{D} = 2 \mid X = x) = \int_0^t F(s-, 0, x)\Lambda_2(\mathrm{d}s \mid x), \\
F(t, -1, x) &= P(\tilde{T} \leq t, \tilde{D} = 0 \mid X = x) = \int_0^t F(s-, 0, x)\Gamma(\mathrm{d}s \mid x).
\end{aligned} \tag{7}$$

Equation (7) implies that a library for the joint survival super learner can by build from three libraries of learners: $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ contain learners for the conditional cause-specific cumulative hazard functions Λ_1 and Λ_2, respectively, and $\mathcal{B}$ contains learners for the conditional cumulative hazard function of the censoring distribution. Based on the Cartesian product of libraries of learners for $(\Lambda_1, \Lambda_2, \Gamma)$ we construct a library $\mathcal{F}$ of learners for F:

$$\mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B}) = \{\varphi_{a_1, a_2, b} : a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2, b \in \mathcal{B}\}, \tag{8}$$

where in correspondence with the relations in equation (7),

$$\begin{aligned}
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, 0, x) &= \exp\{-a_1(\mathcal{D}_n)(s \mid x) - a_2(\mathcal{D}_n)(s \mid x) - b(\mathcal{D}_n)(s \mid x)\}, \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, 1, x) &= \int_0^t \varphi_{a_1,a_2,b}(\mathcal{D}_n)(s-, 0, x)a_1(\mathcal{D}_n)(\mathrm{d}s \mid x), \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, 2, x) &= \int_0^t \varphi_{a_1,a_2,b}(\mathcal{D}_n)(s-, 0, x)a_2(\mathcal{D}_n)(\mathrm{d}s \mid x), \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, -1, x) &= \int_0^t \varphi_{a_1,a_2,b}(\mathcal{D}_n)(s-, 0, x)b(\mathcal{D}_n)(\mathrm{d}s \mid x).
\end{aligned}$$

Notably, the libraries $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}$ can be constructed using standard software for survival analysis. In `R`, for instance, we can construct Cox models as learners using the `survival`-package [Therneau, 2022], and we can construct random survival forests as learners using the `randomForestSRC`-package [Ishwaran and Kogalur, 2025].

To evaluate how well a function F predicts the process η we use the integrated Brier score [Graf et al., 1999] $\bar{B}_\tau(F, O) = \int_0^\tau B_t(F, O)\,\mathrm{d}t$, where B_t is the Brier score [Brier et al., 1950] at time $t \in [0, \tau]$,

$$B_t(F, O) = \sum_{l=-1}^{2} \left(F(t, l, X) - \mathbb{1}\{\eta(t) = l\}\right)^2.$$

The Brier score is here the squared prediction error across all of the four states. Based on a split of a data set $\mathcal{D}_n$ into K disjoint approximately equally sized subsets (c.f., Section 3), each learner $\varphi_{a_1,a_2,b}$ in the library $\mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$ is evaluated using the cross-validated loss,

$$\hat{R}_n(\varphi_{a_1,a_2,b}; \bar{B}_\tau) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|\mathcal{D}_n^k|} \sum_{O_i \in \mathcal{D}_n^k} \bar{B}_\tau\big(\varphi_{a_1,a_2,b}(\mathcal{D}_n^{-k}), O_i\big),$$

and the discrete joint survival super learner is

$$\hat{\varphi}_n = \underset{(a_1,a_2,b)\in \mathcal{A}_1\times\mathcal{A}_2\times\mathcal{B}}{\operatorname{argmin}} \hat{R}_n(\varphi_{a_1,a_2,b}; \bar{B}_\tau).$$

4.2 Obtaining risk predictions

The joint survival super learner estimates the function F which depends on the censoring distribution and is therefore typically not of direct interest in itself. For instance, in the prostate cancer example we consider in Sections 6 and 7, the function F denotes the conditional probability that a patient with a certain set of baseline characteristics will, before time point t, have had tumor recurrence before leaving the study (state 1), have died without tumor recurrence before leaving the study (state 2), have left the study (state -1), or be alive and part of the study without tumor recurrence (state 0). The reference to still being part of the study is irrelevant from a new patient's perspective. We here demonstrate how clinically relevant risk predictions can instead be obtained from the joint survival super learner.

We recall that we work under the assumption of conditional independent censoring and positivity, as introduced in Section 2. Under these assumptions it follows by equations (3) and (4) and the definition of F that

$$\Lambda_j(t, x) = \int_0^t \frac{F(\mathrm{d}s, j, x)}{F(s-, 0, x)}, \quad j \in \{1, 2\}. \tag{9}$$

Cause-specific risk predictions can be obtained from Λ_1 and Λ_2 using the formula [e.g., Benichou and Gail, 1990, Ozenne et al., 2017],

$$Q(T \le t, D = j \mid X = x) = \int_0^t \exp\left\{-\Lambda_1(u \mid x) - \Lambda_2(u \mid x)\right\} \Lambda_j(\mathrm{d}u \mid x), \quad j \in \{1, 2\}. \tag{10}$$

Hence, given the joint survival super learner's estimate of F we can use equation (9) to obtain estimates of the cause-specific cumulative hazard functions Λ_j, which can in turn be used to obtain estimates of the cause-specific risks through equations (1). For instance, in the prostate cancer example, this expression will provide the conditional probability that a

patient with a certain set of baseline characteristics will, before time point t, have had tumor recurrence, have died without tumor recurrence, or be alive without tumor recurrence.

