ANNUAL REVIEWS

Annual Review of Statistics and Its Application

High-Dimensional Survival Analysis: Methods and Applications

Stephen Salerno and Yi Li

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA; email: yili@umich.edu

Annu. Rev. Stat. Appl. 2023. 10:25–49

First published as a Review in Advance on October 6, 2022

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-032921-022127

Keywords

precision medicine, data science, feature screening, machine learning, artificial neural network, statistical inference

Abstract

In the era of precision medicine, time-to-event outcomes such as time to death or progression are routinely collected, along with high-throughput covariates. These high-dimensional data defy classical survival regression models, which are either infeasible to fit or likely to incur low predictability due to overfitting. To overcome this, recent emphasis has been placed on developing novel approaches for feature selection and survival prognostication. In this article, we review various cutting-edge methods that handle survival outcome data with high-dimensional predictors, highlighting recent innovations in machine learning approaches for survival prediction. We cover the statistical intuitions and principles behind these methods and conclude with extensions to more complex settings, where competing events are observed. We exemplify these methods with applications to the Boston Lung Cancer Survival Cohort study, one of the largest cancer epidemiology cohorts investigating the complex mechanisms of lung cancer.

1. INTRODUCTION

Survival analysis is an area of statistics where the random variate is survival time or the time until the occurrence of a specific event, which represents a qualitative change or the transition from one discrete state to another (e.g., alive to deceased). The most often studied event in biomedicine is death, though events of interest in fields ranging from sociology to industry, engineering, finance, and astronomy are widely encountered as well. The goals of survival analysis are to describe the probability of an event occurring by some time, to detect associations between risk factors and events, or to predict survival times based on informative characteristics. What distinguishes survival outcomes from other outcomes is the presence of censoring, meaning that the event of interest may not be observed for all subjects; subjects whose event times are not observed are said to be censored. In practice, the fraction of event times that are censored in a study population can be substantial, prohibiting the direct use of standard regression methods. Estimation methods in survival analysis are built around extracting information from all subjects, censored or not.

In the era of precision medicine, survival outcomes with high-throughput covariates or predictors are routinely collected. These high-dimensional data (i.e., with the number of predictors exceeding the number of observations) challenge classical survival regression models, which are either infeasible to fit or likely to incur low predictability due to overfitting. Recently, emphasis has been placed on developing novel approaches for feature selection and survival prognostication. We review various methods that handle survival outcome data with high-dimensional predictors, highlighting recently developed machine learning approaches for survival prediction. We also discuss recent developments for deep learning in survival settings and introduce some new deep learning techniques in the presence of competing or semicompeting outcomes. A competing risk is an event whose occurrence precludes the occurrence of another event of interest (Austin & Fine 2017), while in a semicompeting setting, the occurrence of a nonterminal event (e.g., disease progression) is subject to a terminal event (e.g., death), but not vice versa (Haneuse & Lee 2016). We illustrate a novel deep learning approach for prediction under semicompeting outcomes and exemplify the method using data from the Boston Lung Cancer Survival Cohort (BLCSC), a large hospital-based cancer epidemiology cohort investigating the molecular mechanisms and clinical pathophysiology of lung cancer (Christiani 2017).

This review is outlined as follows. In Section 2, we provide a brief overview of some key concepts and notation in survival analysis and introduce the necessary prerequisites on which much of the subsequent literature is built. In Section 3, we survey current techniques for fitting survival models with high-dimensional covariates, primarily focusing on methods that perform feature selection under sparsity assumptions. We briefly discuss ultra high-dimensional settings and introduce screening methods, and end this section with a discussion of methods for drawing valid inference with high-dimensional covariates. In Section 4, we turn to machine learning for survival prediction. We first discuss the application of common machine learning concepts in these settings, such as support vector machines (SVMs), recursive partitioning and survival trees, and ensemble learners such as random survival forests. We briefly review artificial neural networks and extend this notion to survival prediction. In Section 5, we review existing deep learning procedures for competing risk analysis, illustrate a new deep learning approach for predicting semicompeting outcomes, and work through the BLCSC study. We conclude with remarks on future work and open areas. **Supplemental Appendixes A and B** tabulate the reviewed methods and their available software, and present additional simulation results.

Supplemental Material >

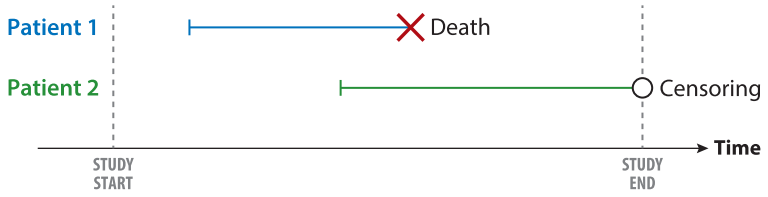

Figure 1

Schematic of observations for two example patients, with different entry times, over the course of a study. The event of interest, death, is observed for patient 1, whereas patient 2 is censored, as the patient is still alive at the end of the study.

2. NOTATION

Consider a study consisting of n subjects. The outcome variable is the time to the event of interest, such as death or cancer progression. Events in other contexts can be bankruptcy, COVID-19 infection, graduation, missing a mortgage payment, and so on. A time zero also needs to be set carefully to have proper biological or practical interpretations when helping to address specific scientific questions. For instance, some common choices of time zero in medical studies include date of birth, time of diagnosis, date of randomization in a clinical trial, or first date receiving a treatment. A unique aspect of survival analysis is that the event may go unobserved for some individuals. In particular, right censoring occurs when a subject's follow-up time ends before the event can be observed (**Figure 1**). Though other types of censoring exist, we focus on right censoring, which happens most often in practice.

We denote the ith subject's survival and censoring times by T_i and C_i, respectively ($i = 1, \ldots, n$), which are nonnegative random variates. For the ith subject, we observe $\boldsymbol{X}_i$, a p-vector of covariates; $Y_i = \min(T_i, C_i)$; and the event indicator $\delta_i = \mathbb{I}(T_i \leq C_i)$, where $\mathbb{I}(\cdot)$ is an indicator function. We assume that subjects are independent from each other and that $T_i \perp C_i$, given $\boldsymbol{X}_i$. Often, the goal of survival analysis is to associate $\boldsymbol{X}_i$ with the distribution of T_i and, in particular, model the conditional hazard function given $\boldsymbol{X}_i$, i.e.,

$$\lambda(t|\boldsymbol{X}_i) = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T_i < t + \Delta | T_i \geq t, \boldsymbol{X}_i), \qquad 1.$$

which measures the instantaneous failure rate at a given time among those who are alive and with $\boldsymbol{X}_i$. Throughout this review, for simplicity, we assume that $\boldsymbol{X}_i$ is time invariant, though in many circumstances extensions to time-dependent $\boldsymbol{X}_i$ are possible.

3. HIGH-DIMENSIONAL SURVIVAL MODELS

In high-dimensional settings, it is not recommended to build prediction models with all of the available features due to the risk of overfitting. A useful strategy is to select only the most vital features under the assumption of sparsity, meaning that most of the potential predictors are unimportant, with nearly no effect on the outcome (Friedman et al. 2010). A key question is how to perform variable selection and estimation simultaneously, and the most widely used approaches fall under the class of regularized regression models. Regularization refers to the addition of a penalty term to the objective function, which shrinks the coefficient estimates toward zero and possibly forces some of them to be exactly zero. This mitigates overfitting and results in parsimonious prediction models (Tibshirani 1996).

3.1. Regularized Cox Models

The approach that dominates survival analysis in the biomedical literature is the Cox (1972) proportional hazards model, famed for presenting both a novel hazard model and a novel concept of partial likelihood. The model links the conditional hazard function (Equation 1) to $\boldsymbol{X}_i$ via

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0(t)\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}),$$

where the baseline hazard, $\lambda_0(t)$, is unspecified, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the coefficient vector of $\boldsymbol{X}_i$ to be estimated, with a fixed $p < n$, by maximizing the partial likelihood,

$$PL(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} PL_i(\boldsymbol{\beta}),$$

with $PL_i(\boldsymbol{\beta})$ being the contribution for subject i who is observed to die:

$$PL_i(\boldsymbol{\beta}) = \Pr(\text{subject } i \text{ dies at } Y_i \mid \text{someone from } \mathcal{R}(Y_i) \text{ dies at } Y_i) = \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\sum_{j\in\mathcal{R}(Y_i)}\exp(\boldsymbol{X}_j^T\boldsymbol{\beta})}$$

where $\mathcal{R}(Y_i) = \{j : Y_j \geq Y_i\}$. In high-dimensional settings—that is, $p > n$ and asymptotically p and n may both go to infinity (Zhao & Yu 2006)—directly optimizing the partial likelihood is not feasible because of overparameterization. Instead, regularized regression adds a penalty term to the negative log partial likelihood, $\ell(\boldsymbol{\beta})$, and optimizes a penalized version of the objective function:

$$-\ell(\boldsymbol{\beta}) + \eta Pen(\boldsymbol{\beta}),$$

where the penalty $Pen(\boldsymbol{\beta})$ is controlled by a positive tuning parameter, η, to be selected through cross-validation. A widely recognized family of penalties is based on the l_q-norm,

$$||\boldsymbol{\beta}||_q = \left(\sum_{j=1}^{p}|\beta_j|^q\right)^{1/q}, \; q \geq 0.$$

Regularization approaches with $Pen(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2^2$, known as ridge regression (Hoerl & Kennard 1970), were applied to the Cox model by Verweij & Van Houwelingen (1994) and returned unique and shrunk estimates. However, ridge regression does not promote sparsity, as it cannot shrink individual coefficients to zero. The least absolute shrinkage and selection operator (lasso) (Tibshirani 1996), with $Pen(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$, penalizes the absolute sum of the coefficient estimates and has been routinely used for producing sparse models. Its application to survival settings (Tibshirani 1997), namely, Cox lasso, has become a widely used approach for high-dimensional survival analysis by performing feature selection and estimation simultaneously (**Figure 2**).

Lasso has several notable statistical properties. It possesses model selection consistency under certain regularity conditions, in particular, the strong irrepresentable condition when p grows much faster than n (i.e., that the absolute sum of coefficients for the regression of any noise variable on signal variables must be strictly smaller than 1) (Zhao & Yu 2006) and has a Bayesian interpretation by viewing $\boldsymbol{\beta}$ as having a double exponential prior (Tibshirani 2009). However, as the lasso penalty term is linear in the size of the coefficients, it leads to biased estimates, especially for the coefficients with large absolute values. To remedy this, Zhang & Lu (2007) proposed the adaptive Cox lasso by utilizing $Pen(\boldsymbol{\beta}) = \sum_j w_j|\beta_j|$, with smaller weights, w_j, assigned to larger coefficients and vice versa. The estimates are $\sqrt{n}$ consistent if $\sqrt{n}\eta = O(1)$ and have oracle properties if $\sqrt{n}\eta \to 0$ and $n\eta \to \infty$. When $p > n$, the author suggested using robust estimates such as ridge regression estimates to determine the w_js.

