ORIGINAL ARTICLE

Journal Section

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

Anders Munch[1] | **Thomas A. Gerds**[1]

[1]Section of Biostatistics, University of Copenhagen

Correspondence
Anders Munch, Section of Biostatistics, University of Copenhagen, 1355 Copenhagen, Denmark
Email: a.munch@sund.ku.dk

Funding information

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 require accurate modeling and an estimator 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 a finite-sample bound on the average price we pay for using cross-validation, and show that this price vanishes asymptotically, up to poly-logarithmic terms, provided that the size of the library does not grow faster than at a polynomial rate in the sample size. We demonstrate the practical utility of our method using prostate cancer data and compare it to existing super learner algorithms for survival analysis using synthesized 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 various research fields, such as 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 risk prediction models 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 for example 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 it 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 article, 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 simultaneously learns prediction models for the event-time and censoring distributions. The joint survival super learner is based on a competing risks model for the observed data, in which censoring is included as a state of its own, such that at any time it is known in which state an individual is. We assume conditionally independent censoring and exploit well-known relationships between the observed data distribution on the one side and the partly unobserved distributions of the event time and the censoring time on the other. Learners for the event-time and censoring hazard functions are then assessed using the integrated Brier score across all states of the observed data. Our estimation framework thus naturally incorporates competing risks, avoids restrictive assumptions on the censoring distribution, and produces an estimator for the censoring distribution. Our approach 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), as we discuss in more detail in Section 3.

To analyse the theoretical properties of the joint survival super learner, we focus on the discrete super learner, which selects the model in the library with the 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 when the library of learners includes at least one consistent learner. We also derive a finite-sample oracle inequality for the discrete joint survival super learner. We demonstrate how to construct a library of learners using common methods for survival analysis and illustrate the use of the joint survival super learner using a prostate cancer data set.

The article is organized as follows. We introduce our notation and framework in Section 2. Section 3 introduces loss-based super learning and discusses other 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 a series 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 the Appendix. Code and an implementation of the joint survival super learner in R (R Core Team, 2024) are available at `https://github.com/amnudn/joint-survival-super-learner`.

2 | NOTATION AND FRAMEWORK

In a competing risks framework (Andersen et al., 2012) with J competing risks, let T be a time to event variable, $D \in \{1, 2, \ldots, J\}$ 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 $[0, \tau] \times \{1, 2, \ldots, J\} \times \mathcal{X}$ such that $(T, D, X) \sim Q$ for some unknown $Q \in \mathcal{Q}$. For $j \in \{1, 2, \ldots, J\}$, 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 from now on that $J = 2$ and 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 given by

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

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 $\mathcal{S} = [0, \tau] \times \{0, 1, 2\} \times \mathcal{X}$ 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

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

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 learners, a cross-validation algorithm, and a loss function for evaluating predictive performance on hold-out samples. Let $\mathcal{D}_n = \{O_i\}_{i=1}^n \in \mathcal{S}^n$ be a data set of i.i.d. observations from $P \in \mathcal{P}$, and $\mathcal{A}$ a collection of candidate learners. 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 $a : \mathcal{S}^n \to \Theta$ which takes a data set as input and returns an estimate $a(\mathcal{D}_n) \in \Theta$. Let $L : \Theta \times \mathcal{S} \to \mathbb{R}_+$ be a loss function, representing the performance of the model $\theta \in \Theta$ at the observation $O \in \mathcal{S}$, 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

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

and depends on both the library of learners and the specific partitioning of the data into cross-validation 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 data 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 recommended for super learning with right-censored data by Polley and van der Laan (2011), under the assumption that data are observed in discrete time. However, the partial log-likelihood loss does not work well as a general purpose measure of performance in hold-out samples when data are observed in continuous time. The reason is that the partial log-likelihood assigns an infinite value to any learner that predicts piecewise constant cumulative hazard functions, if the test set contains event times that are not observed in the training set. For instance, if no competing risks are present, a piecewise constant cumulative hazard function postulates a model for the distribution of the survival times where all probability is assigned to the finite number of time points at which the cumulative hazard function jumps. The likelihood according to such a model is zero at almost all time points, and thus the likelihood of any hold-out sample will almost surely be zero when data are observed in continuous time. 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). One might attempt to resolve this issue by smoothing an estimated cumulative hazard functions to obtain an estimate of the hazard function itself. This is a theoretically unattractive approach, as estimation of a hazard function is much harder than estimation of a cumulative hazard function. In practice, this approach would also introduces the additional problem of tuning a smoothing parameter, which may be infeasible for more complicated estimators like random survival forest, where the smoothing would have to be done conditional on baseline covariates.