We have suggested to implement the joint survival super learner by building a library using learners of the cause-specific cumulative hazard functions, Λ_j, and the cumulative hazard function for censoring, Γ. With this implementation we can directly input the highest ranked tuple of cause-specific hazard functions (Λ_1, Λ_2) provided by the joint survival super learner as input to equation (10).

5 Theoretical guarantees

The use of cross-validation underlying the joint survival super learner is an intuitively reasonable procedure for fair model selection without overfitting. In this section, we follow the works of van der Laan and Dudoit [2003] and van der Vaart et al. [2006] to provide a theoretical justification for this practice in the form of a finite-sample oracle inequality. We begin by demonstrating that minimizing the integrated Brier score, as defined in Section 4, is statistically meaningful, in that perfect minimisation recovers the parameter of the data-generating distribution. Together with our finite-sample oracle inequality (Proposition 5.2 below), this implies that the joint survival super learner is consistent when it is based on a library that includes a consistent learner. Another consequence of our finite-sample oracle inequality is that the joint survival super learner converges at (nearly) the optimal rate achievable within the library. This statement is made precise in Corollary 5.3 and the following discussion.

A sensible loss function should attain the minimal expected value at the parameter corresponding to the data-generating distribution. Loss functions with this property are known as proper scoring rules, and as strictly proper scoring rules if the minimize is unique [Gneiting and Raftery, 2007]. Absence of properness makes it unclear why minimizing the (estimated) expected loss is interesting. Proposition 5.1 is a formal statement of the fact that the integrated Brier score, as defined in our setting (c.f., Section 4), is a strictly proper scoring rule. To state this result, recall that the function F implicitly depends on the data-generating probability measure $P \in \mathcal{P}$ but that this was suppressed in the notation. We now make this dependence explicit by writing F_P for the function determined by a given $P \in \mathcal{P}$ in accordance with equation equation (6). In the following we let $\mathcal{H}_\mathcal{P} = \{F_P : P \in \mathcal{P}\}$.

Proposition 5.1. *If $P \in \mathcal{P}$ then*

$$F_P = \operatorname*{argmin}_{F \in \mathcal{H}_\mathcal{P}} \mathbb{E}_P\left[\bar{B}_\tau(F, O)\right],$$

for all $l \in \{-1, 0, 1, 2\}$, almost all $t \in [0, \tau]$, and P-almost all $x \in \mathcal{X}$.

Proof. See Appendix A.1. □

To evaluate the performance of the joint survival super learner we might benchmark it against the data-generating F_P, as this has smallest expected loss by Proposition 5.1. A more nuanced comparison is to benchmark it against the best learner available given the library and the training data. This is the so-called oracle learner, formally defined as

$$\tilde{\varphi}_n = \operatorname*{argmin}_{\varphi \in \mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})} \tilde{R}_n(\varphi; \bar{B}_\tau), \quad \text{with} \quad \tilde{R}_n(\varphi; \bar{B}_\tau) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_P\left[\bar{B}_\tau\big(\varphi(\mathcal{D}_n^{-k}), O\big) \mid \mathcal{D}_n^{-k}\right],$$

where we use $\mathbb{E}_P$ to denote expectation under the distribution P for a new observation O independent of $\mathcal{D}_n^{-k}$. Like the joint survival super learner, the oracle learner depends on the library of learners and on the specific data partitions, but unlike the joint survival super learner, it also depends on the unknown data-generating distribution.

In the following, we equip the space $\mathcal{H}_{\mathcal{P}}$ with the norm

$$\|F\|_P = \left\{ \sum_{l=-1}^{2} \int_0^\tau \mathbb{E}_P \left[F(t,l,X)^2 \right] \mathrm{d}t \right\}^{1/2}. \tag{11}$$

This norm is equal to the excess risk $\mathbb{E}_P\left[\bar{B}_\tau(F,O)\right] - \mathbb{E}_P\left[\bar{B}_\tau(F_P,O)\right]$ by Lemma A.1 in the Appendix, and is thus a natural performance measure. For simplicity of presentation we take n and the data partitions to be such that $|\mathcal{D}_n^{-k}| = n/K$ with K fixed. We will allow the number of learners to grow with n and write $\mathcal{F}_n = \mathcal{F}(\mathcal{A}_{1,n}, \mathcal{A}_{2,n}, \mathcal{B}_n)$ as short-hand notation emphasing the dependence on the sample size. We now state a finite-sample inequality that bounds the performance of the joint survival super learner relative to that of the oracle learner.

Proposition 5.2. *For all $P \in \mathcal{P}$, $n \in \mathbb{N}$, $k \in \{1,\dots,K\}$, and $\delta > 0$,*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_P\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] \leq (1+2\delta)\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_P\left[\|\tilde{\varphi}_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] \\ + (1+\delta)16K\tau\left(13 + \frac{12}{\delta}\right)\frac{\log(1+|\mathcal{F}_n|)}{n}.\end{aligned}$$

Proof. See Appendix A.2. □

The expectations in Proposition 5.2 reflect a mild abuse of notation, in that they are formally taken with respect to the product measure P^n for the whole data set $\mathcal{D}_n$. This means that we are quantifying the average performance of the joint survival super learner across average training data. As for many finite-sample oracle inequalities, this result is of little direct practical utility because the right hand-side depends on data-dependent, unknown quantities. However, it does quantify how the number of folds, the time horizon, and the number of learners in the library can be expected to influence the performance. The result has the following asymptotic consequences.