Fan & Li (2002) proposed a smoothly clipped absolute deviation (SCAD) penalty, which is a quadratic spline function of $|\beta|$ with knots at η and $\alpha\eta$. Its derivative with respect to $|\beta|$, i.e.,

$$\eta\left\{\mathbb{I}(|\beta| \leq \eta) + \frac{(\alpha\eta - |\beta|)_+}{(\alpha-1)\eta}\mathbb{I}(|\beta| > \eta)\right\}; \quad \alpha > 2, \eta > 0,$$

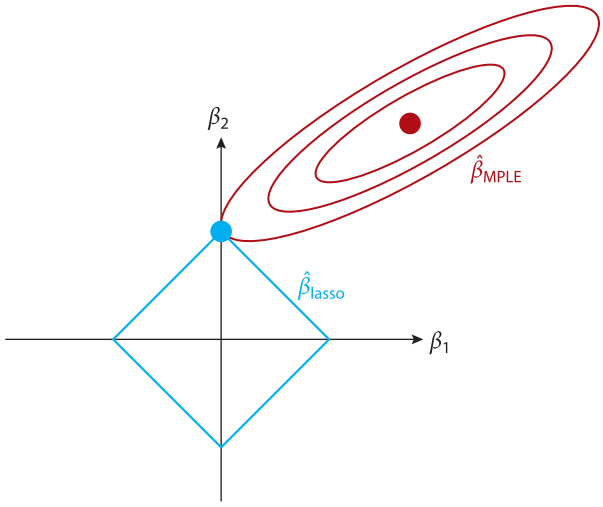

Figure 2

Graphical representation of Cox lasso with two-dimensional predictors. The blue diamond represents the constraint region $|\beta_1| + |\beta_2| \leq s$ for a given s. $\hat{\beta}_{\text{MPLE}}$ and $\hat{\beta}_{\text{lasso}}$ represent the maximum partial likelihood and Cox lasso estimates, respectively, and the red ellipses are contours of the partial likelihood function. As shown, subject to the l_1 constraint, $\hat{\beta}_{\text{lasso}}$ is shrunk to zero compared with $\hat{\beta}_{\text{MPLE}}$, and Cox lasso estimates β_1 to be exactly zero.

may more clearly show the role of the penalty in regularizing estimating equations (Fan & Li 2002). While the SCAD penalty retains the penalization rate of lasso for small coefficients, it relaxes the rate of penalization smoothly as the absolute value of the coefficient increases. Asymptotically, the SCAD penalty yields $\sqrt{n}$-consistent estimates (with a proper rate of η) and possesses oracle properties (if $\sqrt{n}\eta \to 0$ and $n\eta \to \infty$). Strong oracle properties for lasso and SCAD were established by Bradic et al. (2011), who further proposed a class of nonconvex penalization procedures for the Cox model. Nonconvex regularization, including SCAD, is appealing as it obtains support recovery properties under much weaker assumptions than for l_1 penalization (Loh & Wainwright 2017).

Another extension is the elastic net penalty for Cox models (Wu 2012), which combines the lasso and ridge penalties but, unlike lasso, is capable of selecting more predictors than the sample size (Zou & Hastie 2005). This notion was generalized by Vinzamuri & Reddy (2013) with the kernel elastic net Cox regression model, which replaced the ridge penalty with $\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}$. Here, $\boldsymbol{\Sigma}$ is a $p \times p$ radial basis function kernel matrix of predictors, which measures pairwise similarity between predictors. This penalty is meant to encourage correlated predictors to have similar strengths on survival prediction. Other regularized Cox methods include the group Cox lasso (Kim et al. 2012), which selects groups of related covariates as a whole, and the fused lasso (Chaturvedi et al. 2014), which penalizes both the coefficient estimates and their successive differences for ordered features (**Table 1**).

3.2. The Dantzig Selector for Survival Data

Candès & Tao (2007) proposed another type of regularized estimator known as the Dantzig selector for linear regression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad 2.$$

Table 1 Examples of regularized Cox regression methods and their penalty terms

Method	Penalty	Constraints
Ridge	$\|\|\boldsymbol{\beta}\|\|_2^2$	NA
Lasso	$\|\|\boldsymbol{\beta}\|\|_1$	NA
Elastic net	$\alpha\|\|\boldsymbol{\beta}\|\|_1 + (1-\alpha)\|\|\boldsymbol{\beta}\|\|_2^2$	$0 < \alpha < 1$
Adaptive lasso	$\sum_j w_j\|\beta_j\|$	$w_j \geq 0$
SCAD ($Pen(\|\beta\|)$)	$Pen'(\|\beta\|) = \eta\left\{\mathbb{I}(\|\beta\| \leq \eta) + \frac{(\alpha\eta - \|\beta\|)_+}{(\alpha-1)\eta}\mathbb{I}(\|\beta\| > \eta)\right\}$	$\alpha > 2, \eta > 0$
Group lasso	$\sum_g \|\|\boldsymbol{\beta}_g\|\|_1$	$\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gj_g})^T$
Fused lasso	$\sum_j\|\beta_j\|$ and $\sum_j\|\beta_j - \beta_{j-1}\|$	NA

Abbreviations: Lasso, least absolute shrinkage and selection operator; NA, not applicable; SCAD, smoothly clipped absolute deviation.

where $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ are an $n \times 1$ vector of responses, an $n \times p$ covariate matrix, a $p \times 1$ vector of coefficients and an $n \times 1$ vector of zero-mean residual errors, respectively. It estimates $\boldsymbol{\beta}$ by solving

$$\min ||\boldsymbol{\beta}||_1$$

$$\text{subject to } ||\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})||_\infty \leq \eta_Q,$$

where $\eta_Q > 0$ is a tuning parameter. Empowered by linear programming, the Dantzig selector offers a useful alternative as a regularized estimating equation approach. As a dual problem of lasso, it often produces the same solution path (Candès & Tao 2007).

On the other hand, accelerated failure time (AFT) models have become a useful alternative to Cox models due to their ease of interpretation (Saikia & Barman 2017). An AFT model links the (log transformed) survival time to covariates via a linear model

$$\log(T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + e_i, \qquad 3.$$

where the log transformation ensures the parameter space of $\boldsymbol{\beta}$ is unconstrained, and the distribution of the errors, e_i, induces a distribution for T_i (**Table 2**). For parametric AFT models, maximum likelihood estimation can be used for inference. When e_i's distribution is unspecified, the models are semiparametric and the maximum likelihood estimates are difficult to obtain, as the likelihood involves infinite-dimensional parameters. With a fixed $p < n$, Buckley & James (1979) proposed an estimating equation approach by imputing the censored outcomes and solving a least squares problem.

For AFT models with high-dimensional predictors, one cannot directly apply lasso estimation, as the objective function again involves infinite-dimensional parameters. Motivated by the work of Candès & Tao (2007) for regularized least squares estimation, Li et al. (2014) used Buckley-James

Table 2 Specifications of various parametric accelerated failure time models

Distribution of e_i	Induced distribution of T_i
Normal distribution	Log-normal distribution
Extreme value distribution	Weibull distribution
Logistic distribution	Log-logistic distribution

e_i is the residual term, and T_i is the survival time.

imputation to express AFT estimation as a least squares problem and then applied the Dantzig selector:

$$\min ||\boldsymbol{\beta}||_1$$

$$\text{subject to } ||\boldsymbol{X}^T\mathbf{P}_n\{\boldsymbol{T}^*(\boldsymbol{\beta}) - \boldsymbol{X}\boldsymbol{\beta}\}||_\infty \leq \eta_Q,$$

where $T_i^*(\boldsymbol{\beta}) = \log(Y_i) + (1-\delta_i)\frac{\int_{e_i(\boldsymbol{\beta})}^{\infty}\hat{S}(s,\boldsymbol{\beta})\mathrm{d}s}{\hat{S}(e_i(\boldsymbol{\beta}),\boldsymbol{\beta})}$ are imputed outcomes with the Kaplan-Meier estimate, $\hat{S}(\cdot,\boldsymbol{\beta})$, based on $\{e_i(\boldsymbol{\beta}) = \log(Y_i) - \boldsymbol{\beta}^T\boldsymbol{X}_i, \delta_i\}, i = 1,\ldots,n$, and $\boldsymbol{T}^*(\boldsymbol{\beta}) = (T_1^*(\boldsymbol{\beta}),\ldots,T_n^*(\boldsymbol{\beta}))^T$. Here, the projection matrix $\mathbf{P}_n = \mathbf{I}_n - 11^T/n$, where $\mathbf{I}_n$ and 1 are an $n \times n$ identity matrix and an $n \times 1$ vector of 1s, respectively, is for centering covariates to avoid estimation of the intercept or the expectation of e_i in Equation 3. An iterative approach is necessary because this is not a linear programming problem, and, like the Dantzig selector for linear models, estimates may be biased and may not possess the oracle property.

To address this, Li et al. (2014) considered an adaptive version of the Dantzig selector with data-driven weights that vary inversely with the magnitude of coefficients. They showed that the weighted Dantzig selector has model selection consistency and oracle properties. On the other hand, a Dantzig selector for the Cox model was proposed by Antoniadis et al. (2010) based on partial likelihood score equations.

Note that in ultra high-dimensional settings where $p \gg n$, penalized variable selection methods such as those described may incur high computational costs, numeric instability, and poor reproducibility (Fan & Lv 2008). As such, variable screening is a crucial first step in identifying predictive biomarkers and reducing the dimensionality of the feature space before applying regularized methods. Feature screening methods such as sure independence screening (Fan & Lv 2008) fit marginal regression models for each covariate one at a time, choose a threshold, and retain those covariates with magnitudes of marginal effects above the threshold. In the ultra high-dimensional survival settings, additional censoring issues need to be addressed. Recent advancements in survival feature screening have included sure independence screening (Fan et al. 2010), principled sure independence screening (Zhao & Li 2012), score test screening (Zhao & Li 2014), concordance measure–based screening (Ma et al. 2017), Buckley-James assisted sure screening (Liu et al. 2020), conditional screening (Kang et al. 2017, Hong et al. 2018b), integrated power density screening (Hong et al. 2018a), L_q-norm screening (Hong et al. 2020), and forward regression (Hong et al. 2019, Pijyan et al. 2020). A focused review of survival feature screening is provided by Hong & Li (2017).

3.3. Inference with High-Dimensional Covariates

As simultaneous estimation and inference is challenging within the high-dimensional survival framework, we review the limited methods available for drawing inference in this area. More broadly, high-dimensional regression inference methods largely fall under postselection inference and debiased lasso estimation. Further challenges arise in that postselection inference is conditional on the selected subset and does not account for variation in model selection. Several authors used debiased lasso (Javanmard & Montanari 2014, Van de Geer et al. 2014, Yu et al. 2018) to draw inference; however, these methods require estimation of the inverse of a $p \times p$ information matrix, which is a daunting task, especially when $p > n$ (Xia et al. 2021, 2022).