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). However, 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 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, 2023, Appendix A.4).

4 | THE JOINT SURVIVAL SUPER LEARNER

The main idea of the joint survival super learner is to specify libraries of learners for the hazard functions Λ_1, Λ_2, and Γ, and to exploit the relations in equation (2) to define a joint loss function. The joint survival super learner thus evaluates a tuple of learners for $(\Lambda_1, \Lambda_2, \Gamma)$ based on how well they jointly predict the observed data and the discrete joint survival super learner chooses the best performing tuple. 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 and was observed, value 2 the state where a competing risk has occurred and was observed, 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 $x \in \mathcal{X}$.

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 F 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. Taking the Cartesian product of these libraries, we obtain a library Φ of learners for F:

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

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

$$\begin{aligned}
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, 0, x) &= \exp\{-a_1(\mathcal{D}_n)(s \mid x) - a_2(\mathcal{D}_n)(s \mid x) - b(\mathcal{D}_n)(s \mid x)\}, \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, 1, x) &= \int_0^t \varphi_{a_1,a_2,b}(\mathcal{D}_n)(s-, 0, x)a_1(\mathcal{D}_n)(\mathrm{d}s \mid x), \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, 2, x) &= \int_0^t \varphi_{a_1,a_2,b}(\mathcal{D}_n)(s-, 0, x)a_2(\mathcal{D}_n)(\mathrm{d}s \mid x), \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t, -1, x) &= \int_0^t \varphi_{a_1,a_2,b}(\mathcal{D}_n)(s-, 0, x)b(\mathcal{D}_n)(\mathrm{d}s \mid x).
\end{aligned} \tag{9}$$

Notably, the libraries $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}$ can be constructed using standard software for survival analysis. For example, in the `R` software we can specify various ways to include covariates in a Cox regression model and fit learners of the hazard functions using the `survival`-package (Therneau, 2022), and we can specify hyper parameters of a random

survival forest and derive learners of the hazard functions 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) evaulated at time τ:

$$\bar{B}_\tau(F, O) = \int_0^\tau \sum_{l=-1}^{2} (F(t, l, X) - \mathbb{1}\{\eta(t) = l\})^2 \, \mathrm{d}t.$$

Here, the integrand is the average Brier score at time t across the four states (Brier et al., 1950). 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 $\Phi(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$ is evaluated using the cross-validated loss,

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

and the discrete joint survival super learner is the best performing tuple of hazard functions:

$$(\hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}, \hat{\Gamma}_n) = \underset{(a_1,a_2,b) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{B}}{\operatorname{argmin}} \hat{R}_n(\varphi_{a_1,a_2,b}; \bar{B}_\tau). \tag{10}$$

We provide a summary of the procedure for obtaining the joint survival super learner in Algorithm 1.

Cause-specific risk predictions can be obtained from the joint survival super learner (10) by substituting into the well-known formula (e.g., Benichou and Gail, 1990; Ozenne et al., 2017),

$$\hat{Q}_n(T \le t, D = j \mid X = x) = \int_0^t \exp\left\{-\hat{\Lambda}_{1n}(u \mid x) - \hat{\Lambda}_{2n}(u \mid x)\right\} \hat{\Lambda}_{jn}(\mathrm{d}u \mid x), \quad j \in \{1, 2\}. \tag{11}$$

Furthermore, the joint survival super learner provides an estimator of the censoring distribution:

$$\hat{G}_n(T \le t \mid X = x) = \exp\left\{-\hat{\Gamma}_n(t \mid x)\right\}.$$

Algorithm 1: The joint survival super learner using the integrated Brier score

Input: Data set $\mathcal{D}_n$, libraries of learners $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}$, number of folds $K \in \mathbb{N}$, and time horizon τ.

Output: Selected tuple of learner $(\hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}, \hat{\Gamma}_n) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{B}$.