Corollary 5.3. *Assume that $|\mathcal{F}_n| = O(n^q)$, for some $q \in \mathbb{N}$ and that there exists a sequence $\varphi_n \in \mathcal{F}_n$, $n \in \mathbb{N}$, such that $\mathbb{E}_P\left[\|\varphi_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] = C_P + O(n^{-\alpha})$, for some $\alpha \leq 1$ and $C_P \geq 0$.*

(a) If $\alpha = 1$ then $\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_P\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] = C_P + O(\log(n)^{1+\varepsilon}n^{-1})$, $\forall \varepsilon > 0$.

(b) If $\alpha < 1$ then $\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_P\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] = C_P + O(n^{-\alpha})$.

Proof. See Appendix A.2. □

While Proposition 5.2 provided a precise finite-sample bound on the average price we pay for using cross-validation, Corollary 5.3 states that this price is asymptotically vanishing, up to poly-logarithmic terms, as long as the number of learners in the library grows with sample size at a polynomial rate. The situation $C_P = 0$ corresponds to a setting where the library includes a consistent learner. Cases (a) and (b) correspond to situations where

the oracle learner achieves, respectively, a parametric or non-parametric asymptotic rate of convergence.

To illustrate the content of Corollary 5.3, consider first a situation where we use a library with an increasing number of semi-parametric Cox models with different interaction terms, as well as several Poisson regression models based on different discretisations of the time scale. Each of these models will independently achieve a parametric rate of convergence, and hence item (a) of Corollary 5.3 states that the joint survival super learner based on this library will achieve a near-parametric rate of convergence. The constant C_P can be taken equal to the distance to the least false model in the library, and so the joint survival super learner will approximate the least false model in the library at a near-parametric rate. Another situation appears if we add more flexible models to the library, such as Cox lasso or random survival forests. These models typically converge at non-parametric rates, with the fastest rate depending on the unknown data-generating distribution. Item (b) of Corollary 5.3 shows that the joint survival super learner achieves the same convergence rate as the best-performing algorithm in the library, without any knowledge of the data-generating distribution.

6 Numerical experiments

In this section we report results from a simulation study where we consider estimation of the conditional survival function. In the first part, we compare the joint survival super learner to two IPCW based discrete super learners that use either the Kaplan-Meier estimator or a Cox model to estimate the censoring probability [Gonzalez Ginestet et al., 2021]. In the second part we compare the joint survival super learner to the super learner proposed by Westling et al. [2021].

In both parts we use the same data-generating mechanism. We generate data according to a distribution motivated from a real data set in which censoring depends on the baseline covariates. We simulate data based on the prostate cancer study of Kattan et al. [2000]. The outcome of interest is the time to tumor recurrence, and five baseline covariates are used to predict outcome: prostate-specific antigen (PSA, ng/mL), Gleason score sum (GSS, values between 6 and 10), radiation dose (RD), hormone therapy (HT, yes/no) and clinical stage (CS, six values). The study was designed such that a patient's radiation dose depended on when the patient entered the study [Gerds et al., 2013]. This in turn implies that the time of censoring depends on the radiation dose. The data were re-analysed in [Gerds et al., 2013] where a sensitivity analysis was conducted based on simulated data. Here we use the same simulation setup, where event and censoring times are generated according to parametric Cox-Weibull models estimated from the original data, and the covariates are generated according to either marginal Gaussian normal or binomial distributions estimated from the original data [c.f., Gerds et al., 2013, Section 4.6]. We refer to this simulation setting as 'dependent censoring'. We also considered a simulation setting where data were generated in the same way, except that censoring was generated completely independently. We refer to this simulation setting as 'independent censoring'.

For all super learners we use a library consisting of three learners: The Kaplan-Meier estimator [Kaplan and Meier, 1958, Gerds, 2019], a Cox model with main effects [Cox, 1972, Therneau, 2022], and a random survival forest [Ishwaran et al., 2008, Ishwaran and Kogalur, 2025]. We use the same library to learn the outcome distribution and the censoring distribution. The three learners in our library of learners can be used to learn the cumulative hazard functions of the outcome and the censoring distribution. The latter works by training the learner on the data set $\mathcal{D}_n^c$, where $\mathcal{D}_n^c = \{O_i^c\}_{i=1}^n$ with $O_i^c = (\tilde{T}_i, 1 - \Delta_i, X_i)$. When we

say that we use a learner for the cumulative hazard function of the outcome to learn the cumulative hazard function of the censoring time, we mean that the learner is trained on $\mathcal{D}_n^c$.

We compare the joint survival super learner to two IPCW based super learners: The first super learner, called IPCW(Cox), uses a Cox model with main effects to estimate the censoring probabilities, while the second super learner, called IPCW(KM), uses the Kaplan-Meier estimator to estimate the censoring probabilities. The Cox model for the censoring distribution is correctly specified in both simulation settings while the Kaplan Meier estimator only estimates the censoring model correctly in the simulation setting where censoring is independent. Both IPCW super learners are fitted using the `R`-package `riskRegression` [Gerds et al., 2023]. The IPCW super learners use the integrated Brier score up to a fixed time horizon (36 months). The marginal risk of the event before this time horizon is $\approx 24.6\%$. Under the 'dependent censoring' setting the marginal censoring probability before the time horizon is $\approx 61.9\%$. Under the 'independent censoring' setting the marginal censoring probability before this time horizon is $\approx 38.7\%$.