3.3.1. Selection-assisted partial regression and smoothing. To address this challenge, Fei et al. (2019) proposed selection-assisted partial regression and smoothing (SPARES) to draw inference for high-dimensional linear models (Equation 2) with $p > n$. Under this framework,

model selection and partial regression are conducted separately on partitioned data, and multiple sample splittings or bootstraps are used to account for variations in variable selection and estimation. Specifically, given data $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$ and a variable selection procedure $\mathcal{S}_\eta$, data are split into equally sized $\mathcal{D}_1$ and $\mathcal{D}_2$. Denote the variables selected by $\mathcal{S}_\eta$ on $\mathcal{D}_2$ as $S = \mathcal{S}_\eta(\mathcal{D}_2)$. On $\mathcal{D}_1 = (\boldsymbol{X}^1, \boldsymbol{Y}^1)$ and for any $j \in \{1, \ldots, p\}$, $\boldsymbol{Y}^1$ is regressed on $\boldsymbol{X}^1_{S\cup j}$ to estimate β^0_j by

$$\tilde{\beta}_j = \left\{ (\boldsymbol{X}^1_{S\cup j}{}^T \boldsymbol{X}^1_{S\cup j})^{-1} \boldsymbol{X}^1_{S\cup j}{}^T \boldsymbol{Y}^1 \right\}_j, \tag{4.}$$

where $\{\cdot\}_j$ denotes the estimate corresponding to variable j. Equation 4 is termed the partial regression estimator (Fei et al. 2019). Set $\hat{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$. The rationale behind this idea is that if $\Pr(S \supset S_0) \to 1$, where S_0 is the true active set, the one-time partial regression in Equation 4 returns a consistent estimate for β^0_j, regardless of whether $j \in S$. However, the one-time estimator is highly variable, depends heavily on S and the specific data split, and does not account for variation in the variables selected. To address this, repeated applications of data splitting and partial regression are carried out over B random splits. Denote by $\hat{\boldsymbol{\beta}}^b$ the estimate of $\boldsymbol{\beta}$ based on the bth resample ($b = 1, \ldots, B$). The SPARES estimator is

$$\hat{\boldsymbol{\beta}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{\beta}}^b.$$

To draw inference, a nonparametric delta method (Van der Vaart 2000, Efron 2014) is used to estimate the standard error of $\hat{\beta}_j$ ($j = 1, \ldots, p$) as $\hat{\text{se}}^B_j = [\sum_{i=1}^n \hat{\text{cov}}^2_{ij}]^{1/2}$, where $\hat{\text{cov}}_{ij}$ is the sample covariance between I_{bi} and $\hat{\beta}^b_j$, with I_{bi} indicating whether subject i is included in the bth resample (used for partial regression). Approximate 95% confidence intervals are given by $\hat{\beta}_j \pm 1.96\, \hat{\text{se}}^B_j$, while a two-sided p-value testing $H_0 : \beta_j = 0$ is given by $2 \times [1 - \Phi(|\hat{\beta}_j| / \hat{\text{se}}^B_j)]$, where $\Phi(\cdot)$ is the distribution function of $N(0, 1)$. SPARES provides a novel inference technique that converts a high-dimensional problem to a low-dimensional regression. It is valid with general selection methods, including lasso, SCAD, screening, and boosting, as long as they possess selection consistency or the relaxed sure screening condition of Fei & Li (2021). Furthermore, this approach is not sensitive to the tuning parameter η in $\mathcal{S}_\eta$ and can be extended to analyze censored outcomes, as detailed below.

3.3.2. High-dimensional censored quantile regression. As opposed to the Cox and AFT models, censored quantile regression permits the effects of covariates to vary across quantile levels, thus accommodating potentially heterogeneous impacts of certain risk predictors. For $\tau \in (0, 1)$, the τth quantile is a value at or below which a τ-fraction of population lies. Denote the τth conditional quantile of $\tilde{T} = \log(T)$ given $\boldsymbol{X}_i$ by $Q_{\tilde{T}}(\tau | \boldsymbol{X}_i)$. The censored quantile regression model (Powell 1986, Portnoy 2003) stipulates that

$$Q_{\tilde{T}}(\tau | \boldsymbol{X}_i) = \beta_0(\tau) + \boldsymbol{X}_i^T \boldsymbol{\beta}(\tau), \tag{5.}$$

where $\boldsymbol{\beta}(\tau)$ is a vector of quantile-specific regression coefficients. With a fixed $p < n$, Peng & Huang (2008) proposed a class of martingale-based estimating equations to estimate $\boldsymbol{\beta}(\tau)$ in Equation 5 over a fine grid of quantile levels—that is, $\Gamma_m = \{\nu = \tau_0, \tau_1, \ldots, \tau_m = \tau_U\}$, where $0 < \tau_U < 1$ is an upper bound for estimability. Specifically, $\boldsymbol{\beta}(\tau_k)$ values, where $\tau_k \in \Gamma_m$, can be sequentially and consistently estimated by solving

$$\sum_{i=1}^{n} \boldsymbol{X}_i \left(N_i \left(\boldsymbol{X}_i^T \boldsymbol{\beta}(\tau_k) \right) - \sum_{r=0}^{k-1} \int_{\tau_r}^{\tau_{r+1}} \mathbb{I}[\log Y_i \geq \boldsymbol{X}_i^T \check{\boldsymbol{\beta}}(\tau_r)] \mathrm{d}H(u) \right) = 0, \tag{6.}$$

where $N_i(t) = \mathbb{I}(\log Y_i \leq t, \delta_i = 1)$ and $H(u) = -\log(1 - u)$ (Peng & Huang 2008).

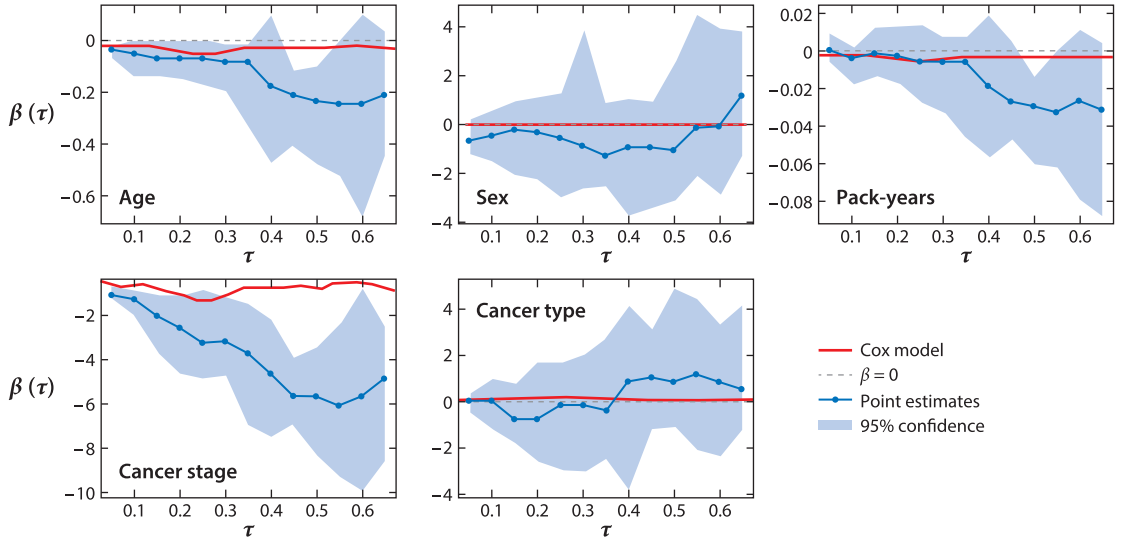

Figure 3

Example censored quantile regression analysis on $n = 153$ patients from the BLCSC study. Estimated local quantile measures (Portnoy 2003) from a Cox model are shown in red, and the reference $\beta = 0$ is given in black. Also shown are the point estimates (*blue curves*) of the quantile-specific regression coefficients and their 95% confidence intervals (*lighter blue shaded regions*). Figure adapted with permission from Hong et al. (2019, figure 2). Abbreviation: BLCSC, Boston Lung Cancer Survival Cohort.

As a concrete example, with a subset of 153 patients from the BLCSC study, Hong et al. (2019) fit a censored quantile regression model that linked the conditional quantile of overall survival to age (years), sex (0: female, 1: male), pack-years, cancer type (0: adenocarcinoma, 1: nonadenocarcinoma), and cancer stage (0: stage 1, 1: stage 2 or above). **Figure 3** displays the point estimates of the quantile-specific regression coefficients and their 95% confidence intervals.

While methods have been proposed to deal with variable selection for high-dimensional censored quantile regression (HDCQR), including penalized quantile regression (Wang et al. 2013a), adaptive penalized quantile regression (Zheng et al. 2013), model-free variable screening (He et al. 2013), and stochastic integral-based estimating equations (Zheng et al. 2018), none could draw statistical inference with HDCQR. Belloni et al. (2019) provided postselection inference in high-dimensional quantile regression at some fixed quantile levels; however, the method cannot handle censored outcomes.

To address this issue, Fei et al. (2021) proposed a fused HDCQR method, which utilizes a variable selection procedure for HDCQR (Zheng et al. 2018) to reduce the dimension of the data and applied partial regression to estimate the effect of each predictor, regardless of whether it is selected. Estimates are aggregated based on multiple data splits and selections. Specifically, when $p > n$, fused HDCQR adapts the SPARES procedure and fits low-dimensional censored quantile regressions using Equation 6. With B random sample splittings, these estimates, denoted by $\tilde{\beta}_j^b, b = 1, \ldots, B$, are aggregated to form the fused HDCQR estimator,

$$\hat{\beta}_j(\tau_k) = \frac{1}{B}\sum_{b=1}^{B} \tilde{\beta}_j^b(\tau_k),\ \tau_k \in \Gamma_m,$$
$$\hat{\beta}_j(\tau) = \hat{\beta}_j(\tau_k),\ \tau_{k-1} \le \tau < \tau_k,\ k = 1, \ldots, m,$$

and a functional delta method (Van der Vaart 2000) can be applied to estimate the variance of $\hat{\beta}_j(\tau_k)$. The fused HDCQR procedure involves repeated fitting of low-dimensional regressions,

which is computationally feasible and can estimate and conduct hypothesis testing for the heterogeneous effects of various risk factors.

The use of fused HDCQR is illustrated with the BLCSC data by studying the differential impacts of genetic variants on different quantiles of survival times. For example, Fei et al. (2021) focused on 2,002 candidate single nucleotide polymorphisms (SNPs) residing in 14 well-known lung cancer related genes and investigated how each SNP played a different role among high-risk (i.e., lower quantiles of overall survival) versus low-risk (i.e., higher quantiles of overall survival) cancer survivors. With the fused HDCQR approach, the estimated coefficient of active smoking ranged from -0.42 ($p = 0.0011$) to -0.53 ($p = 0.0005$) as τ changed from 0.2 to 0.5, and then it increased to -0.31 ($p = 0.038$) as τ changed to 0.7, suggesting that active smoking might be more harmful to the high- or median-risk groups than the low-risk group of patients. The results resonated with the strong need to develop effective smoking cession programs among high-risk populations (Barbeau et al. 2006). Furthermore, SNP AX.37793583_T remained significant throughout $\tau = 0.2$ to $\tau = 0.7$, but its estimated coefficient decreased from 2.75 ($\tau = 0.2$) to 1.39 ($\tau = 0.7$), indicating its heterogeneous impacts on survival—i.e., stronger protective effect at lower quantiles and vice versa, which could not be detected using traditional Cox models (Fei et al. 2021).

4. MACHINE LEARNING TECHNIQUES FOR SURVIVAL PREDICTION

Significant work has gone into the development of machine learning algorithms that can accommodate survival data. These nonparametric learning approaches can handle nonlinear relationships or higher-order interaction that would otherwise be costly using classical methods, and they can improve accuracy in prediction for survival outcomes.