Partition the data set $\mathcal{D}_n$ randomly into K subsets $\{\mathcal{D}_n^1, \ldots, \mathcal{D}_n^K\}$ of approximately equal size.

for $k = 1, \ldots, K$ **do**

Fit all learners in all libraries to the k'th training data to obtain the collections of fitted learners:

$$\hat{\mathcal{A}}_1^{-k} = \{a_1(\mathcal{D}_n^{-k}) : a_1 \in \mathcal{A}_1\},$$

$$\hat{\mathcal{A}}_2^{-k} = \{a_2(\mathcal{D}_n^{-k}) : a_2 \in \mathcal{A}_2\},$$

$$\hat{\mathcal{B}}^{-k} = \{b(\mathcal{D}_n^{-k}) : b \in \mathcal{B}\},$$

where $\mathcal{D}_n^{-k} = \mathcal{D}_n \setminus \mathcal{D}_n^k$.

for $(a_1, a_2, b) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{B}$ **do**

Use equation (9) and the fitted learners $a_1(\mathcal{D}_n^{-k}) \in \hat{\mathcal{A}}_1^{-k}$, $a_2(\mathcal{D}_n^{-k}) \in \hat{\mathcal{A}}_2^{-k}$, and $b(\mathcal{D}_n^{-k}) \in \hat{\mathcal{B}}^{-k}$ to obtain the fitted F-learner, $\varphi_{a_1,a_2,b}(\mathcal{D}_n^{-k})$.

Calculate the estimated risk in the k'th hold out data using the integrated Brier score:

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

end

end

for $(a_1, a_2, b) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{B}$ **do**

Calculate the cross-validated risk: $\hat{R}_n(\varphi_{a_1,a_2,b}; \bar{B}_\tau) = \frac{1}{K} \sum_{k=1}^K \hat{R}_n^k(\varphi_{a_1,a_2,b}; \bar{B}_\tau)$.

end

return $(\hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}, \hat{\Gamma}_n) = \operatorname{argmin}_{(a_1,a_2,b) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{B}} \hat{R}_n(\varphi_{a_1,a_2,b}; \bar{B}_\tau)$

5 | THEORETICAL GUARANTEES

Cross-validation is the backbone of super learning and an intuitively reasonable procedure for fair model selection without overfitting. In this section, we adapt the work of van der Laan and Dudoit (2003) and van der Vaart et al. (2006) and provide a theoretical justification for the joint survival super learner 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 proper, in the sense that minimization recovers the parameter of the data-generating distribution. Together with our finite-sample oracle inequality (Proposition 2 below), this implies that the joint survival super learner is consistent when it is based on a library that includes at least one consistent learner. Another consequence of our finite-sample oracle inequality is that the joint survival super learner converges (nearly) at the optimal rate achievable within the library of learners. This statement is made precise in Corollary 3 and the following discussion. Proofs are deferred to the Appendix.

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 called proper, and strictly proper if the minimizer is unique (Gneiting and Raftery, 2007). Absence of properness makes it unclear why minimizing the (estimated) expected loss is interesting. Proposition 1 states that the integrated Brier score as defined in Section 4 is a strictly proper scoring rule. To establish this result, recall that the function F implicitly depends on the data-generating probability measure

$P \in \mathcal{P}$ but that this was so-far 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 $\mathcal{F}_{\mathcal{P}} = \{F_P : P \in \mathcal{P}\}$.

Proposition 1 *If $P \in \mathcal{P}$ then*

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

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

The discrete joint survival super learner defined in (10) provides an estimate of the function F

$$\hat{\varphi}_n = \varphi_{\hat{\lambda}_{1n}, \hat{\lambda}_{2n}, \hat{\Gamma}_n}$$

which is obtained by substituting $(\hat{\lambda}_{1n}, \hat{\lambda}_{2n}, \hat{\Gamma}_n)$ for (a_1, a_2, b) into the structural equations (9). To evaluate the performance of $\hat{\varphi}_n$ we benchmark it against the data-generating distribution F_P, which according to Proposition 1 has the smallest expected loss. Another useful theoretical benchmark is the so-called oracle learner which is the best learner included in the library of learners and formally defined by

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

where we use $\mathbb{E}_P$ to denote the expectation under the distribution P for a new observation O which is independent of the data $\mathcal{D}_n^{-k}$. Like the joint survival super learner, the oracle learner depends on the library of learners and on the actual partition of the data, but unlike the joint survival super learner, it also depends on the unknown data-generating distribution. It is hence not available in practice and serves only as a theoretical benchmark.