Each super learner provides a learner for the cumulative hazard function for the outcome of interest. From the cumulative hazard function a risk prediction model can be obtained as described in Section 4.2 (with the special case of $\Lambda_2 = 0$). We measure the performance of each super learner by calculating the index of prediction accuracy (IPA) [Kattan and Gerds, 2018] at a fixed time horizon (36 months) for the risk prediction model provided by the super learner. The IPA is 1 minus the ratio between the model's Brier score and the null model's Brier score, where the null model is the model that does not use any covariate information. The IPA is approximated using a large ($n = 20,000$) independent data set of uncensored data. As a benchmark we calculate the performance of the risk prediction model chosen by the oracle selector, which uses the large data set of uncensored event times to select the learner with the highest IPA.

The results are shown in Figure 1. We see that in the scenario where censoring depends on the covariates, using the Kaplan-Meier estimator to estimate the censoring probabilities provides a risk prediction model with an IPA that is lower than the risk prediction model provided by the joint survival super learner. The performance of the risk prediction model selected by the joint survival super learner is similar to the risk prediction model selected by the IPCW(Cox) super learner which a priori uses a correctly specified model for the censoring distribution. Both these risk prediction models are close to the performance of the oracle, except for small sample sizes.

We next compare the joint survival super learner to the super learner survSL [Westling et al., 2021]. This is another super learner which like the joint survival super learner works without a pre-specified censoring model. Both the joint survival super learner and survSL provide a prediction model for the event time outcome and also for the probability of being censored. Hence, we compare the performance of these methods with respect to both the outcome and the censoring distribution. Again we use the IPA to quantify the predictive performance.

The results are shown in Figures 2 and 3. We see that for most sample sizes, the joint survival super learner selected prediction models for both censoring and outcome which have similar or higher IPA compared to the prediction models selected by survSL.

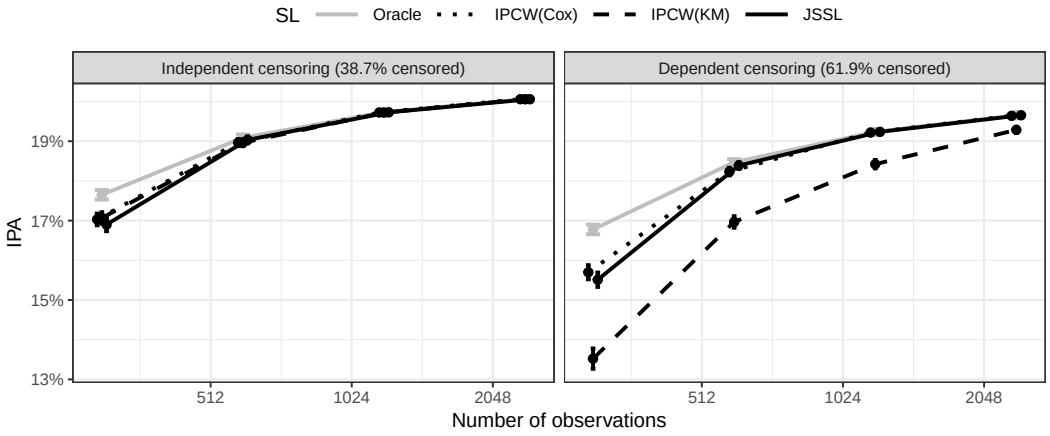

Figure 1: For the risk prediction models provided by each of the super learners, the IPA is plotted against sample size. The results are averages across 1000 simulated data sets and the error bars are used to quantify the Monte Carlo uncertainty. JSSL denotes the joint survival super learner.

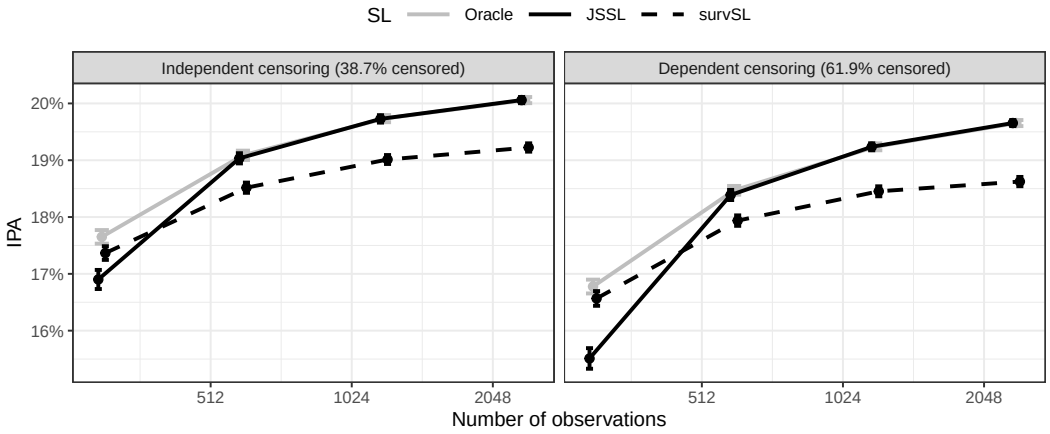

Figure 2: For the risk prediction models of the outcome provided by each of the super learners, the IPA at the fixed time horizon is plotted against sample size. The results are averages across 1000 repetitions and the error bars are used to quantify the Monte Carlo uncertainty.