4.1. Support Vector Machines

SVMs fall under the supervised learning family (Vapnik et al. 1995, Noble 2006) and seek to find a hyperplane that provides maximal separation between groups (**Figure 4**). Specifically, consider a binary outcome $Y_i \in \{-1, 1\}$ for each individual i, with a corresponding p-dimensional covariate vector, $\boldsymbol{X}_i$. The goal of SVMs is to identify a hyperplane, $H(\psi, a) = \{\mathbf{v} \in \mathbb{R}^p | \langle \psi, \mathbf{v} \rangle + a = 0\}$, separating these two groups so that the margin, $2/||\psi||$, can be maximized, where $\psi \in \mathbb{R}^p$ is the slope vector and $\langle \cdot, \cdot \rangle$ denotes the inner product (**Figure 4**). Often, the two classes may not be separable in the original feature space within $\mathbb{R}^p$, and we use $\mathbf{F}(\cdot)$ to map the original predictors to a higher-dimensional space where the outcomes can be distinguished, in which case, the hyperplane to deal with is $H(\psi, a) = \{\mathbf{v} \in \mathbb{R}^p | \langle \psi, \mathbf{F}(\mathbf{v}) \rangle + a = 0\}$ and, with slight overuse of notation, the dimension of ψ is the same as that of $\mathbf{F}(\mathbf{v})$. In practice, $\mathbf{F}(\cdot)$ does not have to be obtained explicitly and $\langle \psi, \mathbf{F}(\mathbf{v}) \rangle$ can be calculated by using a reproducing kernel (Wahba et al. 1999). We further introduce a slack variable, $\xi_i = [1 - Y_i\{\langle \psi, \mathbf{F}(\boldsymbol{X}_i) \rangle + a\}]_+$, to dictate the degree to which the ith data point is misclassified, as illustrated in **Figure 4**.

SVMs have been extended to model continuous time-to-event data, which are prone to censoring, by predicting the survival time to be $\langle \psi, \mathbf{F}(\boldsymbol{X}_i) \rangle + a$. Van Belle et al. (2007) formulated the survival SVM based on the rank concordance between the prediction and observed survival time, Y_i, among comparable individuals in the presence of censoring. Specifically, they introduced a comparability indicator, $v_{ij} = \delta_i \mathbb{I}(Y_i < Y_j)$, such that the ordering of the observed survival times for subjects i, j can only be determined when $v_{ij} = 1$. For a comparable pair with $v_{ij} = 1$, a concordance in rank is reached if and only if $\langle \psi, \mathbf{F}(\boldsymbol{X}_j) \rangle - \langle \psi, \mathbf{F}(\boldsymbol{X}_i) \rangle > 0$. Allowing varying degrees of pairwise slacks, i.e., when $\langle \psi, \mathbf{F}(\boldsymbol{X}_j) \rangle - \langle \psi, \mathbf{F}(\boldsymbol{X}_i) \rangle \leq 0$ with $v_{ij} = 1$, across comparable pairs,

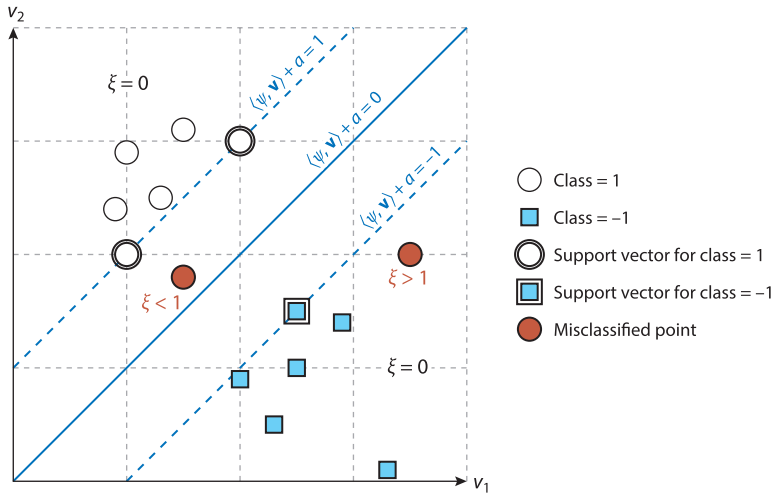

Figure 4

A support vector machine to distinguish binary outcomes with two-dimensional covariates $\mathbf{v} = (v_1, v_2)^T$ by a linear separating line. The solid line represents the optimal hyperplane separating the data, while the dotted lines denote the maximal margin defined by the support vectors (*open double circles*) for one group (*white circles*) versus the other (*cyan squares*). Misclassified points are labeled in red, with corresponding magnitudes for slack variables, ξ.

Van Belle et al. (2007) proposed to solve

$$\min_{\psi,\xi} \frac{1}{2}\|\psi\|^2 + \gamma \sum_{(i,j):Y_i<Y_j} v_{ij}\xi_{ij}$$

$$\text{subject to } \langle \psi, \mathbf{F}(\boldsymbol{X}_j)\rangle - \langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle \geq -\xi_{ij}$$

$$\text{and } \xi_{ij} \geq 0, i, j = 1, \ldots, n,$$

where ξ_{ij}s are pair-specific slacks, whose summation is to be minimized, and $\gamma > 0$ is a regularization parameter controlling the maximal margin and misclassification penalties. This formulation can be shown to maximize the Harrell rank-based concordance index (Harrell et al. 1982). Hence, it is termed the rank-based SVM approach for survival data and does not estimate the intercept a. An alternative regression approach (Shivaswamy et al. 2007) aimed to find a prediction, $\langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle + a$, for continuous survival times by identifying a hyperplane that best fit the data that are subject to censoring (Smola & Schölkopf 2004):

$$\min_{\psi,a,\xi,\xi^*} \frac{1}{2}\|\psi\|^2 + \gamma \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right)$$

$$\text{subject to } Y_i - \langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle - a \leq \xi_i,$$

$$\delta_i \left(\langle \psi, \mathbf{F}(\boldsymbol{X}_i)\rangle + a - Y_i\right) \leq \xi_i^*,$$

$$\text{and } \xi_i, \xi_i^* \geq 0.$$

With censoring indicators incorporated into the constraints, the formulation utilizes available information from both censored and noncensored observations. To make full use of the strengths of both approaches, Van Belle et al. (2011) and Pölsterl et al. (2015) further proposed hybrid approaches, combining the penalties imposed by both methods.

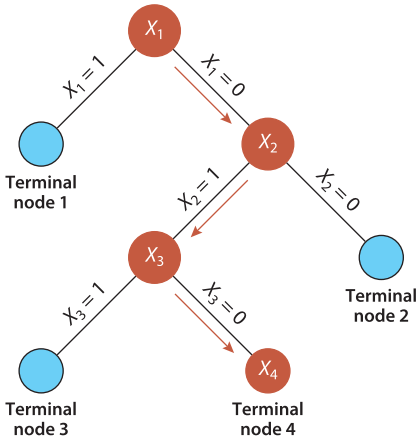

Figure 5

Illustration of a classification tree with three binary covariates that yields four terminal nodes.

4.2. Tree-Based Methods

While SVMs are adept at estimating nonlinear relationships, they do not scale well for large datasets and often underperform when the outcomes are noisy. Also, there may be no clear interpretations for classifying data points above or below the estimated hyperplane (Somvanshi et al. 2016). Decision trees are an alternative for classifying patients that provide an intuitive interpretation of the hierarchical relationships between predictors. Broadly, classification and regression trees are umbrella terms for a set of recursive partitioning algorithms, which predict the group membership (classification) or target value (regression) for an observation based on a set of binary decision rules (**Figure 5**).

Gordon & Olshen (1985) first presented survival trees, and Ciampi et al. (1986, 1987) solidified the notion and established splitting criteria based on the log-rank and likelihood ratio test statistics, respectively, gaining predictive accuracy and interpretability. A recursive partitioning algorithm for generating a survival tree is given as follows:

1. Discretize each covariate to be a binary variable (categorical variables with m levels are expressed as $m - 1$ dummy variables).
2. For every binary covariate, $X_j, j = 1, \ldots, p$, compute the log-rank statistic to test the difference between the survival curves for the two groups defined by X_j.
3. Choose the covariate—say, X_{j^*}—with the largest significant test statistic and partition the full sample (i.e., the root node) into two groups (child nodes) based on X_{j^*}.
4. Repeat steps 2 and 3 for each subset (child node) until reaching the terminal nodes—that is, no covariates produce a significant test statistic and there are enough events (exceeding a prespecified number) in each terminal node.

The resulting terminal nodes split the original sample into distinct groups, which are deemed more homogeneous within groups than across groups, and will output survival estimates via Kaplan-Meier estimation in each group. Further variations in splitting are based on metrics that accommodate censored data, either by minimizing within-node homogeneity or maximizing between-node heterogeneity. For example, these metrics can be martingale residuals (Therneau et al. 1990) or deviance residuals (LeBlanc & Crowley 1992). With an established splitting

criterion, to select a final tree, either a full survival tree is grown and pruned or a stopping rule is applied in backward or forward selection (Bou-Hamad et al. 2011).

4.3. Ensemble Learners

While survival trees provide a fast and intuitive means of studying hierarchical relationships of predictors with outcomes, they are prone to overfitting and high variability (Steingrimsson et al. 2016, Hu & Steingrimsson 2018). Ensemble learners overcome the instability issues by using techniques such as bagging, boosting, and random forests.

4.3.1. Bagging. Bootstrap aggregation, or bagging, refers to a means of training an ensemble learner by resampling the data with replacement, training weak learners (e.g., individual survival trees) in parallel, and combining these results over the multiple bootstrapped samples (Breiman 1996). It has three steps:

1. Bootstrapping: Resample from the original data of size n with replacement to form a new sample also of size n, and obtain B such samples.
2. Parallel training: With each bootstrap sample, $b = 1, \ldots, B$, independently train the weak learners in parallel.
3. Aggregation: Combine the B individual predictions by averaging over them or by taking a majority vote.

Bagging for survival trees was first proposed by Hothorn et al. (2004); in contrast to bagging for classification trees, aggregation is done by averaging survival predictions, rather than a majority vote. Each survival tree is grown so that every terminal node has enough events, which are used to predict the survival function node-wise at each terminal node. Then, for any newcomer, the predictions are averaged over the individual trees to yield an ensemble prediction of their survival function (**Figure 6**).

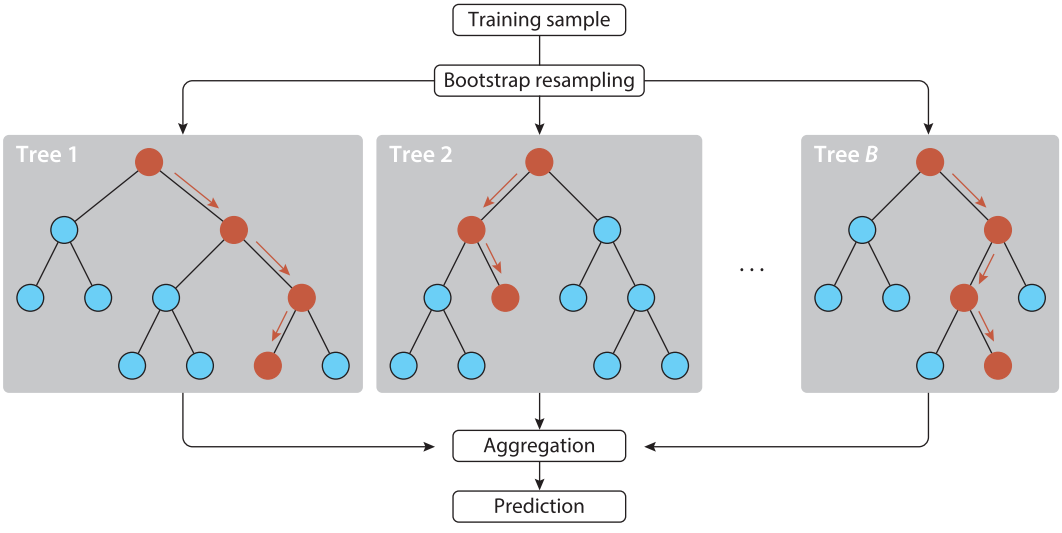

Figure 6

Ensemble learning with bootstrap aggregation (bagging) for survival trees.

