Machine Learning for Survival Analysis: A Survey

PING WANG, Virginia Tech
YAN LI, University of Michigan, Ann Arbor
CHANDAN K. REDDY, Virginia Tech

Survival analysis is a subfield of statistics where the goal is to analyze and model data where the outcome is the time until an event of interest occurs. One of the main challenges in this context is the presence of instances whose event outcomes become unobservable after a certain time point or when some instances do not experience any event during the monitoring period. This so-called *censoring* can be handled most effectively using survival analysis techniques. Traditionally, statistical approaches have been widely developed in the literature to overcome the issue of censoring. In addition, many machine learning algorithms have been adapted to deal with such censored data and tackle other challenging problems that arise in real-world data. In this survey, we provide a comprehensive and structured review of the statistical methods typically used and the machine learning techniques developed for survival analysis, along with a detailed taxonomy of the existing methods. We also discuss several topics that are closely related to survival analysis and describe several successful applications in a variety of real-world application domains. We hope that this article will give readers a more comprehensive understanding of recent advances in survival analysis and offer some guidelines for applying these approaches to solve new problems arising in applications involving censored data.

CCS Concepts: • **Mathematics of computing** → **Survival analysis**; • **Computing methodologies** → **Machine learning**; • **Information systems** → **Data mining**;

Additional Key Words and Phrases: Machine learning, survival analysis, censoring, regression, hazard rate, Cox model, concordance index, survival data

ACM Reference format:
Ping Wang, Yan Li, and Chandan k. Reddy. 2019. Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv.* 51, 6, Article 110 (February 2019), 36 pages.
https://doi.org/10.1145/3214306

1 INTRODUCTION

Due to the development of various new data acquisition and big data technologies, the ability to collect a wide variety of data and monitor observations over long periods is now reality in many different disciplines. For most of these real-world applications, the primary objective is to obtain

This material is based upon work supported by, or in part by, the U.S. National Science Foundation grants IIS-1707498, IIS-1619028 and IIS-1838730.
Authors' addresses: P. Wang is with the Department of Computer Science, Virginia Tech, 900 N. Glebe Road, Arlington, VA, 22203. email: ping@vt.edu; Y. Li is with the Machine Intelligence Research Sector, Alibaba DAMO Academy, 500 108th Ave NE Bellevue, WA 98004; email: yl.yy6993@alibaba-inc.com; C. K. Reddy is with the Department of Computer Science, Virginia Tech, 900 N. Glebe Road, Arlington, VA, 22203. email: reddy@cs.vt.edu

0360-0300/2019/02-ART110 $15.00
https://doi.org/10.1145/3214306

a better estimate of the time a particular event of interest occurs. One of the main challenges for such time-to-event data is that censored instances where the event of interest is not observed either due to time limitations or losing track during the observation period are not uncommon. It often happens that instances that have experienced an event (or are labeled as an event) occur but information about the outcome variable for the remaining instances is only available up to a specific time point. This makes it impractical to apply predictive algorithms directly using the statistical and machine learning approaches normally used to analyze the survival data. Survival analysis, which is an important subfield of statistics, provides various mechanisms to handle such censored data problems when they arise in modeling complex data. This is also referred to as *time-to-event data* when modeling a particular event of interest is the main objective and is very common in real-world application domains.

In addition to the difficulty of handling censored data, there are other unique challenges that must be overcome when performing predictive modeling with survival data. A number of researchers have begun to work on developing new computational algorithms that are capable of handling such complex challenges. To address these practical concerns, some related works have adapted machine learning methods to solve survival analysis problems, and machine learning researchers have developed a suite of sophisticated and effective algorithms that either complement or compete with the traditional statistical methods. However, in spite of the importance of these problems and their relevance to various real-world applications, research on this topic is scattered across several different disciplines. Moreover, few surveys are available in the literature on this topic, and, to the best of our knowledge, there has not yet been a comprehensive review of survival analysis and its recent developments from a machine learning perspective. Almost all of these existing survey articles focus solely on the statistical methods utilized and either completely ignore or barely mention recent advances in machine learning in this field. One of the earliest surveys (Chung et al. 1991) gives an overview of the statistical survival analysis methods and describes their application in criminology for predicting the time until recidivism. Most of the existing books on survival analysis (Kleinbaum and Klein 2006; Lee and Wang 2003; Allison 2010) introduce this topic from a traditional statistical perspective rather than a machine learning standpoint. Recently, however, this has begun to change and Cruz and Wishart (2006) and Kourou et al. (2015) both discussed applications in cancer prediction and included a comparison of several machine learning techniques.