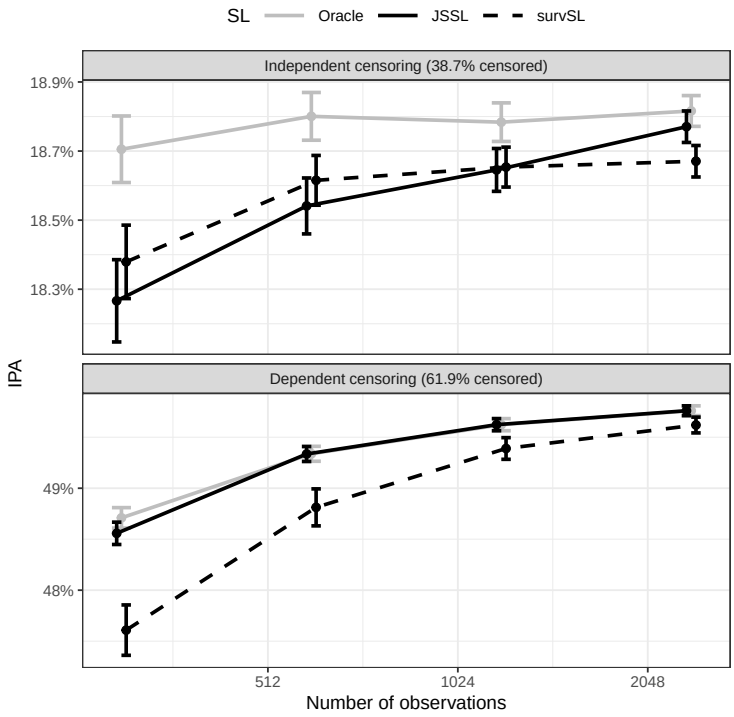

Figure 3: For the risk prediction models of the censoring model provided by each of the super learners, the IPA at the fixed time horizon is plotted against sample size. The results are averages across 1000 repetitions and the error bars are used to quantify the Monte Carlo uncertainty. JSSL denotes the joint survival super learner. JSSL denotes the joint survival super learner.

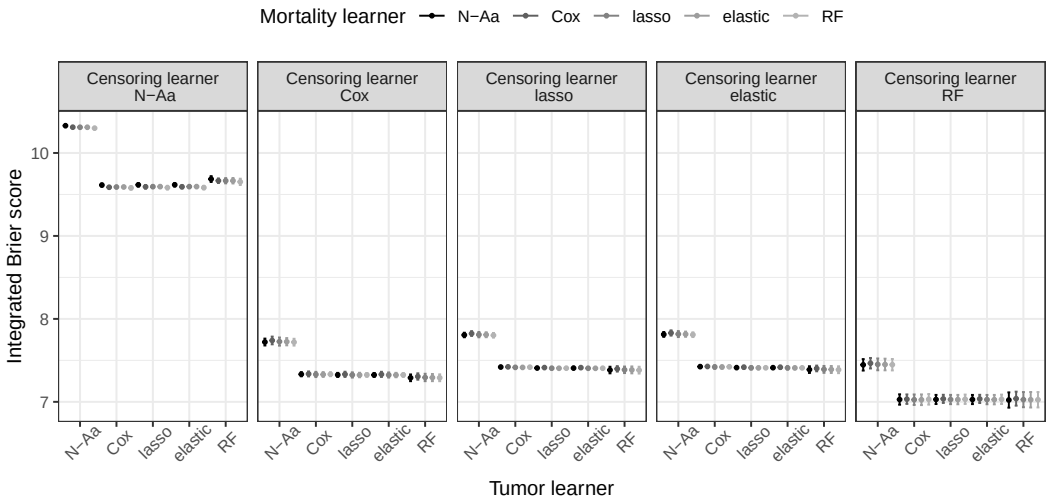

Figure 4: The results of applying the 125 combinations of learners to the prostate cancer data set. The error bars are based on five repetitions using different splits. We refer to learners of Λ_1, Λ_2, and Γ as 'Tumor learner', 'Mortality learner', and 'Censoring learner', respectively.

7 Prostate cancer study

In this section we use the prostate cancer data of Kattan et al. [2000] to illustrate the use of the joint survival super learner in the presence of competing risks. We have introduced the data in Section 6. The data consists of 1,042 patients who are followed from start of followup until tumor recurrence, death without tumor recurrence or end of followup (censored) whatever came first. We use the joint survival super learner to rank libraries of learners for the cause-specific cumulative hazard functions of tumor recurrence, death without tumor recurrence, and censoring. The libraries of learners each include five learners: the Nelson-Aalen estimator, three Cox regression models (unpenalized, lasso, elastic net) each including additive effects of the 5 covariates (c.f., Section 6), and a random survival forest. We use the same set of learners to learn the cumulative hazard function of tumor recurrence Λ_1, the cumulative hazard function of death without tumor recurrence Λ_2, and the cumulative hazard function of the conditional censoring distribution Γ.