4.3.2. Boosting. In a similar vein, boosting trains a series of weak learners with the goal of aggregating them into a better ensemble learner (Bühlmann & Hothorn 2007). Hothorn et al. (2006) proposed a gradient boosting algorithm for survival settings. Consider a mortality risk prediction based on covariates $\boldsymbol{X}_i$. For an M-step gradient boosting algorithm, a prediction, $\mathcal{F}_m(\boldsymbol{X}_i)$, is made at each step, say $m = 1, \ldots, M$, based on a previous prediction, $\mathcal{F}_{m-1}(\boldsymbol{X}_i)$, and an additional weak learner $f_m(\boldsymbol{X}_i)$, which is the projection of the residual error of $\mathcal{F}_{m-1}(\boldsymbol{X}_i)$ to the space spanned by $\boldsymbol{X}_i$,

$$\mathcal{F}_m(\boldsymbol{X}_i) = \mathcal{F}_{m-1}(\boldsymbol{X}_i) + w_m f_m(\boldsymbol{X}_i),$$

where $0 < w_m \leq 1$ (e.g., $w_m = 0.1$) is the step size, the residual error refers to the gradient of the loss function [e.g., the negative log partial likelihood function in a survival setting, evaluated at $\mathcal{F}_{m-1}(\boldsymbol{X}_i)$], and the number of steps, M, can be viewed as a tuning parameter.

Boosting has two notable differences from bagging. First, boosting trains weak learners sequentially, updating the weights placed on learners iteratively, whereas in bagging individual weak learners such as survival trees are trained independently and in parallel, which are aggregated via majority voting or averaging. Second, boosting is applicable to settings where learners have low variability and high bias, as the performance is improved by redistributing the weights. In contrast, bagging is often applied when individual learners exhibit high variability but low bias, as it reduces variations arising from individual trees.

4.3.3. Random forests. Yet another class of ensemble learners are random forests (Breiman 2001), which, like bagging, aggregate predictions from individual trees generated over bootstrap resampled datasets. However, unlike bagging, random forests randomly select a subset of features, say $p' < p$ features, when generating each tree and use them for the individual tree's growth. By doing so, random forests reduce correlations among individual trees, leading to gains in accuracy (Breiman 2001). The choice of p' is problem-specific, so it can also be viewed as a tuning parameter. In survival settings, Ishwaran et al. (2008) aggregated the survival predictions arising from each tree by averaging the predicted cumulative hazard functions into an ensemble prediction.

Further notable developments include those of Ishwaran et al. (2011), who extended random survival forests to high dimensions by incorporating regularization; Ishwaran & Lu (2019), who provided standard errors and confidence intervals for variable importance; and Steingrimsson et al. (2019), who proposed censoring unbiased regression trees and ensembles.

4.4. Deep Learning and Artificial Neural Networks

Deep learning has emerged as a powerful tool for risk prediction. This work stems from artificial neural networks that tried to mirror how the human brain functions (Rosenblatt 1958), wherein nodes (or neurons) are connected in a network as a weighted sum of inputs through a series of affine transformations and nonlinear activations.

A fully connected, feed-forward artificial neural network is made up of L layers, with k_l neurons in the lth layer ($l = 1, \ldots, L$) (**Figure 7**). With an input, network predictions are made based on an L-fold composite function, $f_L \circ f_{L-1} \circ \cdots \circ f_1(\cdot)$, with $(g \circ f)(\cdot) = g(f(\cdot))$. At the lth layer, $f_l(\cdot)$ is defined as

$$f_l(\mathbf{v}) = \sigma_l(\mathbf{W}_l \mathbf{v} + \mathbf{b}_l) \in \mathbb{R}^{k_l},$$

where $\mathbf{v}$ is a $k_{l-1} \times 1$ input vector fed from the $(l-1)$th layer, $\sigma_l(\cdot) : \mathbb{R}^{k_l} \to \mathbb{R}^{k_l}$ is an activation function, $\mathbf{W}_l$ is a $k_l \times k_{l-1}$ weight matrix, $\mathbf{b}_l$ is a $k_l \times 1$ bias vector, and the 0th layer is the input layer. Typical choices of $\sigma_l(\cdot)$ include the sigmoid function or the rectified linear unit activation

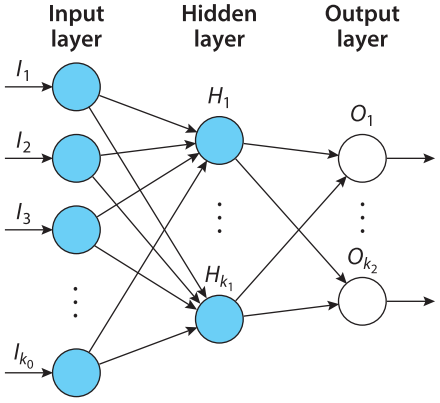

Figure 7

Diagram of a feed-forward, fully connected two-layer artificial neural network, including the hidden (first) and output (second) layer. The input (0th) layer is not counted as a real neural network layer.

function (ReLU), that is, $\sigma_l(\mathbf{b}) = \max(0, \mathbf{b})$, where $\mathbf{b} \in \mathbb{R}^{k_l}$ and $\max(0, \cdot)$ is operated componentwise.

For survival prediction, several deep learning approaches have emerged, beginning with the seminal work of Faraggi & Simon (1995), who adopted a fully connected, feed-forward neural network to extend the Cox model to nonlinear predictions. Other feed-forward neural networks (Liestbl et al. 1994, Brown et al. 1997, Biganzoli et al. 1998, Eleuteri et al. 2003) use the survival status as a training label and output predicted survival probabilities. Further developments have been made in Bayesian networks (Lisboa et al. 2003, Bellazzi & Zupan 2008, Fard et al. 2016), convolutional neural networks (Katzman et al. 2017, 2018; Ranganath et al. 2016; Yao et al. 2017), and recurrent neural networks (Yang et al. 2018).

5. PREDICTION FOR COMPETING AND SEMICOMPETING RISKS

Many survival processes in real applications involve multiple competing events. Risk prediction in these settings is an up-and-coming field with many potential developments. We focus on two common competing event settings, competing and semicompeting risks.

5.1. Competing Risks

In a competing risk setting, observing an event type, labeled by $c \in \{1, \ldots, K\}$, effectively eliminates the chance of observing other event types happening to the same individual (Young et al. 2020). For example, when studying the survival of patients with cancer, competing events can be cancer-related death ($c = 1$) or death by cardiac disease ($c = 2$) (**Figure 8**); an individual cannot die of cardiovascular disease once they have died of cancer, and vice versa. Two commonly used statistical metrics are used to characterize the risk of competing events, cause-specific hazard and subdistribution hazard, which target different counterfactual scenarios. The former describes the risk under hypothetical elimination of competing events, while the latter describes the observable risk without elimination of any competing events (Rudolph et al. 2020).

Some authors (Lau et al. 2009, Koller et al. 2012) have stated that the subdistribution hazard is useful for predicting the probability of having an event by a given time and having it be of a type of interest, termed cumulative incidence function (CIF), which reflects an individual's actual risks and prognosis. In the following, we focus on subdistribution hazard, which is derived from CIF,

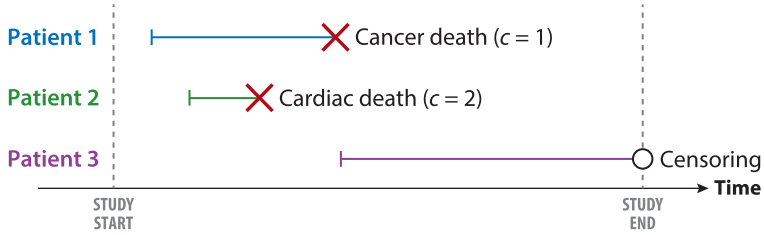

Figure 8

Schematic of observation times for three example patients with competing risks. The red crosses indicate cancer and cardiac death, and the open circle indicates censoring.

i.e., $F_c(t) = \Pr(T_i < t, \mathcal{C}_i = c)$, where $\mathcal{C}_i$ marks the event type for subject i. Specifically, for each event type $c = 1, \ldots, K$, the subdistribution hazard is defined as

$$\lambda_c(t) = \lim_{\Delta \to 0} \frac{\Pr\left(t \leq T_i < t + \Delta, \mathcal{C}_i = c \mid T_i \geq t \cup \{T_i < t \wedge \mathcal{C}_i \neq c\}\right)}{\Delta} = \frac{\mathrm{d}F_c(t)/\mathrm{d}t}{1 - F_c(t)},$$

which denotes the instantaneous risk of failure from event type c among those who have not experienced this type of event. That is, the risk set at t includes those who are event free as well as those who have experienced a competing event (other than type c) by t. The subdistribution hazard model (Fine & Gray 1999) links a subdistribution hazard function to covariates via

$$\lambda_c(t|\boldsymbol{X}_i) = \lambda_{0c}(t) \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}), \qquad 7.$$

where $\lambda_{0c}(t)$ is the baseline subdistribution hazard function for event type c, and $\boldsymbol{\beta}$ specifies the effect of $\boldsymbol{X}_i$ on the probability of event c occurring over time. In fact, the model in Equation 7 implies that $1 - F_c(t|\boldsymbol{X}_i) = \{1 - F_{0c}(t)\}^{\exp(\boldsymbol{X}_i^T \boldsymbol{\beta})}$, where $F_c(t|\boldsymbol{X}_i)$ and $F_{0c}(t)$ are the CIF given $\boldsymbol{X}_i$ and the baseline CIF, respectively.

With high-dimensional predictors, several authors (Kawaguchi et al. 2019, Ha et al. 2014, Ahn et al. 2018) proposed regularized subdistribution hazard models for variable selection, and Hou et al. (2019) further performed inference using a one-step debiased lasso estimator. For prediction, several deep learning methods for competing risks have been proposed based on CIFs. For example, Lee et al. (2018) developed a multi-task network, DeepHit, to nonparametrically estimate $F_c(t|\boldsymbol{X}_i)$ for $c = 1, \ldots, K$. The network is trained to minimize a loss function, which is constructed based on the joint distribution of the first hitting time for competing events of each subject, while ensuring the concordance of estimates across subjects (Harrell et al. 1982)—that is, a patient who died at a given time should have a higher risk at that time than a patient who survived longer. Dynamic DeepHit (Lee et al. 2019) further incorporated longitudinal information for dynamic predictions. Other approaches have included DeepCompete (Aastha & Liu 2020), as well as a hierarchical approach to multi-event survival analysis (Tjandra et al. 2021).