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

$$\|F\|_P = \left\{\sum_{l=-1}^{2} \int_0^\tau \mathbb{E}_P\left[F(t, l, X)^2\right] \mathrm{d}t\right\}^{1/2}. \tag{12}$$

This norm induces a natural performance measure because $\|F - F_P\|_P$ is equal to the excess risk, $\mathbb{E}_P\left[\bar{B}_\tau(F, O)\right] - \mathbb{E}_P\left[\bar{B}_\tau(F_P, O)\right]$, as shown in Lemma 4 in the Appendix. For simplicity of presentation, we assume that all folds of the data partition have equal size, $|\mathcal{D}_n^{-k}| = n/K$ for a fixed number of folds K. We allow the number of learners to grow with n and write $\Phi_n = \Phi(\mathcal{A}_{1n}, \mathcal{A}_{2n}, \mathcal{B}_n)$ as short-hand notation emphasizing 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 2 *For all $P \in \mathcal{P}$, $n \in \mathbb{N}$, $k \in \{1, \ldots, K\}$, and $\delta > 0$,*

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

The expectation in Proposition 2 is taken with respect to the product measure P^n for the data set $\mathcal{D}_n$. This means that we are quantifying the average performance of the joint survival super learner across all training data of size n. A corresponding quantity was called the expected true error rate in Efron and Tibshirani (1997). As with 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 3 *Assume that* $|\Phi_n| = O(n^q)$, *for some* $q \in \mathbb{N}$ *and that there exists a sequence* $\varphi_n \in \Phi_n$, $n \in \mathbb{N}$, *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* $\mathbb{E}_P\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] = C_P + O(\log(n)^{1+\varepsilon} n^{-1})$, $\forall \varepsilon > 0$.
(b) *If* $\alpha < 1$, *then* $\mathbb{E}_P\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_P\|_P^2\right] = C_P + O(n^{-\alpha})$.

Proposition 2 provided a finite-sample bound on the average price we pay for using cross-validation, and Corollary 3 states that this price vanishes asymptotically, up to poly-logarithmic terms, provided that the size of the library does not grow faster than at a polynomial rate in the sample size. The situation $C_P = 0$ corresponds to a setting in which the library includes a consistent learner. Cases (a) and (b) correspond to situations where the oracle learner achieves a parametric or non-parametric asymptotic rate of convergence, respectively.

To illustrate the content of Corollary 3, consider first a situation where we use a library with an increasing number of Cox regression models. Each of these models will achieve a parametric rate of convergence, to possibly different least-false cumulative hazard functions Hjort (1992), and hence item (a) of Corollary 3 states that the joint survival super learner based on this library will achieve a near-parametric rate of convergence. C_P can be set to be the distance between the data-generating distribution and 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 slower rates, where the fastest achievable rate depends on the data-generating distribution. Item (b) of Corollary 3 shows that the joint survival super learner achieves the same convergence rate as the best-performing learner in the library, without any knowledge of the data-generating distribution.

The norm defined in equation (12) operates on functions F which are features of the observed data distribution. This means that Proposition 2 and Corollary 3 provide guarantees in terms of a how well the function $\hat{\varphi}_n$ predicts the observed data. Ideally, we would like performance guarantees for, e.g., the selected learners $\hat{\Lambda}_{jn}$ or the derived risk-prediction learner $\hat{Q}_n$ defined in equation (11). There is a one-to-one correspondence between the learner $\hat{\varphi}_n$ and the tuple of learners $(\hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}, \hat{\Gamma}_n)$ through equations (3)-(4) and (7), and we expect that the performance guarantees provided for $\hat{\varphi}_n$ will in many cases translate into similar performance guarantees for each element of the tuple $(\hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}, \hat{\Gamma}_n)$. We do not investigate this further theoretically, but investigate is empirically in our numerical experiments in Section 6.

6 | NUMERICAL EXPERIMENTS

The numerical experiments have two aims. The first aim is to demonstrate that the joint survival super learner can outperform the IPCW based discrete super learners of (Gonzalez Ginestet et al., 2021) which pre-specify a potentially misspecified model for the censoring mechanism. The second aim is to show that the discrete joint survival super

learner can compete and outperform the ensemble super learner proposed by Westling et al. (2021).