This gives a library consisting of $5^3 = 125$ learners for the conditional state occupation probability function F defined in equation (6). We use five folds for training and testing the models, and we repeat training and evaluation five times with different splits. The integrated Brier score (defined in Section 4) for all learners are shown in Figure 4. We see that the prediction performance is mostly affected by the choice of learner for the censoring distribution. Several combinations of learners give similar performance as measured by the integrated Brier score, as long as a random forest is used to model the censoring distribution.

8 Discussion

A major advantage of the joint survival super learner is that the performance of each combination of learners can be estimated without additional nuisance parameters. A potential drawback of our approach is that we are evaluating the loss of the learners on the level of the

observed data distribution while the target of the analysis is either the event time distribution, or the censoring distribution, or both. Specifically, the finite-sample oracle inequality in Proposition 5.2 concerns the function F, which is a feature of $P \in \mathcal{P}$, while what we are typically interested in is Λ_j or S, which are features of $Q \in \mathcal{Q}$. We emphasize that while the joint survival super learner provides us with estimates of Λ_j and Γ based on libraries $\mathcal{A}_j$ and $\mathcal{B}$, the performance of these learners is not assessed directly for their respective target parameters, but only indirectly via the performance of F. For settings without competing risks, our numerical studies suggest that measuring the performance of F also leads to good performance for estimation of S.

Our proposed super learner can be implemented with a broad library of learners and using existing software. Furthermore, while the library $\mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$ consists of $|\mathcal{A}_1||\mathcal{A}_2||\mathcal{B}|$ many learners, we only need to fit $|\mathcal{A}_1| + |\mathcal{A}_2| + |\mathcal{B}|$ many learners in each fold. To evaluate the performance of each learner we need to perform $|\mathcal{A}_1||\mathcal{A}_2||\mathcal{B}|$ many operations to calculate the integrated Brier score in each hold-out sample, one for each combination of the fitted models, but these operations are often negligible compared to fitting the models. Hence the joint survival super learner is essentially not more computationally demanding than any procedure that uses super learning to learn Λ_1, Λ_2, and Γ separately. While our proposal is based on constructing the library $\mathcal{F}$ from libraries for learning Λ_1, Λ_2, and Γ, it could also be of interest to consider learners that estimate F directly.

In our numerical studies, we only considered learners of Λ_j and Γ that provide cumulative hazard functions which are piece-wise constant in the time argument. This simplifies the calculation of F as the integrals in equation (7) reduce to sums. When Λ_j or Γ are absolutely continuous in the time argument, calculating F is more involved, but we expect that a good approximation can be achieved by discretisation.

Our original motivation for developing the joint survival super learner was for use within the framework of targeted or debiased machine learning – a general methodology that combines flexible, data-adaptive estimation with asymptotically valid inference for low-dimensional target parameters [van der Laan and Rose, 2011, Chernozhukov et al., 2018]. In settings with right-censored competing risks, the relevant nuisance parameters often include the cause-specific and censoring cumulative hazard functions [e.g., van der Laan and Robins, 2003, Rytgaard and van der Laan, 2022]. The joint survival super learner immediately provides estimates of these nuisance parameters and is hence particularly well suited for targeted and debiased machine learning. We leave the study of the joint survival super learner in the context of targeted and debiased machine learning for a future paper.

References

P. K. Andersen, J. P. Klein, and S. Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 2003.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

J. M. Begun, W. J. Hall, W.-M. Huang, and J. A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, 11(2):432–452, 1983.

J. Benichou and M. H. Gail. Estimates of absolute cause-specific risk in cohort studies. *Biometrics*, pages 813–826, 1990.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

G. W. Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Routledge, 1996.

M. F. Gensheimer and B. Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

T. A. Gerds. *prodlim: Product-Limit Estimation for Censored Event History Analysis*, 2019. URL https://CRAN.R-project.org/package=prodlim. R package version 2019.11.13.

T. A. Gerds and M. W. Kattan. *Medical risk prediction models: with ties to machine learning*. CRC Press, 2021.

T. A. Gerds and M. Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–2184, 2013.

T. A. Gerds, J. S. Ohlendorff, and B. Ozenne. *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*, 2023. URL https://CRAN.R-project.org/package=riskRegression. R package version 2023.03.22.

R. D. Gill, M. J. van der Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

M. K. Golmakani and E. C. Polley. Super learner for survival data prediction. *The International Journal of Biostatistics*, 16(2):20190065, 2020.

P. Gonzalez Ginestet, A. Kotalik, D. M. Vock, J. Wolfson, and E. E. Gabriel. Stacked inverse probability of censoring weighted bagging: A case study in the infcarehiv register. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(1):51–65, 2021.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 1999.

X. Han, M. Goldstein, A. Puli, T. Wies, A. Perotte, and R. Ranganath. Inverse-weighted survival games. *Advances in Neural Information Processing Systems*, 34, 2021.

T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

H. Ishwaran and U. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2025. URL https://cran.r-project.org/package=randomForestSRC. R package version 3.3.3.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

M. W. Kattan and T. A. Gerds. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, 2018.

M. W. Kattan, M. J. Zelefsky, P. A. Kupelian, P. T. Scardino, Z. Fuks, and S. A. Leibel. Pretreatment nomogram for predicting the outcome of three-dimensional conformal radiotherapy in prostate cancer. *Journal of clinical oncology*, 18(19):3352–3359, 2000.