5.2. Semicompeting Risks

Semicompeting risk problems, a variant of competing risk problems, have commonly been encountered in clinical studies. By semicompeting, we mean that the occurrence of one event, i.e., a nonterminal event, is subject to the occurrence of another, terminal event, but not vice versa (**Figure 9**). As the nonterminal event (e.g., cancer progression) is often a strong precursor to the terminal event (death), semicompeting events are often related and, hence, the terminal event may informatively censor the nonterminal event (Jazić et al. 2016). To overcome such informative censoring, researchers either consider only the terminal event (i.e., mortality) or a composite outcome such as progression-free survival, that is, time to progression or death, whichever comes first.

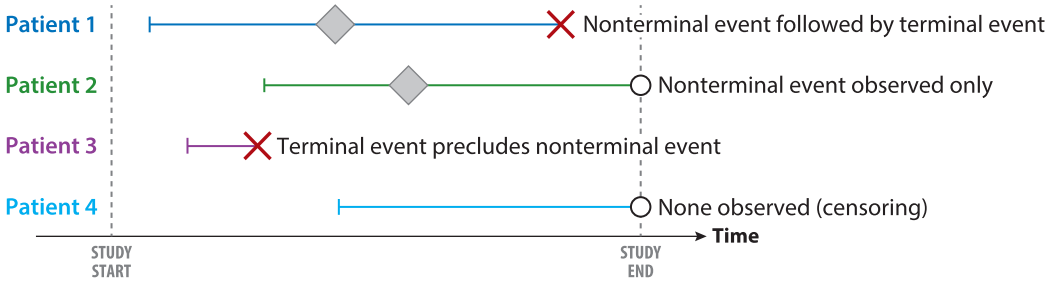

Figure 9

Schematic of four example patients with semicompeting risks. Diamonds indicate nonterminal events, crosses indicate terminal events, and open circles indicate censoring.

What is lacking in these models is how to model a predictor's potentially different roles in disease progress and death, and they ignore crucial information about the sojourn time between progression and death. Even in settings where the nonterminal and terminal event times are only modestly correlated, failing to acknowledge this sojourn time may lead to incorrect inference or biased predictions (Crilly et al. 2021).

5.2.1. The illness-death model. Central to the formulation of the semicompeting problem is the illness-death model, a compartment-type model for the rates at which individuals transition from being event-free (e.g., from time of diagnosis) to progression or to death or from progression to death (Andersen et al. 2012). Letting T_{i1}, T_{i2}, and C_i denote the times to the nonterminal and terminal events and to the censoring time, respectively, we observe $(Y_{i1},\ \delta_{i1},\ Y_{i2},\ \delta_{i2},\ \boldsymbol{X}_i)$, $i = 1,\ldots,n$, where $Y_{i2} = \min(T_{i2},\ C_i)$, $\delta_{i2} = \mathbb{I}(T_{i2} \le C_i)$, $Y_{i1} = \min(T_{i1},\ Y_{i2})$, $\delta_{i1} = \mathbb{I}(T_{i1} \le Y_{i2})$, and $\boldsymbol{X}_i$ is a p-vector of covariates. The hazards for each subject at t (since diagnosis) are defined and modeled as follows:

$$\lambda_1(t \mid \gamma_i, \boldsymbol{X}_i) = \lim_{\Delta \to 0} \frac{\Pr(t \le T_{i1} < t + \Delta | T_{i1} \ge t, T_{i2} \ge t, \gamma_i, \boldsymbol{X}_i)}{\Delta} = \gamma_i \lambda_{01}(t) \exp\left\{h_1(\boldsymbol{X}_i)\right\}, \quad 8.$$

$$\lambda_2(t \mid \gamma_i, \boldsymbol{X}_i) = \lim_{\Delta \to 0} \frac{\Pr(t \le T_{i2} < t + \Delta | T_{i1} \ge t, T_{i2} \ge t, \gamma_i, \boldsymbol{X}_i)}{\Delta} = \gamma_i \lambda_{02}(t) \exp\left\{h_2(\boldsymbol{X}_i)\right\}, \quad 9.$$

$$\lambda_3(t \mid t_1, \gamma_i, \boldsymbol{X}_i) = \lim_{\Delta \to 0} \frac{\Pr(t \le T_{i2} < t + \Delta | T_{i1} = t_1, T_{i2} \ge t, \gamma_i, \boldsymbol{X}_i)}{\Delta} = \gamma_i \lambda_{03}(t - t_1) \exp\left\{h_3(\boldsymbol{X}_i)\right\}, \quad 10.$$

where $0 < t_1 < t$, and Equations 8, 9, and 10, respectively, correspond to the transitions from diagnosis to progression prior to death, from diagnosis to death prior to progression, and from progression (that happens at t_1) to death (Haneuse & Lee 2016). Here, $\gamma_i \overset{\text{i.i.d.}}{\sim} \Gamma(1/\theta, 1/\theta), i = 1,\ldots,n$ (i.e., both shape and rate are $1/\theta$ so that the mean and variance are, respectively, 1 and θ), is a patient-specific frailty that models the dependence among these three transition processes within subject i—that is, a larger value of θ reflects a stronger dependence. In addition, $\lambda_{0g}(\cdot)$, $g = 1, 2, 3$, are the baseline hazard functions for the three state transitions, respectively, and $h_g(\cdot)$, $g = 1, 2, 3$, are log-risk functions which relate a patient's covariates to each potential transition. The λ_{0g} functions can be taken to be Weibull functions or piecewise constant with jumps at the distinct observed event times. Given Equations 8–10, and by integrating out the frailty term,

Reeder et al. (2022) derived the marginal likelihood based on n independent subjects as

$$\mathcal{L} = \prod_{i=1}^{n} \{\lambda_{1i}(Y_{i1})\}^{\delta_{i1}} \{\lambda_{2i}(Y_{i1})\}^{(1-\delta_{i1})\delta_{i2}} \{\lambda_{3i}(Y_{i2} - Y_{i1})\}^{\delta_{i1}\delta_{i2}} \left(1 + \theta^{-1}\right)^{\delta_{i1}\delta_{i2}} \times \left[1 + \theta^{-1} \{\Lambda_{1i}(Y_{i1}) + \Lambda_{2i}(Y_{i1}) + \Lambda_{3i}(Y_{i2} - Y_{i1})\}\right]^{-\theta-\delta_{i1}-\delta_{i2}}, \quad 11.$$

where $\lambda_{gi}(s) = \lambda_{0g}(s) \exp\{h_g(\boldsymbol{X}_i)\}$ and $\Lambda_{gi}(t) = \int_0^t \lambda_{gi}(s)\mathrm{d}s$ for $g = 1, 2, 3$.

5.2.2. A new deep learning approach for semicompeting risks. We propose a multi-task deep neural network for semicompeting risks (DNN-SCR), by using Equation 11 as the objective function with potentially high-dimensional covariates. DNN-SCR consists of three risk-specific subnetworks, respectively corresponding to the three possible state transitions, and a finite set of trainable parameters for specifying the baseline hazards (i.e., the ϕ parameters in **Figure 10**) if we specify Weibull baseline hazards, $\lambda_{0g}(s) = \phi_{g1}\phi_{g2}s^{\phi_{g2}-1}$ for $g = 1, 2, 3$, in Equations 8–11 as well as the dependence among the three transition processes (i.e., θ in **Figure 10**). Instead of the classical models, we opt for flexible, nonparametric estimation of $h_g(\cdot)$, $g = 1, 2, 3$, to better capture potential nonlinear dependencies of covariates on semicompeting events and to maximize the predictive accuracy.

In particular, we design three neural network subarchitectures to estimate the h functions nonparametrically as outputs. For identifiability, we require $h_g(\mathbf{0}) = 0$, $g = 1, 2, 3$, where $\mathbf{0}$ is a $p \times 1$ vector of 0's. Each subnetwork is a fully connected feed-forward neural network with ReLU activation functions and a linear activation in the final layer (**Figure 10**). The number of hidden layers and nodes per layer, as well as the dropout fraction, regularization, and learning rates are optimized as hyperparameters over a grid of values based on predictive performance. We implement

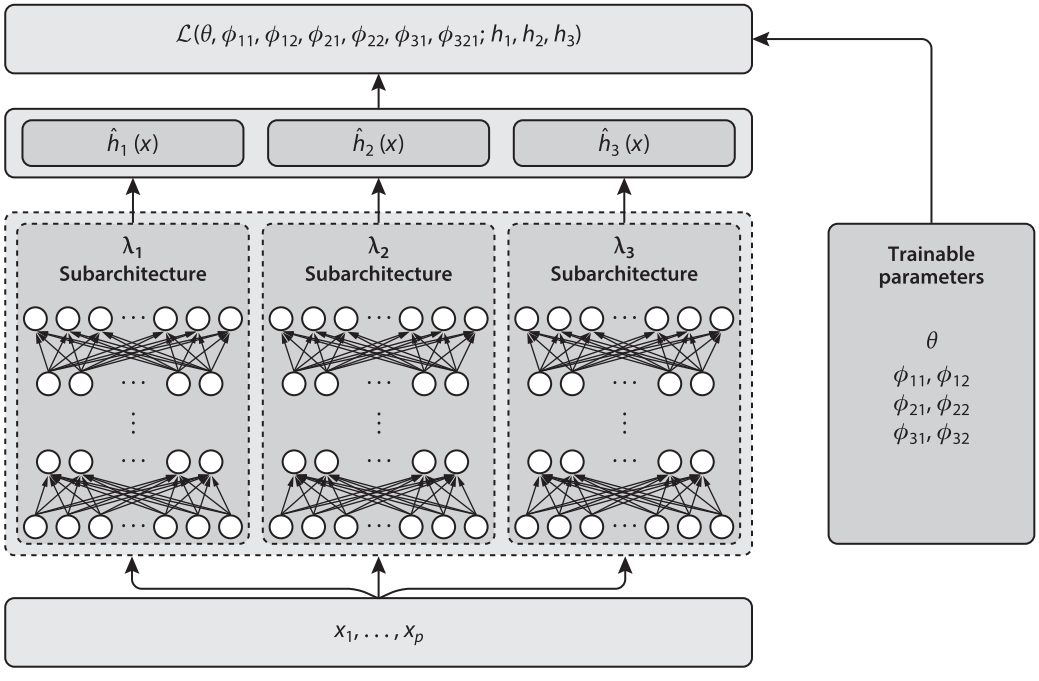

Figure 10

Architecture for the proposed semicompeting risk deep neural network.

our approach using the R interface for the deep learning library TensorFlow (Abadi et al. 2015), with model building and fitting done using the Keras application programming interface (API) (Chollet & Keras Team 2015). Finite-dimensional parameter training is done via the Gradient-Tape API (Agrawal et al. 2019) for automatic differentiation. Intensive simulations have indicated the new method predicts the risks well (**Supplemental Appendix A**).