For the numerical experiments we have synthesized the prostate cancer data of Kattan et al. (2000) by fitting a hierarchical structural equation model under parametric assumptions. The outcome of interest is the time from randomization until the combined endpoint tumour recurrence or all-cause death. Five baseline covariates are used to predict the outcome risk: 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. 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 (Bender et al., 2005) 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 of the covariates. We refer to this simulation setting as 'independent censoring'.

For all super learners, we use a library consisting of three learners: The Nelson-Aalen estimator (Andersen et al., 2012), a Cox regression model with additive effects of the covariates (Cox, 1972), and a random survival forest (Ishwaran et al., 2008). We use the same library to learn the cumulative hazard functions of the outcome and the censoring time, respectively. Specifically, to obtain estimates of the cumulative censoring hazard function we fit the learners to the modified data set $\{(\tilde{T}_i, 1-\Delta_i, X_i)\}_{i=1}^n$ where the roles of censoring and outcome are exchanged.

We compare the joint survival super learner to two IPCW based super learners: The first super learner, called IPCW(Cox), uses a Cox regression model with additive effects of the five covariates 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 thus 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, we obtain a risk prediction model as described in Section 4, see in particular equation (11) with the special case of $\Lambda_2 = 0$. We measure the performance of the risk prediction model provided by 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. For a risk prediction model $r : \mathcal{X} \to [0,1]$, IPA at time τ is

$$1 - \frac{\mathbb{E}_Q\left[(r(X) - \mathbb{1}\{T \le \tau\})^2\right]}{\mathbb{E}_Q\left[(Q(T \le \tau) - \mathbb{1}\{T \le \tau\})^2\right]}.$$

We chose IPA as a performance measure because it is proper, incorporates both discrimination and calibration, and is easy to interpret as it measures the relative performance gain compared to the null model which does not use any baseline information. The definition of IPA involves the uncensored survival time T, which is not available in practice. However, in the numerical studies, this quantity is available because we know the data-generating mechanism used to generate T. In practice, we Monte Carlo approximate the IPA by generating a large ($n = 20,000$) independent data set

of uncensored survival times, and calculate the empirical version of the IPA in there. As a benchmark, we calculate the performance of the risk prediction model chosen by the oracle selector, which has the highest IPA in the simulation setting.

The results for the first aim 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.

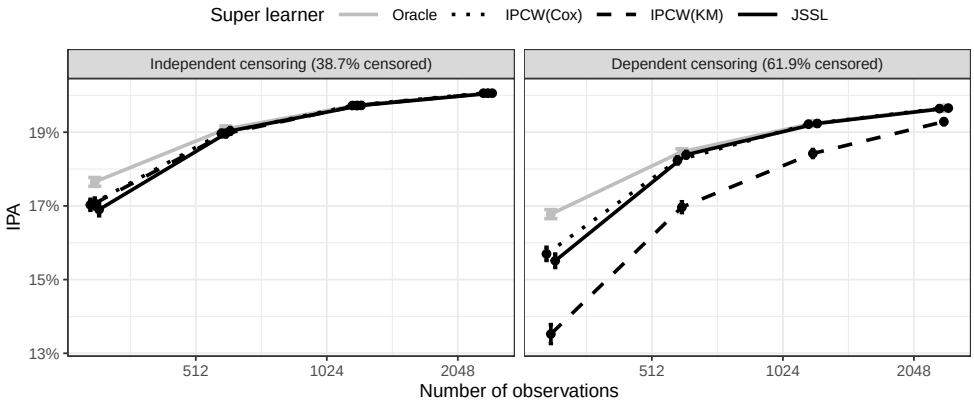

FIGURE 1 For the risk prediction models provided by each of the super learners, the IPA is plotted against sample size. The results are averages across 1000 simulated data sets and the error bars are used to quantify the Monte Carlo uncertainty. JSSL denotes the joint survival super learner.

For the second aim, we consider the super learner *survSL* proposed by (Westling et al., 2021) which like the joint survival super learner does not require a pre-specified censoring model. Both methods provide estimates of the event-time and censoring distributions and hence we compare their performance with respect to both the outcome and the censoring distribution. Again we use the IPA to quantify the predictive performance.

The results for the second aim are shown in Figures 2 and 3. We see that for most sample sizes, the joint survival super learner has similar or higher IPA compared to survSL with respect to both the prediction of the censoring and the outcome risks. We note that the advantage of the joint survival super learner in this particular simulation setting might be due that it is a discrete super learner, whereas survSL combines the learners.