J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

S. Keles, M. van der Laan, and S. Dudoit. Asymptotically optimal model selection method with right censored outcomes. *Bernoulli*, 10(6):1011–1037, 2004.

E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

C. Lee, W. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

D. K. Lee, N. Chen, and H. Ishwaran. Boosted nonparametric hazards with time-dependent covariates. *Annals of Statistics*, 49(4):2101, 2021.

Y. Li, K. S. Xu, and C. K. Reddy. Regularized parametric regression for high-dimensional survival analysis. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 765–773. SIAM, 2016.

P. Liu, S. Sawhney, U. Heide-Jørgensen, R. R. Quinn, S. K. Jensen, A. Mclean, C. F. Christiansen, T. A. Gerds, and P. Ravani. Predicting the risks of kidney failure and death in adults with moderate to severe chronic kidney disease: multinational, longitudinal, population based, cohort study. *British Medical Journal*, 385, 2024.

U. B. Mogensen and T. A. Gerds. A random forest approach for competing risks based on pseudo-values. *Statistics in medicine*, 32(18):3102–3114, 2013.

A. M. Molinaro, S. Dudoit, and M. J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1): 154–177, 2004.

A. Munch. *Targeted learning with right-censored data*. Phd thesis, University of Copenhagen, 2024. URL https://publichealth.ku.dk/about-the-department/biostat/phd-theses/2023_munch.pdf.

B. Ozenne, A. L. Sørensen, T. Scheike, C. Torp-Pedersen, and T. A. Gerds. riskregression: Predicting the risk of an event using Cox regression models. *R Journal*, 9(2):440–460, 2017.

J. Pfanzagl and W. Wefelmeyer. *Contributions to a general asymptotic statistical theory*. Springer, 1982.

E. C. Polley and M. J. van der Laan. Super learning for right-censored data. In M. J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 249–258. Springer, 2011.

H. C. Rytgaard and M. J. van der Laan. Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Analysis*, pages 1–30, 2022.

H. C. Rytgaard, F. Eriksson, and M. J. van der Laan. Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*, 2021.

M. C. Sachs, A. Discacciati, Å. H. Everhov, O. Olén, and E. E. Gabriel. Ensemble prediction of time-to-event outcomes with competing risks: A case-study of surgical complications in Crohn's disease. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 68(5):1431–1446, 2019.

J. A. Steingrimsson, L. Diao, and R. L. Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 2019.

T. M. Therneau. *A Package for Survival Analysis in R*, 2022. URL https://CRAN.R-project.org/package=survival. R package version 3.4-0.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, 2003.

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

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

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.

P. J. Verweij and H. C. van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–2314, 1993.

T. Westling, A. Luedtke, P. Gilbert, and M. Carone. Inference for treatment-specific survival curves using machine learning. *arXiv preprint arXiv:2106.06602*, 2021.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

J. Yao, X. Zhu, F. Zhu, and J. Huang. Deep correlational learning for survival prediction from multi-modality data. In *International conference on medical image computing and computer-assisted intervention*, pages 406–414. Springer, 2017.

A Proofs

A.1 Strictly properness

Define $\bar{B}_{\tau,P}(F,o) = \bar{B}_\tau(F,o) - \bar{B}_\tau(F_P,o)$ and $R_P(F) = \mathbb{E}_P\left[\bar{B}_{\tau,P}(F,O)\right]$, where the integrated Brier score $\bar{B}_\tau$ was defined in Section 4. Recall the norm $\|\cdot\|_P$ defined in equation (11).

Lemma A.1. $R_P(F) = \|F - F_P\|_P^2$.

Proof. For any $t \in [0,\tau]$ and $l \in \{-1,0,1,2\}$ we have

$$\begin{aligned}
&\mathbb{E}_P\left[(F(t,l,X) - \mathbb{1}\{\eta(t)=l\})^2\right] \\
&= \mathbb{E}_P\left[(F(t,l,X) - F_P(t,l,X) + F_P(t,l,X) - \mathbb{1}\{\eta(t)=l\})^2\right] \\
&= \mathbb{E}_P\left[(F(t,l,X) - F_P(t,l,X))^2\right] + \mathbb{E}_P\left[(F_P(t,l,X) - \mathbb{1}\{\eta(t)=l\})^2\right] \\
&\quad + 2\,\mathbb{E}_P\left[(F(t,l,X) - F_P(t,l,X))(F_P(t,l,X) - \mathbb{1}\{\eta(t)=l\})\right] \\
&= \mathbb{E}_P\left[(F(t,l,X) - F_P(t,l,X))^2\right] + \mathbb{E}_P\left[(F_P(t,l,X) - \mathbb{1}\{\eta(t)=l\})^2\right],
\end{aligned}$$

where the last equality follows from the tower property. Hence, using Fubini, we have

$$\mathbb{E}_P\left[\bar{B}_\tau(F,O)\right] = \|F - F_P\|_P^2 + \mathbb{E}_P\left[\bar{B}_\tau(F_P,O)\right].$$

□

Proof of Proposition 5.1. The result follows from Lemma A.1. □

A.2 Oracle inequalities

Recall that we use $\mathcal{F}_n$ to denote a library of learners for the function F, and that $\hat{\varphi}$ and $\tilde{\varphi}$ denotes, respectively, the discrete super learner and the oracle learner for the library $\mathcal{F}_n$, c.f., Section 4.