The primary purpose of this survey article is therefore to provide a comprehensive and structured overview of various machine learning methods for survival analysis, along with a discussion of the traditional statistical methods. We demonstrate the commonly used evaluation metrics and advanced related formulations that are typically investigated in connection with this research topic. We will discuss a detailed taxonomy of all the survival analysis methods developed in traditional statistics, as well as those that have been proposed more recently by the machine learning community. We will also provide links to various implementations and source codes which will enable readers to delve more deeply into the methods discussed in this article. Finally, we will discuss various applications of survival analysis.

The rest of this article is organized as follows. We will begin with a brief review of the basic concepts, notations, and definitions that are necessary to comprehend the survival analysis algorithms and provide a formal problem statement for the survival analysis problem in Section 2. A taxonomy of the existing survival analysis methods, including both statistical and machine learning methods, will also provide a holistic view of existing works in the area of survival analysis. We will then move on to review the existing well-studied representative conventional statistical methods, including nonparametric, semi-parametric, and parametric models, in Section 3. Section 4 describes several basic machine learning approaches, including survival trees, Bayesian methods, support

vector machines, and neural networks that have been specifically developed for survival analysis. Different kinds of advanced machine learning algorithms such as ensemble learning, transfer learning, multitask learning, and active learning for handling survival data will also be discussed. Section 5 focuses on appropriate evaluation metrics for survival models. In addition to the survival analysis algorithms, some interesting topics related to this topic have attracted considerable attention in various fields. In Section 6, several related concepts such as early prediction and complex events will be discussed. Various data transformation techniques such as uncensoring and calibration, both of which are typically used in conjunction with existing predictive methods, will also be mentioned briefly. A discussion of relevant topics in complex event analysis, such as competing risks and recurrent events, will also be provided. In Section 7, some real-world applications of survival analysis methods will be briefly explained, and more insights into these application domains are provided. In Section 8, details regarding various implementations and software packages for survival analysis methods are discussed. Finally, Section 9 concludes our discussion.

2 DEFINITION OF SURVIVAL ANALYSIS

In this section, we begin by presenting the basic notations and terminologies used in this article. We then provide an illustrative example that explains the structure of survival data and give a more formal problem statement for survival analysis. Finally, we will also provide a complete taxonomy of the existing survival analysis methods that have been reported in the literature, including both conventional statistical methods and machine learning approaches. The objective here is to provide a holistic view of the field of survival analysis and equip readers with the basic knowledge about the methods used in this field before moving on to consider the detailed algorithms.

2.1 Survival Data and Censoring

During the study of a survival analysis problem, it is possible that the events of interest are not observed for some instances. This may be because of either the limited observation time window or missing traces caused by other uninterested events. This concept is known as *censoring* (Klein and Moeschberger 2005). We can broadly categorize censoring into three groups based on the reason that the censoring has occurred (Lee and Wang 2003): (i) *right-censoring*, where the observed survival time is less than or equal to the true survival time; (ii) *left-censoring*, where the observed survival time is greater than or equal to the true survival time; and (iii) *interval censoring*, where we only know that the event occurs during a given time interval. Note that the true event occurrence time is unknown in all three cases. Of these, right-censoring is the most common scenario, and it arises in many practical problems (Marubini and Valsecchi 2004), hence survival data with right-censoring will be the main focus of this article. Table 1 describes the notations used in this article.

For a survival problem, the time to the event of interest (T) is known precisely only for those instances where the event occurs during the study period. For the remaining instances, since we may lose track of them during the observation time or their time to event is greater than the observation time, we can only have the censored time (C), which may be the time of withdrawn, lost, or the end of the observation. These are considered to be censored instances in the context of survival analysis. In other words, we can only observe either survival time (T_i) or censored time (C_i) but not both for any given instance i. If, and only if, $y_i = min(T_i, C_i)$ can be observed during the study, the dataset is said to be right-censored. In a survival analysis problem with right-censored instances, the censoring time is also a random variable since the subjects enter the study randomly and the time of withdrawal or loss of tracking is also random. Thus, in this article, we assume that the censoring occurs randomly in the survival problems. For the sake of brevity, in this article, we will refer to randomly occurring right-censoring as *censoring* from now on.