7 | PROSTATE CANCER STUDY

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 consist of 1,042 patients who are followed from start of followup until tumour recurrence ($n = 268$), death without tumour recurrence ($n = 29$), or censored ($n = 745$), whatever came first. We use the joint survival super learner to rank libraries of learners for the cause-specific cumulative hazard functions of tumour recurrence, death without tumour recurrence, and censoring.

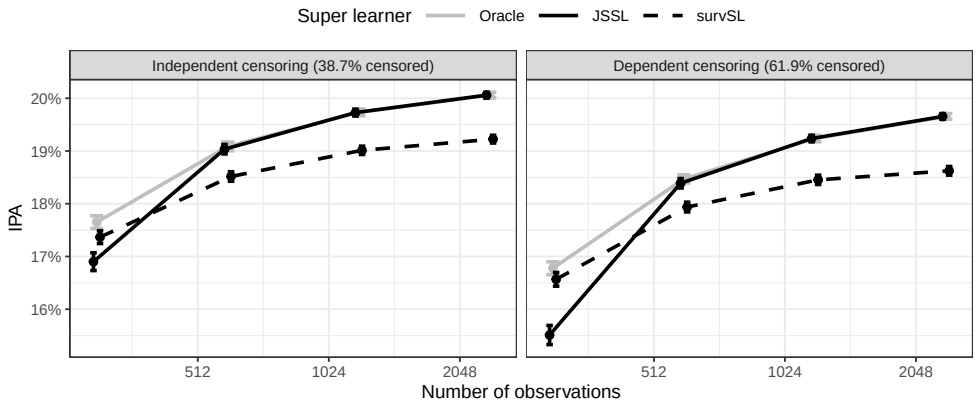

FIGURE 2 For the risk prediction models of the outcome provided by each of the super learners, the IPA at the fixed time horizon is plotted against sample size. The results are averages across 1000 repetitions and the error bars are used to quantify the Monte Carlo uncertainty. JSSL denotes the joint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

The three 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 baseline covariates, and a random survival forest. The Nelson-Aalen estimator is estimated without covariates and serves as a benchmark model which guarantees that the joint survival super learner is not worse than a model which predicts the same probability to all individuals. The other four learners use the five baseline covariates listed in Section 6 to predict the three cumulative hazard functions of time to tumour recurrence Λ_1, time to death without tumour recurrence Λ_2, and time to censoring Γ. The resulting library consists of $5^3 = 125$ learners. We use five folds for training and testing the models, repeat training and evaluation five times with different splits, and obtain the discrete joint survival super learner as the combination with the best average five-fold integrated Brier score across the five repetitions. Table 1 shows integrated Brier scores for the conditional state occupation probability function F as defined in Section 4, evaluated 3 years after randomization for a selection of the 125 learners. We see that among the five best combinations, the random survival forest is always selected for Γ and that the differences for different learners of Λ_1 and Λ_2 are small. To illustrate the comparative performance of the discrete joint survival super learner, we also split the data randomly into a training set with $n = 658$ individuals and a test set with the remaining $n = 384$ individuals. We fit the discrete joint survival super learner (five repetitions of five-fold cross-validation) and for comparison pre-specified cause-specific Cox regression models to the training set. In the training set, the discrete joint survival super learner chooses the unpenalized Cox regression model for tumor recurrence, the elastic net Cox regression for death without tumor recurrence, and the random survival forest for censoring. Figure 4 compares the 3-year risk predictions from the pre-specified Cox model and the discrete joint survival super learner in the test set.