Proof of Proposition 5.2. Minimizing the loss $\bar{B}_\tau$ is equivalent to minimizing the loss $\bar{B}_{\tau,P}$, so the discrete super learner and oracle according to $\bar{B}_\tau$ and $\bar{B}_{\tau,P}$ are identical. By Lemma A.1, $R_P(F) \geq 0$ for any $F \in \mathcal{H}_\mathcal{P}$, and so using Theorem 2.3 from [van der Vaart et al., 2006] with $p = 1$, we have that for all $\delta > 0$,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_P\left[R_P(\hat{\varphi}_n(\mathcal{D}_n^{-k}))\right] \\
&\leq (1+2\delta)\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_P\left[R_P(\tilde{\varphi}_n(\mathcal{D}_n^{-k}))\right] \\
&\quad + (1+\delta)\frac{16K}{n}\log(1+|\mathcal{F}_n|)\sup_{F\in\mathcal{H}_\mathcal{P}}\left\{M(F) + \frac{v(F)}{R_P(F)}\left(\frac{1}{\delta}+1\right)\right\}
\end{aligned}$$

where for each $F \in \mathcal{H}_{\mathcal{P}}$, $(M(F), v(F))$ is some Bernstein pair for the function $o \mapsto \bar{B}_{\tau,P}(F, o)$. As $\bar{B}_{\tau,P}(F, \cdot)$ is uniformly bounded by τ for any $F \in \mathcal{H}_{\mathcal{P}}$, it follows from section 8.1 in [van der Vaart et al., 2006] that $(\tau, 1.5\,\mathbb{E}_P\,[\bar{B}_{\tau,P}(F, O)^2])$ is a Bernstein pair for $\bar{B}_{\tau,P}(F, \cdot)$. Now, for any $a, b, c \in \mathbb{R}$ we have

$$\begin{aligned}(a-c)^2 - (b-c)^2 &= (a-b+b-c)^2 - (b-c)^2 \\ &= (a-b)^2 + (b-c)^2 + 2(b-c)(a-b) - (b-c)^2 \\ &= (a-b)\{(a-b) + 2(b-c)\} \\ &= (a-b)\{a+b-2c\},\end{aligned}$$

so using this with $a = F(t, l, x)$, $b = F_P(t, l, x)$, and $c = \mathbb{1}\{\eta(t) = l\}$, we have by Jensen's inequality

$$\begin{aligned}&\mathbb{E}_P\,[\bar{B}_{\tau,P}(F, O)^2] \\ &\leq 2\tau\,\mathbb{E}_P\left[\sum_{l=-1}^{2}\int_0^\tau \left\{(F(t,l,X) - \mathbb{1}\{\eta(t)=l\})^2 - (F_P(t,l,X) - \mathbb{1}\{\eta(t)=l\})^2\right\}^2 \mathrm{d}t\right] \\ &= 2\tau\,\mathbb{E}_P\left[\sum_{l=-1}^{2}\int_0^\tau (F(t,l,X) - F_P(t,l,X))^2 \right. \\ &\qquad\qquad \left. \times \{F(t,l,X) + F_P(t,l,X) - 2\mathbb{1}\{\eta(t)=l\}\}^2 \mathrm{d}t\right] \\ &\leq 8\tau\,\mathbb{E}_P\left[\sum_{l=-1}^{2}\int_0^\tau (F(t,l,X) - F_P(t,l,X))^2 \mathrm{d}t\right]. \\ &= 8\tau\|F - F_P\|_P^2.\end{aligned}$$

Thus when $v(F) = 1.5\,\mathbb{E}_P\,[\bar{B}_{\tau,P}(F, O)^2]$ we have by Lemma A.1

$$\frac{v(F)}{R_P(F)} = 1.5\frac{\mathbb{E}_P\,[\bar{B}_{\tau,P}(F, O)^2]}{\mathbb{E}_P\,[\bar{B}_{\tau,P}(F, O)]} \leq 12\tau,$$

and so using the Bernstein pairs $(\tau, 1.5\,\mathbb{E}_P\,[\bar{B}_{\tau,P}(F, O)^2])$ we have

$$\sup_{F \in \mathcal{H}_{\mathcal{P}}}\left\{M(F) + \frac{v(F)}{R_P(F)}\left(\frac{1}{\delta} + 1\right)\right\} \leq \tau\left(13 + \frac{12}{\delta}\right).$$

For all $\delta > 0$ we thus have

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_P\left[R_P(\hat{\varphi}_n(\mathcal{D}_n^{-k}))\right] \leq &(1+2\delta)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_P\left[R_P(\tilde{\varphi}_n(\mathcal{D}_n^{-k}))\right] \\ &+ (1+\delta)\log(1+|\mathcal{F}_n|)\tau\frac{16K}{n}\left(13 + \frac{12}{\delta}\right),\end{aligned}$$

and then the final result follows from Lemma A.1. □

Proof of Corollary 5.3. By definition of the oracle and Lemma A.1,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_P\left[\|\tilde{\varphi}_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] \leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_P\left[\|\varphi_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] = \mathbb{E}_P\left[\|\varphi_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right],$$

for all $n \in \mathbb{N}$, where the last equality follows because all the training sets $\mathcal{D}_n^{-k}$ have the same distribution. The result then follows from Proposition 5.2 by letting δ grow to zero with n, for instance as $\delta_n = \log(n)^{-\varepsilon}$ for some $\varepsilon > 0$. □