Table 1. Notations Used in this Article

Notations	Descriptions
P	Number of features
N	Number of instances
X	$\mathbb{R}^{N\times P}$ feature vector
X_i	$\mathbb{R}^{1\times P}$ covariate vector of instance i
T	$\mathbb{R}^{N\times 1}$ vector of event times
C	$\mathbb{R}^{N\times 1}$ vector of last follow-up times
y	$\mathbb{R}^{N\times 1}$ vector of observed time which is equal to $min(T, C)$
δ	$N \times 1$ binary vector for event status
β	$\mathbb{R}^{P\times 1}$ coefficient vector
$f(t)$	Death density function
$F(t)$	Cumulative event probability function
$S(t)$	Survival probability function
$h(t)$	Hazard function
$h_0(t)$	Baseline hazard function
$H(t)$	Cumulative hazard function

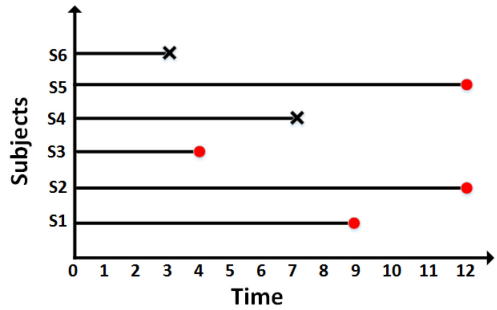

Fig. 1. An illustration demonstrating the survival analysis problem.

Figure 1 shows an example to help provide a better understanding of the definition of censoring and the structure of survival data. Six instances are observed in this study over a 12-month period, and the event occurrence information during this time period is recorded. From Figure 1, we can see that only subjects S4 and S6 actually experienced the event (marked by 'X') during the follow-up time, and the observed time for them will be the event time. As the event did not occur within the 12-month monitoring period for subjects S1, S2, S3, and S5, these are considered to be censored and are thus marked as red dots in the figure. More specifically, subjects S2 and S5 are censored since no event occurred during the study period, while subjects S1 and S3 are censored due to withdrawal or the follow-up being lost within the study time period.

Problem Statement: For a given instance i, represented by a triplet (X_i, y_i, δ_i), where $X_i \in \mathbb{R}^{1\times P}$ is the feature vector; δ_i is the binary event indicator (i.e., $\delta_i = 1$ for an uncensored instance and $\delta_i = 0$ for a censored instance); and y_i denotes the observed time and is equal to the survival time T_i for an uncensored instance and C_i for a censored instance; that is,

$$y_i = \begin{cases} T_i & \text{if } \delta_i = 1 \\ C_i & \text{if } \delta_i = 0 \end{cases}. \tag{1}$$

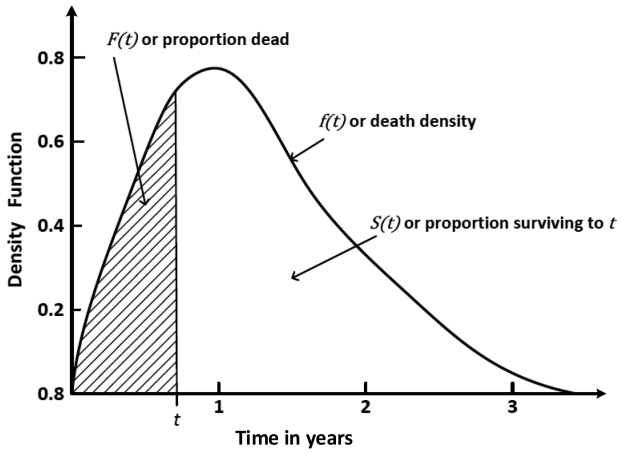

Fig. 2. Relationships between the different entities $f(t)$, $F(t)$, and $S(t)$.

T_i is a latent value for censored instances since these instances did not experience any event during the observation time period.

The goal of survival analysis is to estimate the time to the event of interest T_j for a new instance j with feature predictors denoted by X_j. Note that in survival analysis problems the value of T_j will be both non-negative and continuous.

2.2 Survival and Hazard Function

The ***survival function***, which is used to represent the probability that the time to the event of interest is not earlier than a specified time t (Lee and Wang 2003; Klein and Moeschberger 2005), is one of the primary goals in survival analysis. Conventionally, the survival function is represented by S, which is given as follows:

$$S(t) = Pr(T \geq t). \tag{2}$$

The survival function monotonically decreases with t, and the initial value is 1 when $t = 0$, which signifies that, at the beginning of the observation, 100% of the observed subjects survive; in other words, none of the events of interest has occurred.

In contrast, the *cumulative death distribution function* $F(t)$, which represents the probability that the event of interest occurs earlier than t, is defined as $F(t) = 1 - S(t)$, and the ***death density function*** can be obtained as $f(t) = \frac{d}{dt}F(t)$ for continuous cases, and $f(t) = [F(t + \Delta t) - F(t)]/\Delta t$, where Δt denotes a small time interval, for discrete cases. Figure 2 shows the relationships between these functions.