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-

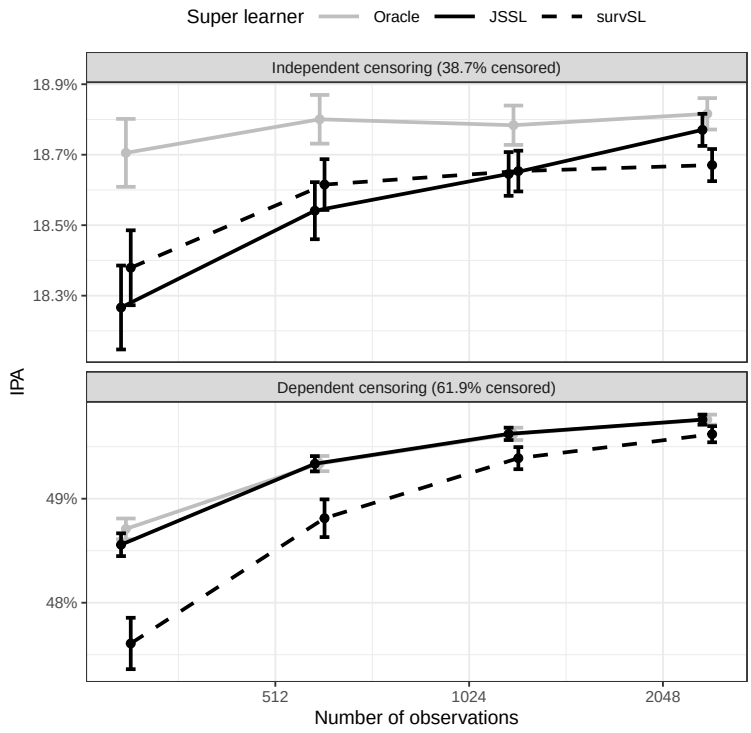

FIGURE 3 For the risk prediction models of the censoring model provided by each of the super learners, the IPA at the fixed time horizon is plotted against sample size. The results are averages across 1000 repetitions and the error bars are used to quantify the Monte Carlo uncertainty. JSSL denotes the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

time distribution, or the censoring distribution, or both. Our numerical experiments suggest that our approach does provide estimates of the conditional survival functions which perform well, also when the predictive performance is measured against the true survival function with no censoring present.

Alternatives to using a performance measure defined with respect the observed data are the use of IPCW loss functions, censoring unbiased transformations, or pseudo-values. As mentioned in Section 3, the drawback of these approaches is that they all need a pre-specified estimator of the censoring distribution, and hence these methods are not immediately applicable if we do not in advance know how to model the censoring distribution. We note that for the special case where the partial log-likelihood loss can be used, this loss function, like our suggested approach, also measures performance with respect to a feature defined by the observed data distribution (e.g., Hjort, 1992; Whitney et al., 2019). We do not know of any method that would allow us to evaluate performance of a risk-prediction model in censored data without either modeling additional nuisance parameters (such as the censoring distribution) or measuring performance directly with respect to the observed data.

A relevant application of the joint survival super learner is within the framework of targeted learning (van der Laan and Rose, 2011), also known as debiased machine learning (Chernozhukov et al., 2018), – a general methodology that combines flexible, data-adaptive estimation of nuisance parameters with asymptotically valid inference for low-

TABLE 1 The results of applying the 125 combinations of learners to the prostate cancer data set. Shown are the best 5 combinations and selected intermediate ranks. The 'Loss' is the integrated Brier score evaluated at 3 years and 'SD' is the standard deviation across the five repetitions of five-fold cross-validation. The discrete joint survival super learner chooses rank 1.

Rank	Cause 1	Cause 2	Censored	Loss	SD
1	elastic net	elastic net	random forest	7.0205	0.030977
2	lasso	elastic net	random forest	7.0209	0.031035
3	Cox	elastic net	random forest	7.0224	0.030871
4	elastic net	lasso	random forest	7.0225	0.030170
5	lasso	lasso	random forest	7.0228	0.030237
25	random forest	random forest	lasso	7.3845	0.026975
50	elastic net	random forest	lasso	7.3974	0.021290
75	lasso	Cox	lasso	7.4059	0.024660
100	Nelson-Aalen	Cox	elastic net	7.8300	0.016719
125	Nelson-Aalen	Nelson-Aalen	Nelson-Aalen	10.3298	0.003289

dimensional target parameters. For example, the methods in van der Laan and Robins (2003) and Rytgaard and van der Laan (2022) target the average treatment effect in a survival setting, which require estimates of the cause-specific cumulative hazard functions and the censoring cumulative hazard function. The joint survival super learner can be used to estimate these nuisance parameters and is more generally well suited for targeted and debiased machine learning with right-censored data.

We have focused on a discrete version of the joint survival super learner, but it is of interest to extend the method to a proper ensemble learner, where learners are combined, e.g., through stacking. There are at least two possible directions for constructing an ensemble version of the joints survival super leaner. One option is to construct a single convex combination of the F-learners $\varphi \in \Phi(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$. Another, perhaps more interesting option, is to construct three separate convex combinations of the learners in $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}$. How such an ensemble should be build and implemented is an interesting topic for future research.