Supplemental Material >

Revisiting the BLCSC study, we exemplify our method by studying the impact of clinical and genetic predictors on disease progression and mortality. The subset includes 5,296 patients with nonsmall cell lung cancer diagnosed between June 1983 and October 2021. Also included in the dataset are patients' characteristics, namely, age at diagnosis (years), sex (0: male, 1: female), race (0: other, 1: white), ethnicity (0: non-Hispanic, 1: Hispanic), height (meters), weight (kilograms), smoking status (0: never, 1: former, 2: current), pack-years, cancer stage (1–4), and two indicators of genetic mutations [epidermal growth factor receptor (EGFR) and Kirsten rat sarcoma viral oncogene homolog (KRAS)]. Semicompeting events of cancer progression and death are documented in the data; the date of progression is the date of the first source evidence, including exam, radiology report or pathology. Progression followed by death is observed in 111 patients (2%), progression but alive at the last follow-up date is observed in 224 patients (4%), and death prior to progression is observed among 1,916 patients (36%).

To investigate the dependence of disease progression on death and predict the transition processes, we fit the models in Equations 8–10 via a DNN-SCR. Specifically, we assume Weibull baseline hazards $\lambda_{0g}(s)$, $g = 1, 2, 3$, and $\gamma_i \overset{\text{i.i.d.}}{\sim} \Gamma(1/\theta, 1/\theta)$ (as specified following Equation 10) and let $\boldsymbol{X}_i$ be the ith patient's characteristics, $i = 1, \ldots, 5,296$. We then use DNN-SCR to optimize the objective function (Equation 11) in order to output the estimates of the finite-dimensional parameters (ϕs and θ) and the predicted h_g, $g = 1, 2, 3$ (log risk estimates), for any covariate values.

We estimate the frailty variance, θ, to be 3.15 (bootstrapped 95% confidence interval: 3.02–3.29), which suggests that progression is indeed correlated with death. **Figure 11** depicts the h functions (log risks) for the effect of age at diagnosis on each state transition, stratified by sex and initial cancer stage while fixing the other covariates to be at their sample means or modes. There seems to exist a nonlinear effect of age that differs by transition, cancer stage, and sex.

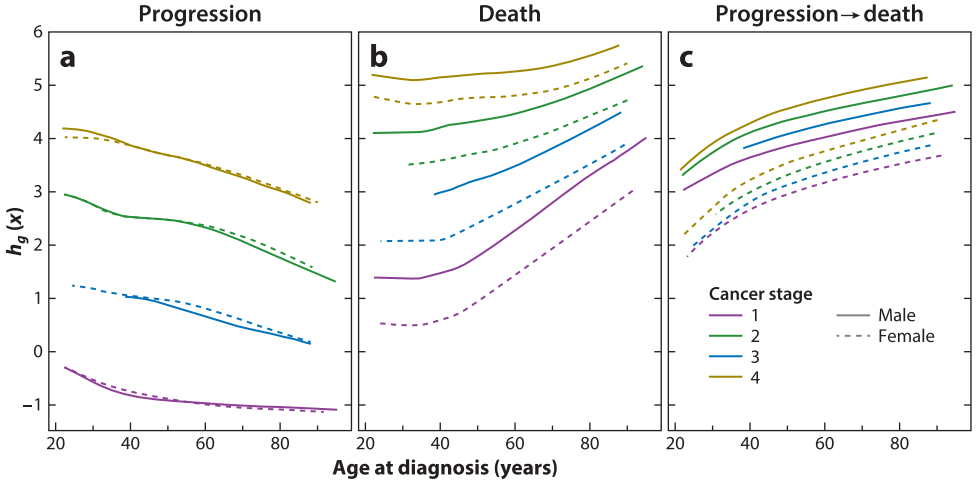

Figure 11

Log risk functions of age at diagnosis on each state transition, stratified by sex (solid versus dashed lines) and initial cancer stage (line color).

Figure 11*a* shows that younger age and more advanced stage are associated with higher hazards for progression; for the transitions from diagnosis or progression to death (**Figure 11*b*,*c***), older age is associated with higher hazards; interestingly, while sex does not seem to play a role in disease progression (**Figure 11*a***), male patients are more likely to die than female patients after diagnosis (**Figure 11*b***) or after progression (**Figure 11*c***). Finally, more advanced stage is associated with higher hazards for all the transitions.

6. CONCLUSIONS

We have presented various methods for analyzing survival outcome data with high-dimensional predictors. We first provided a primer on time-to-event data and the unique features of survival analysis that make it distinct from other areas of statistics. We then reviewed regularized approaches for extending classical models such as the Cox, AFT, and censored quantile regression models, which lay the foundation for much of the subsequent work in this field, to high-dimensional settings. We briefly touched on feature screening for ultra high-dimensional predictors and discussed high-dimensional inference with survival data. Finally, we focused on machine learning for survival prediction and concluded with methods at the forefront of the field of prognostication with competing event data.

Supplemental Material >

This review is intended to provide a road map for readers interested in high-dimensional survival analysis (see **Supplemental Appendix B** for a tabulation of the reviewed methods and their available software), though our review is by no means exhaustive. This is an exciting and rapidly evolving field, with many open questions and new developments. For example, progress in survival predictions with high-dimensional predictors, including deep learning, active learning, and transfer learning, will open new avenues to interdisciplinary breakthroughs in biomedical research and methods with greater prognostic utility. Also, our review is mainly focused on frequentist methods, and in the past decade, a significant portion of Bayesian works (Annest et al. 2009, Bonato et al. 2011, Lee 2011, Wang et al. 2013b, Pungpapong 2021) have appeared, which make the field even more exciting. This review pays tribute to the late Sir D.R. Cox, whose work in survival analysis has fundamentally changed the paradigm of biomedical research and will continue to impact future research for years to come.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank our long-term collaborator, Dr. David C. Christiani, for providing the BLCSC data and Dr. Xinan Wang for helpful discussion of the BLCSC application results. We thank Dr. Ingrid van Keilegom and a reviewer for many helpful suggestions that significantly improved the quality of the article. The work is partially supported by grants from the National Institutes of Health.

LITERATURE CITED

Aastha PH, Liu Y. 2020. DeepCompete: A deep learning approach to competing risks in continuous time domain. *AMIA Annu. Symp. Proc.* 2020:177–86

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, et al. 2015. TensorFlow: large-scale machine learning on heterogeneous systems. *Software Platform for Machine Learning*. **https://www.tensorflow.org/**

Agrawal A, Modi A, Passos A, Lavoie A, Agarwal A, et al. 2019. TensorFlow Eager: a multi-stage, Python-embedded DSL for machine learning. *Proc. Mach. Learn. Syst.* 1:178–89

Ahn KW, Banerjee A, Sahr N, Kim S. 2018. Group and within-group variable selection for competing risks data. *Lifetime Data Anal.* 24(3):407–24

Andersen PK, Borgan O, Gill RD, Keiding N. 2012. *Statistical Models Based on Counting Processes*. New York: Springer

Annest A, Bumgarner RE, Raftery AE, Yeung KY. 2009. Iterative Bayesian model averaging: a method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformat.* 10:72

Antoniadis A, Fryzlewicz P, Letué F. 2010. The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* 37(4):531–52

Austin PC, Fine JP. 2017. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat. Med.* 36(27):4391–400

Barbeau EM, Li Y, Calderon P, Hartman C, Quinn M, et al. 2006. Results of a union-based smoking cessation intervention for apprentice iron workers (United States). *Cancer Causes Control* 17(1):53–61

Bellazzi R, Zupan B. 2008. Predictive data mining in clinical medicine: current issues and guidelines. *Int. J. Med. Informat.* 77(2):81–97

Belloni A, Chernozhukov V, Kato K. 2019. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *J. Am. Stat. Assoc.* 114(526):749–58

Biganzoli E, Boracchi P, Mariani L, Marubini E. 1998. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat. Med.* 17(10):1169–86

Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do KA. 2011. Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* 27(3):359–67

Bou-Hamad I, Larocque D, Ben-Ameur H. 2011. A review of survival trees. *Stat. Surv.* 5:44–71

Bradic J, Fan J, Jiang J. 2011. Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Stat.* 39(6):3092–120

Breiman L. 1996. Bagging predictors. *Mach. Learn.* 24(2):123–40

Breiman L. 2001. Random forests. *Mach. Learn.* 45(1):5–32

Brown SF, Branford AJ, Moran W. 1997. On the use of artificial neural networks for the analysis of survival data. *IEEE Trans. Neural Netw.* 8(5):1071–77

Buckley J, James I. 1979. Linear regression with censored data. *Biometrika* 66(3):429–36

Bühlmann P, Hothorn T. 2007. Boosting algorithms: regularization, prediction and model fitting. *Stat. Sci.* 22(4):477–505

Candès E, Tao T. 2007. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.* 35(6):2313–51

Chaturvedi N, de Menezes RX, Goeman JJ. 2014. Fused lasso algorithm for Cox proportional hazards and binomial logit models with application to copy number profiles. *Biometrical J.* 56(3):477–92

Chollet F, Keras Team. 2015. Keras: deep learning for humans. *Deep Learning Software*. **https://github.com/fchollet/keras**

Christiani DC. 2017. *The Boston Lung Cancer Survival Cohort*. Grant Proposal, Harvard Univ., Cambridge, MA. **http://grantome.com/grant/NIH/U01-CA209414-01A1**

Ciampi A, Chang CH, Hogg S, McKinney S. 1987. Recursive partition: a versatile method for exploratory-data analysis in biostatistics. In *Biostatistics*, ed. IB MacNeill, GJ Umphrey, A Donner, VK Jandhyala, pp. 23–50. New York: Springer

Ciampi A, Thiffault J, Nakache JP, Asselain B. 1986. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Comput. Stat. Data Anal.* 4(3):185–204

Cox DR. 1972. Regression models and life tables (with discussion). *J. R. Stat. Soc. Ser. B* 34(2):187–220

Crilly CJ, Haneuse S, Litt JS. 2021. Predicting the outcomes of preterm neonates beyond the neonatal intensive care unit: What are we missing? *Pediatr. Res.* 89(3):426–45

Efron B. 2014. Estimation and accuracy after model selection. *J. Am. Stat. Assoc.* 109(507):991–1007

Eleuteri A, Tagliaferri R, Milano L, De Placido S, De Laurentiis M. 2003. A novel neural network-based survival analysis model. *Neural Netw.* 16(5–6):855–64

Fan J, Feng Y, Wu Y. 2010. High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*, ed. JO Berger, TT Cai, IM Johnstone, pp. 70–86. Beachwood, OH: Inst. Math. Stat.