In survival analysis, another commonly used function is the ***hazard function*** ($h(t)$), which may also be referred to as the *force of mortality*, the *instantaneous death rate,* or the *conditional failure rate* (Dunn and Clark 2009). The hazard function does not indicate the chance or probability that the event of interest will occur, but instead represents the likelihood of the event occurring at time t given that no event has occurred before time t. Mathematically, the hazard function is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \tag{3}$$

Similar to $S(t)$, $h(t)$ is also a non-negative function. While all the survival functions $S(t)$ decrease over time, the hazard function can have a variety of shapes. Consider the definition of $f(t)$. As

this can also be expressed as $f(t) = -\frac{d}{dt}S(t)$, the hazard function can be represented as:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}S(t) \cdot \frac{1}{S(t)} = -\frac{d}{dt}[lnS(t)]. \tag{4}$$

Thus, the survival function defined in Equartion (2) can be rewritten as

$$S(t) = exp(-H(t)), \tag{5}$$

where $H(t) = \int_0^t h(u)du$ represents the *Cumulative Hazard Function* (CHF) (Lee and Wang 2003).

2.3 Taxonomy of Survival Analysis Methods

Broadly speaking, survival analysis methods can be classified into two main categories: statistical methods and machine learning based methods. Statistical methods share a common goal with machine learning methods in that both are expected to make predictions of the survival time and estimate the survival probability at the estimated survival time. However, the former focus more on characterizing both the distributions of the event times and the statistical properties of the parameter estimation by estimating the survival curves, while the latter focus primarily on the prediction of event occurrence at a given time by combining the power of traditional survival analysis methods with various machine learning techniques. Machine learning methods are usually applied to high-dimensional problems, while statistical methods are generally developed to handle low-dimensional data. Machine learning methods for survival analysis offer more effective algorithms because of their ability to analyze survival problems using both statistical methods and machine learning methods, thus taking advantage of recent developments in machine learning and optimization to learn the dependencies between covariates and survival times in different ways.

Depending on the assumptions made and the way parameters are used in the model, the traditional statistical methods can be subdivided into three categories: (i) non-parametric models, (ii) semi-parametric models, and (iii) parametric models. Machine learning algorithms such as survival trees, Bayesian methods, neural networks, and support vector machines, all of which have become increasingly popular in recent years, are included under a separate branch. Several advanced machine learning methods, including ensemble learning, active learning, transfer learning, and multitask learning methods, are also included in this category. The overall taxonomy also includes several research topics related to survival analysis such as complex events, data transformation, and early prediction. A complete taxonomy of these survival analysis methods is shown in Figure 3.

3 TRADITIONAL STATISTICAL METHODS

In this section, we introduce three different types of statistical methods that are commonly used to estimate the survival/hazard functions: non-parametric, semi-parametric, and parametric methods. Table 2 shows both the advantages and disadvantages of each type based on theoretical and experimental analysis and lists the specific methods included. The detailed characteristics of each statistical survival algorithm are summarized in Supplemental Table 1, including the assumptions made in each model, the optimization techniques used, and its computational complexity.

Non-parametric methods are more efficient when there is no underlying distribution for the event time or the proportional hazard assumption does not hold. In non-parametric methods, an empirical estimate of the survival function is obtained using the Kaplan-Meier (KM) method, Nelson-Aalen (NA) estimator, or Life-Table (LT) method. More generally, any KM estimator for the survival probability at the specified survival time will be a product of the same estimate up to the previous time and the observed survival rate for that given time. Thus, the KM method is also sometimes referred to as a *product-limit method* (Kaplan and Meier 1958; Lee and Wang 2003). The