[Generalizing our proposal to ensemble learning is ... At least two strategies could be pursued. First, ensemble for F. More interestingly, jointly build ensembles for all element of the tuple.]

Appendix

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 (12).

Lemma 4 $R_P(F) = \|F - F_P\|_P^2$.

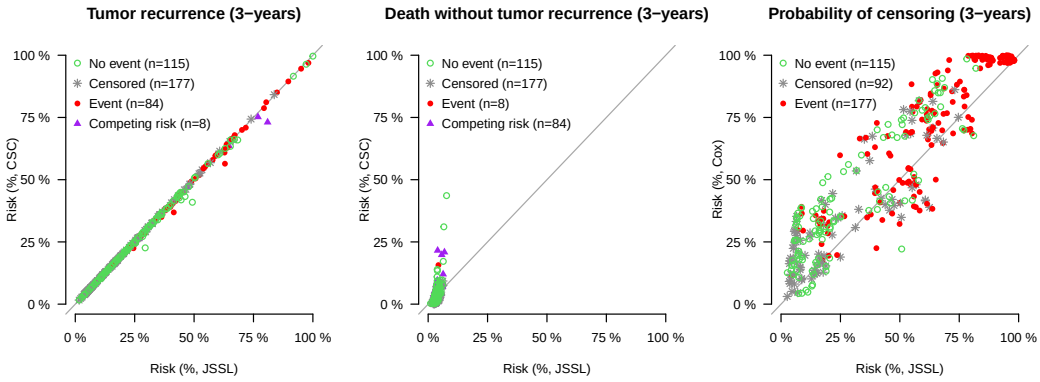

FIGURE 4 Comparison of risk predictions of the discrete joint survival super learner and pre-specified Cox regressions in an independent test data set.

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

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

of Proposition 1 The result follows from Lemma 4.

Recall that we use Φ_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 Φ_n, c.f., Section 4.

of Proposition 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 4, $R_P(F) \geq 0$ for any $F \in \mathcal{F}_\mathcal{P}$, and so using Theorem 2.3 from (van der Vaart et al., 2006) with $p = 1$, we have that for all $\delta > 0$,

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

where for each $F \in \mathcal{F}_{\mathcal{P}}$, $(M(F), v(F))$ 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 $F \in \mathcal{F}_{\mathcal{P}}$, 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

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

Thus when $v(F) = 1.5\,\mathbb{E}_P\,[\bar{B}_{\tau,P}(F, O)^2]$ we have by Lemma 4

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

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

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

For all $\delta > 0$ we thus have

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

which is equivalent to

$$\mathbb{E}_P\left[R_P(\hat{\varphi}_n(\mathcal{D}_n^{-k}))\right] \le (1+2\delta)\,\mathbb{E}_P\left[R_P(\tilde{\varphi}_n(\mathcal{D}_n^{-k}))\right] \quad + (1+\delta)\log(1+|\Phi_n|)\tau\frac{16K}{n}\left(13+\frac{12}{\delta}\right),$$

for any $k \in \{1, \ldots, K\}$, because we have assumed that $|\mathcal{D}_n^{-k}| = n/K$ for all k, and hence the expectations on the right- and left-hand side above do not depend on k. The final result then follows from Lemma 4.

of Corollary 3 By definition of the oracle and Lemma 4,

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

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 2 by letting δ grow to zero with n, for instance as $\delta_n = \log(n)^{-\varepsilon}$ for some $\varepsilon > 0$.

references

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

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

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

Bender, R., Augustin, T. and Blettner, M. (2005) Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, **24**, 1713–1723.

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

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

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

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

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

Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548–560.

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

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

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

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

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

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

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

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

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

Gonzalez Ginestet, P., Kotalik, A., Vock, D. M., Wolfson, J. and Gabriel, E. E. (2021) 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**, 51–65.

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

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

Hjort, N. L. (1992) On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique*, 355–387.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Liu, P., Sawhney, S., Heide-Jørgensen, U., Quinn, R. R., Jensen, S. K., Mclean, A., Christiansen, C. F., Gerds, T. A. and Ravani, P. (2024) 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**.

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

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

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

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

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

R Core Team (2024) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

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

Sachs, M. C., Discacciati, A., Everhov, Å. H., Olén, O. and Gabriel, E. E. (2019) 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**, 1431–1446.

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

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

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

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

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

Whitney, D., Shojaie, A. and Carone, M. (2019) Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **34**, 591.

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

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