Fan J, Li R. 2002. Variable selection for Cox's proportional hazards model and frailty model. *Ann. Stat.* 30(1):74–99

Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* 70(5):849–911

Faraggi D, Simon R. 1995. A neural network model for survival data. *Stat. Med.* 14(1):73–82

Fard MJ, Wang P, Chawla S, Reddy CK. 2016. A Bayesian perspective on early stage event prediction in longitudinal data. *IEEE Trans. Knowledge Data Eng.* 28(12):3126–39

Fei Z, Li Y. 2021. Estimation and inference for high dimensional generalized linear models: a splitting and smoothing approach. *J. Mach. Learn. Res.* 22:58

Fei Z, Zheng Q, Hong HG, Li Y. 2021. Inference for high-dimensional censored quantile regression. *J. Am. Stat. Assoc.* **https://doi.org/10.1080/01621459.2021.1957900**

Fei Z, Zhu J, Banerjee M, Li Y. 2019. Drawing inferences for high-dimensional linear models: a selection-assisted partial regression and smoothing approach. *Biometrics* 75(2):551–61

Fine JP, Gray RJ. 1999. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* 94(446):496–509

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33(1):1–22

Gordon L, Olshen RA. 1985. Tree-structured survival analysis. *Cancer Treatment Rep.* 69(10):1065–69

Ha ID, Lee M, Oh S, Jeong JH, Sylvester R, Lee Y. 2014. Variable selection in subdistribution hazard frailty models with competing risks data. *Stat. Med.* 33(26):4590–604

Haneuse S, Lee KH. 2016. Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circ. Cardiovasc. Qual. Outcomes* 9(3):322–31

Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. 1982. Evaluating the yield of medical tests. *JAMA* 247(18):2543–46

He X, Wang L, Hong HG. 2013. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Stat.* 41(1):342–69

Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67

Hong HG, Chen X, Christiani DC, Li Y. 2018a. Integrated powered density: screening ultrahigh dimensional covariates with survival outcomes. *Biometrics* 74(2):421–29

Hong HG, Chen X, Kang J, Li Y. 2020. The L_q-norm learning for ultrahigh-dimensional survival data: an integrative framework. *Stat. Sin.* 30(3):1213–33

Hong HG, Christiani DC, Li Y. 2019. Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine. *Precision Clin. Med.* 2(2):90–99

Hong HG, Kang J, Li Y. 2018b. Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Anal.* 24(1):45–71

Hong HG, Li Y. 2017. Feature selection of ultrahigh-dimensional covariates with survival outcomes: a selective review. *Appl. Math.* 32(4):379–96

Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. 2006. Survival ensembles. *Biostatistics* 7(3):355–73

Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. 2004. Bagging survival trees. *Stat. Med.* 23(1):77–91

Hou J, Bradic J, Xu R. 2019. Inference under fine-gray competing risks model with high-dimensional covariates. *Electron. J. Stat.* 13(2):4449–507

Hu C, Steingrimsson JA. 2018. Personalized risk prediction in clinical oncology research: applications and practical issues using survival trees and random forests. *J. Biopharmaceut. Stat.* 28(2):333–49

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. 2008. Random survival forests. *Ann. Appl. Stat.* 2(3):841–60

Ishwaran H, Kogalur UB, Chen X, Minn AJ. 2011. Random survival forests for high-dimensional data. *Stat. Anal. Data Mining* 4(1):115–32

Ishwaran H, Lu M. 2019. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat. Med.* 38(4):558–82

Javanmard A, Montanari A. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* 15(1):2869–909

Jazić I, Schrag D, Sargent DJ, Haneuse S. 2016. Beyond composite endpoints analysis: semicompeting risks as an underutilized framework for cancer research. *J. Natl. Cancer Inst.* 108(12):djw154

Kang J, Hong HG, Li Y. 2017. Partition-based ultrahigh-dimensional variable screening. *Biometrika* 104(4):785–800

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. 2017. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. arXiv:1606.00931v3 [stat.ML]

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* 18:24

Kawaguchi ES, Shen JI, Li G, Suchard MA. 2019. A fast and scalable implementation method for competing risks data with the R package fastcmprsk. arXiv:1905.07438 [stat.CO]

Kim J, Sohn I, Jung SH, Kim S, Park C. 2012. Analysis of survival data with group lasso. *Commun. Stat. Simul. Comput.* 41(9):1593–605

Koller MT, Raatz H, Steyerberg EW, Wolbers M. 2012. Competing risks and the clinical community: irrelevance or ignorance? *Stat. Med.* 31(11–12):1089–97

Lau B, Cole SR, Gange SJ. 2009. Competing risk regression models for epidemiologic data. *Am. J. Epidemiol.* 170(2):244–56

LeBlanc M, Crowley J. 1992. Relative risk trees for censored survival data. *Biometrics* 48:411–25

Lee C, Yoon J, Van Der Schaar M. 2019. Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans. Biomed. Eng.* 67(1):122–33

Lee C, Zame WR, Yoon J, van der Schaar M. 2018. DeepHit: a deep learning approach to survival analysis with competing risks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, ed. SA McIlraith, KQ Weinberger, pp. 2314–21. Menlo Park, CA: AAAI

Lee KH. 2011. *Bayesian variable selection in parametric and semiparametric high dimensional survival analysis*. PhD Thesis, Univ. Mo., Columbia, Mo.

Li Y, Dicker L, Zhao SD. 2014. The Dantzig selector for censored linear regression models. *Stat. Sin.* 24(1):251–58

Liestbl K, Andersen PK, Andersen U. 1994. Survival analysis and neural nets. *Stat. Med.* 13(12):1189–200

Lisboa PJ, Wong H, Harris P, Swindell R. 2003. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artif. Intel. Med.* 28(1):1–25

Liu Y, Chen X, Li G. 2020. A new joint screening method for right-censored time-to-event data with ultra-high dimensional covariates. *Stat. Methods Med. Res.* 29(6):1499–513

Loh PL, Wainwright MJ. 2017. Support recovery without incoherence: a case for nonconvex regularization. *Ann. Stat.* 45(6):2455–82

Ma Y, Li Y, Lin H. 2017. Concordance measure-based feature screening and variable selection. *Stat. Sin.* 27:1967–85

Noble WS. 2006. What is a support vector machine? *Nat. Biotechnol.* 24(12):1565–67

Peng L, Huang Y. 2008. Survival analysis with quantile regression models. *J. Am. Stat. Assoc.* 103(482):637–49

Pijyan A, Zheng Q, Hong HG, Li Y. 2020. Consistent estimation of generalized linear models with high dimensional predictors via stepwise regression. *Entropy* 22(9):965

Pölsterl S, Navab N, Katouzian A. 2015. Fast training of support vector machines for survival analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015 Porto, Portugal, September 711, 2015 Proceedings, Part I*, ed. A Appice, P Pereira Rodrigues, V Santos Costa, C Soares, J Gama, A Jorge, pp. 243–59. New York: Springer

Portnoy S. 2003. Censored regression quantiles. *J. Am. Stat. Assoc.* 98(464):1001–12

Powell JL. 1986. Censored regression quantiles. *J. Econom.* 32(1):143–55

Pungpapong V. 2021. Incorporating biological networks into high-dimensional Bayesian survival analysis using an ICM/M algorithm. *J. Bioinformat. Comput. Biol.* 19(05):2150027

Ranganath R, Perotte A, Elhadad N, Blei D. 2016. Deep survival analysis. *PMLR* 56:101–14
Reeder HT, Lu J, Haneuse S. 2022. Penalized estimation of frailty-based illness-death models for semi-competing risks. arXiv:2202.00618 [stat.ME]
Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65(6):386–408
Rudolph JE, Lesko CR, Naimi AI. 2020. Causal inference in the face of competing events. *Curr. Epidemiol. Rep.* 7(3):125–31
Saikia R, Barman MP. 2017. A review on accelerated failure time models. *Int. J. Stat. Syst.* 12(2):311–22
Shivaswamy PK, Chu W, Jansche M. 2007. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 655–60. New York: IEEE
Smola AJ, Schölkopf B. 2004. A tutorial on support vector regression. *Stat. Comput.* 14(3):199–222
Somvanshi M, Chavan P, Tambade S, Shinde SV. 2016. A review of machine learning techniques using decision tree and support vector machine. In *2016 International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–7. New York: IEEE
Steingrimsson JA, Diao L, Molinaro AM, Strawderman RL. 2016. Doubly robust survival trees. *Stat. Med.* 35(20):3595–612
Steingrimsson JA, Diao L, Strawderman RL. 2019. Censoring unbiased regression trees and ensembles. *J. Am. Stat. Assoc.* 114(525):370–83
Therneau TM, Grambsch PM, Fleming TR. 1990. Martingale-based residuals for survival models. *Biometrika* 77(1):147–60
Tibshirani RJ. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58:267–88
Tibshirani RJ. 1997. The lasso method for variable selection in the Cox model. *Stat. Med.* 16:385–95
Tibshirani RJ. 2009. Univariate shrinkage in the Cox model for high dimensional data. *Stat. Appl. Genet. Mol. Biol.* 8:21
Tjandra D, He Y, Wiens J. 2021. A hierarchical approach to multi-event survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 591–99. Menlo Park, CA: AAAI
Van Belle V, Pelckmans K, Suykens J, Van Huffel S. 2007. *Support vector machines for survival analysis.* Paper presented at the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), Plymouth, UK, Jul. 25–27
Van Belle V, Pelckmans K, Van Huffel S, Suykens JA. 2011. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif. Intel. Med.* 53(2):107–18
Van de Geer S, Bühlmann P, Ritov Y, Dezeure R. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.* 42(3):1166–202
Van der Vaart AW. 2000. *Asymptotic Statistics*, Vol. 3. Cambridge, UK: Cambridge Univ. Press
Vapnik V, Guyon I, Hastie T. 1995. Support vector machines. *Mach. Learn.* 20(3):273–97
Verweij PJ, Van Houwelingen HC. 1994. Penalized likelihood in Cox regression. *Stat. Med.* 13(23–24):2427–36
Vinzamuri B, Reddy CK. 2013. Cox regression with correlation based regularization for electronic health records. In *2013 IEEE 13th International Conference on Data Mining*, pp. 757–66. New York: IEEE
Wahba G. 1999. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods: Support Vector Learning*, ed. B Schölkopf, CJC Burges, AJ Smola, pp. 69–87. Cambridge, MA: MIT Press
Wang HJ, Zhou J, Li Y. 2013a. Variable selection for censored quantile regression. *Stat. Sin.* 23(1):145–67
Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA. 2013b. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* 29(2):149–59
Wu Y. 2012. Elastic net for Cox's proportional hazards model with a solution path algorithm. *Stat. Sin.* 22(1):271–94
Xia L, Nan B, Li Y. 2022. Debiased lasso for generalized linear models with a diverging number of covariates. *Biometrics*. In press
Xia L, Nan B, Li Y. 2021. Statistical inference for Cox proportional hazards models with a diverging number of covariates. arXiv:2106.03244 [stat.ME]
Yang G, Cai Y, Reddy CK. 2018. Spatio-temporal check-in time prediction with recurrent neural network based survival analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2976–2983. N.P.: IJCAI

Yao J, Zhu X, Zhu F, Huang J. 2017. Deep correlational learning for survival prediction from multi-modality data. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017*, ed. M Descoteaux, L Maier-Hein, A Franz, P Jannin, DL Collins, S Duchesne, pp. 406–14. New York: Springer

Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA. 2020. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat. Med.* 39(8):1199–236

Yu Y, Bradic J, Samworth RJ. 2018. Confidence intervals for high-dimensional Cox models. arXiv:1803.01150 [stat.ME]

Zhang HH, Lu W. 2007. Adaptive lasso for Cox's proportional hazards model. *Biometrika* 94(3):691–703

Zhao P, Yu B. 2006. On model selection consistency of lasso. *J. Mach. Learn. Res.* 7:2541–63

Zhao SD, Li Y. 2012. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivar. Anal.* 105:397–411

Zhao SD, Li Y. 2014. Score test variable screening. *Biometrics* 70(4):862–71

Zheng Q, Gallagher C, Kulasekera K. 2013. Adaptive penalized quantile regression for high dimensional data. *J. Stat. Plan. Inference* 143(6):1029–38

Zheng Q, Peng L, He X. 2018. High dimensional censored quantile regression. *Ann. Stat.* 46:308–43

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67(2):301–20