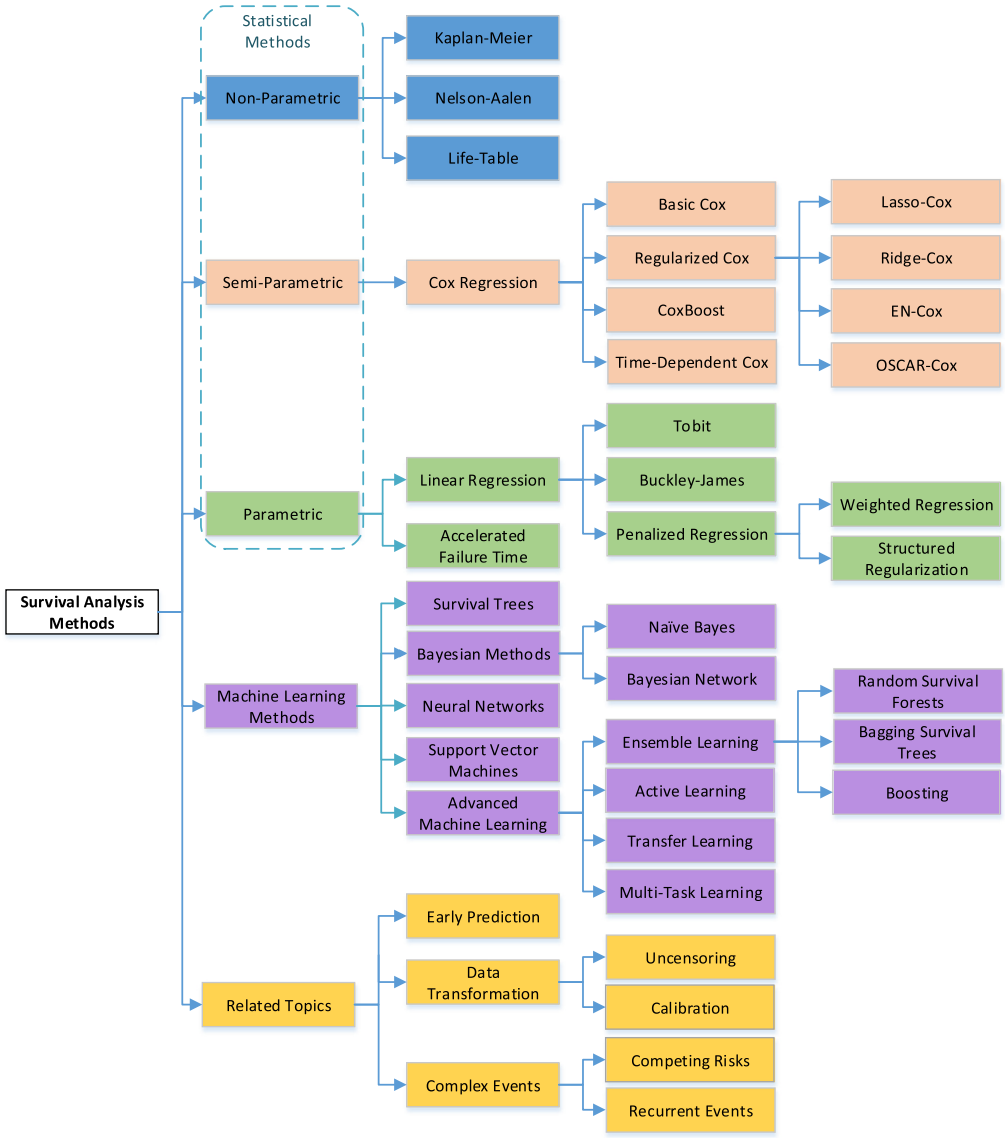

Fig. 3. Taxonomy of the methods developed for survival analysis.

NA method is an estimator based on modern counting process techniques (Andersen et al. 2012), while the LT method (Cutler and Ederer 1958) is the application of the KM method to the interval grouped survival data.

In the semi-parametric category, the Cox model is the most commonly used regression analysis approach for survival data. This differs significantly from other methods since it is built on the proportional hazards assumption and employs partial likelihood for the parameter estimation. Cox regression is described as a semi-parametric method since the distribution of the outcome remains unknown even if it is based on a parametric regression model. Several useful variants of the basic Cox model, such as the penalized Cox models (Zhang and Lu 2007), the CoxBoost algorithm (Binder

Table 2. Summary of Different Types of Statistical Methods for Survival Analysis

Type	Advantages	Disadvantages	Specific methods
Non-parametric	More efficient when no suitable theoretical distributions are known.	Difficult to interpret; yields inaccurate estimates.	Kaplan-Meier Nelson-Aalen Life-Table
Semi-parametric	Knowledge of the underlying distribution of survival times is not required.	The distribution of the outcome is unknown; not easy to interpret.	Cox model Regularized Cox CoxBoost Time-Dependent Cox
Parametric	Easy to interpret, more efficient and accurate when the survival times follow a particular distribution.	When the distribution assumption is violated, it may be inconsistent and can give suboptimal results.	Tobit Buckley-James Penalized regression Accelerated Failure Time

and Schumacher 2008) and the Time-Dependent Cox (TD-Cox) model (Kleinbaum and Klein 2006) have also been proposed in the literature.

Parametric methods are more efficient and accurate for estimation when the time to the event of interest follows a particular distribution that can be specified in terms of certain parameters. It is relatively easy to estimate the time to the event of interest with parametric models, but it becomes awkward or even impossible to do so with the Cox model (Allison 2010). Linear regression is one of the main parametric survival methods, while the Tobit model, Buckley-James regression model, and penalized regression are the most commonly used linear models for survival analysis. Other parametric models such as Accelerated Failure Time (AFT), which models the survival time as a function of covariates (Kleinbaum and Klein 2006), are also widely used. These will be described in turn in this section.

3.1 Non-parametric Models

Among all the functions options, the survival function or its graphical presentation is the most widely used. In 1958, Kaplan and Meier Kaplan and Meier (1958) developed the Kaplan-Meier (KM) Curve , which is also known as the Product-Limit (PL) estimator, to estimate the survival function using the actual length of the observed time. This method is the most widely used for estimating the survival function. Let $T_1 < T_2 < \cdots < T_K$ be a set of distinct ordered event times observed for $N(K \leq N)$ instances. In addition to these event times, there are also censored times for instances whose event times are not observed. For a specific event time T_j $(j = 1, 2, \ldots, K)$, the number of observed events is $d_j \geq 1$, and r_j instances will be considered to be "at risk" since their event time or censored time is greater than or equal to T_j. Note that we cannot simply consider r_j as the difference between r_{j-1} and d_{j-1} due to the censoring. The correct way to obtain r_j is $r_j = r_{j-1} - d_{j-1} - c_{j-1}$, where c_{j-1} represents the number of censored instances during the time period between T_{j-1} and T_j. This means that the conditional probability of surviving beyond time T_j can now be defined as:

$$p(T_j) = \frac{r_j - d_j}{r_j}. \tag{6}$$

Based on this conditional probability, the product-limit estimate of survival function $S(t) = P(T \geq t)$ is given as follows:

$$\hat{S}(t) = \prod_{j:T_j<t} p(T_j) = \prod_{j:T_j<t} \left(1 - \frac{d_j}{r_j}\right). \tag{7}$$

However, if the subjects in the data are grouped into some interval periods according to the time, if the number of subjects is very large, or if the study covers a large population, the LT analysis (Cutler and Ederer 1958) will be a more convenient method. Unlike the KM and LT methods, the NA estimator (Nelson 1972; Aalen 1978) estimates the cumulative hazard function for censored data based on a counting process approach. More details about LT and NA can be found under the supplementary material of this article. It is important to bear in mind that when the time to the event of interest follows a specific distribution, non-parametric methods are less efficient compared to parametric methods.

3.2 Semi-Parametric Models

As a hybrid of the parametric and non-parametric approaches, semi-parametric models can obtain more consistent estimates under a broader range of conditions than parametric models, and more precise estimates than non-parametric methods (Powell 1994). The Cox model (David 1972) is the most commonly used survival analysis method in this category. Unlike parametric methods, no knowledge of the underlying distribution of time to the event of interest is required, but the attributes are assumed to have an exponential influence on the outcome. In this section, we discuss the model in more detail and then describe different variants and extensions of the basic Cox model, such as regularized Cox models, CoxBoost, and the Time-Dependent Cox.

3.2.1 The Basic Cox Model. For a given instance i, represented by a triplet (X_i, y_i, δ_i), the hazard function $h(t, X_i)$ in the Cox model follows the proportional hazards assumption given by

$$h(t, X_i) = h_0(t)exp(X_i\beta), \tag{8}$$

for $i = 1, 2, \ldots, N$, where the *baseline hazard function*, $h_0(t)$, can be an arbitrary non-negative function of time; $X_i = (x_{i1}, x_{i2}, \ldots, x_{iP})$ is the corresponding covariate vector for instance i; and $\beta^T = (\beta_1, \beta_2, \ldots, \beta_P)$ is the coefficient vector. The Cox model is a semi-parametric algorithm since the baseline hazard function, $h_0(t)$, is unspecified. For any two instances X_1 and X_2, the hazard ratio is given by

$$\frac{h(t, X_1)}{h(t, X_2)} = \frac{h_0(t)exp(X_1\beta)}{h_0(t)exp(X_2\beta)} = exp[(X_1 - X_2)\beta]. \tag{9}$$

which means that the hazard ratio is independent of the baseline hazard function. The basic Cox model is a proportional hazards model since the hazard ratio is a constant and all the subjects share the same baseline hazard function. Based on this assumption, the survival function can be computed as follows:

$$S(t) = exp(-H_0(t)exp(X\beta)) = S_0(t)^{exp(X\beta)}, \tag{10}$$

where $H_0(t)$ is the *cumulative baseline hazard function*, and $S_0(t) = exp(-H_0(t))$ represents the baseline survival function. Breslow's estimator (Breslow 1972) is the most widely used method for estimating $H_0(t)$, which is given by

$$\hat{H}_0(t) = \sum_{t_i \leq t} \hat{h}_0(t_i), \tag{11}$$

where $\hat{h}_0(t_i) = 1/\sum_{j \in R_i} e^{X_j\beta}$ if t_i is an event time, otherwise $\hat{h}_0(t_i) = 0$. Here, R_i represents the set of subjects that are at risk at time t_i.

Because the baseline hazard function $h_0(t)$ in the basic Cox model is not specified, it is not possible to fit this model using the standard likelihood function. The hazard function $h_0(t)$ is effectively a nuisance function, while the coefficients β are the parameters of interest in the model. To estimate these coefficients, Cox proposed a partial likelihood (David 1972, 1975) which depends only on the parameter of interest β and is free of the nuisance parameters. The hazard function refers

to the probability that an instance with covariate X fails at time t; the condition that it survives until time t can be expressed by $h(t, X)dt$ with $dt \to 0$. Let J ($J \leq N$) be the total number of events of interest that occurred during the observation period for N instances, and $T_1 < T_2 < \cdots < T_J$ is the distinct ordered time to the event of interest. Without considering ties, let X_j be the corresponding covariate vector for the subject who fails at T_j, and let R_j be the set of risk subjects at T_j. Thus, conditional on the fact that the event occurs at T_j, the individual probability corresponding to covariate X_j can be formulated as follows:

$$\frac{h(T_j, X_j)dt}{\sum_{i \in R_j} h(T_j, X_i)dt}, \tag{12}$$

and the partial likelihood is the product of the probability of each subject. With reference to the Cox assumption and the presence of censoring, the partial likelihood is defined as follows:

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{exp(X_j\beta)}{\sum_{i \in R_j} exp(X_i\beta)} \right]^{\delta_j}. \tag{13}$$

Note that here $j = 1, 2, \ldots, N$. If $\delta_j = 1$, the jth term in the product is the conditional probability; otherwise, when $\delta_j = 0$, the corresponding term is 1, which means that the term will not have any effect on the final product. The coefficient vector $\hat{\beta}$ is estimated by either maximizing this partial likelihood or, equivalently, minimizing the negative *log-partial likelihood* for improving efficiency:

$$LL(\beta) = -\sum_{j=1}^{N} \delta_j \{X_j\beta - log[\sum_{i \in R_j} exp(X_i\beta)]\}. \tag{14}$$

The Maximum Partial Likelihood Estimator (MPLE) (David 1972; Lee and Wang 2003) can be used along with the numerical Newton-Raphson method (Kelley 1999) to iteratively find an estimator $\hat{\beta}$ which minimizes $LL(\beta)$ with time complexity $O(NP^2)$.

3.2.2 Regularized Cox models. With the continuing development of data collection and detection techniques, most real-world domains tend to accumulate high-dimensional data. In some cases, the number of variables (P) in the given data will almost equal, or even exceed, the number of instances (N). It is therefore challenging to build a good prediction model that takes into account all the available features, and, in some cases, the model may provide inaccurate results because of the overfitting problem (van Houwelingen and Putter 2011). This encourages the use of sparsity norms to select vital features in high-dimensional data under the assumption that most of the features are not significant (Friedman et al. 2001). In order to identify the features that are most relevant to the outcome variable among what can be tens of thousands of features, a number of different penalty functions, including lasso, group lasso, fused lasso, and graph lasso, have been developed for prediction models using sparse learning methods. The family of ℓ-norm penalty functions $\ell_\gamma : \mathbb{R}^\gamma \to \mathbb{R}$, which take the form $\ell_\gamma(\beta) = \| \beta \|_\gamma = (\sum_{i=1}^{P} \|\beta_i\|^\gamma)^{\frac{1}{\gamma}}, \gamma > 0$ are the most commonly used of these penalty functions. The smaller the value of γ, the sparser the solution, but when $0 \leq \gamma < 1$, the penalty is nonconvex, which makes the optimization problem more challenging to solve. The regularizers for the most commonly used regularized Cox models are summarized in Table 3.

Lasso-Cox: Lasso (Tibshirani 1996) is an ℓ_1-norm regularizer that performs feature selection and estimates the regression coefficients simultaneously. In Tibshirani (1997), the ℓ_1-norm penalty was incorporated into the log-partial likelihood shown in Equation (14) to obtain the Lasso-Cox algorithm, which inherits the properties of ℓ_1-norm in feature selection.

Table 3. Regularizers Used in Variants of the Cox Model

Regularized Cox models	Regularizers
Lasso-Cox	$\lambda \sum_{p=1}^{P} \lvert\beta_p\rvert$
Ridge-Cox	$\frac{\lambda}{2} \sum_{p=1}^{P} \beta_p^2$
EN-Cox	$\lambda[\alpha \sum_{p=1}^{P} \lvert\beta_p\rvert + \frac{1}{2}(1-\alpha) \sum_{p=1}^{P} \beta_p^2]$
OSCAR-Cox	$\lambda_1 \parallel \beta \parallel_1 + \lambda_2 \parallel T\beta \parallel_1$

There are also several useful extensions of the Lasso-Cox method. The adaptive Lasso-Cox (Zhang and Lu 2007) is based on a penalized partial likelihood with adaptively weighted ℓ_1 penalties $\lambda \sum_{j=1}^{P} \tau_j |\beta_j|$ on the regression coefficients, with small weights τ_j assigned for large coefficients and large weights for small coefficients. In the fused Lasso-Cox (Tibshirani et al. 2005), the coefficients and their successive differences are penalized using the ℓ_1-norm, while in the graphical Lasso-Cox (Friedman et al. 2008), the sparse graphs are estimated using the coordinate descent method by applying an ℓ_1-penalty to the inverse covariance matrix. These extensions then solve survival problems in a similar way to the regular Lasso-Cox model, differing only in incorporating these different ℓ_1 penalties.

Ridge-Cox: The Ridge regression was originally proposed by Hoerl and Kennard Hoerl and Kennard (1970) and was successfully used in the context of Cox regression by Verweij et al. (Verweij and Van Houwelingen 1994). It incorporates an ℓ_2-norm regularizer to select correlated features and shrink their values toward each other. The feature-based regularized Cox method (FEAR-Cox) (Vinzamuri and Reddy 2013) uses a feature-based non-negative valued regularizer $R(\beta) = |\beta|^T M|\beta|$ for the modified least squares formulation of the Cox regression; the cyclic coordinate descent method is used to solve this optimization problem, where $M \in \mathbb{R}^{P \times P}$ (P is the number of features) is a positive semi-definite matrix. Ridge-Cox is a special case of FEAR-Cox when M is the identity matrix.

EN-Cox: Elastic Net (EN), which combines the ℓ_1 and squared ℓ_2 penalties, has the potential to perform feature selection and deal with the correlation between the features simultaneously (Zou and Hastie 2005). The EN-Cox method was proposed by Simon et al. Simon et al. (2011), using the EN penalty term shown in Table 3 with $0 \leq \alpha \leq 1$ and introducing the log-partial likelihood function in Equation (14). Unlike Lasso-Cox, EN-Cox can select more than N features if $N \leq P$.

The Kernel Elastic Net (KEN) algorithm (Vinzamuri and Reddy 2013), which as the name suggests uses the concept of kernels, compensates for the drawbacks of the EN-Cox, which is only partially effective at dealing with the correlated features in survival data. KEN-Cox it builds a kernel similarity matrix for the feature space in order to incorporate pairwise feature similarity into the Cox model. The regularizer used in KEN-Cox is defined as $\lambda\alpha||\beta||_1 + \lambda(1-\alpha)\beta^T K\beta$, where K is a symmetric kernel matrix with $K_{ij} = exp(- \parallel x_i - x_j \parallel_2^2 / 2\sigma^2)(i, j = 1, \ldots, P)$ as its entries. Note that the equation for the KEN-Cox method includes both smooth and non-smooth ℓ_1 terms.

OSCAR-Cox: The modified graph Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) (Yang et al. 2012; Ye and Liu 2012) regularizer is incorporated in the basic Cox model as the OSCAR-Cox algorithm (Vinzamuri and Reddy 2013), which can perform variable selection for highly correlated features in the regression problem. The main advantage of using an OSCAR regularizer is that it tends to have equal coefficients for the features that relate to the outcome in similar ways. At the same time, it gains the advantages of both individual sparsity because of the ℓ_1 norm and group sparsity due to the ℓ_∞ norm. The regularizer used in the formulation of OSCAR-Cox is given in Table 3, where T is the sparse symmetric edge set matrix generated by building a graph structure that considers each feature as an individual node. In this way, a pairwise feature regularizer can be incorporated into the basic Cox regression framework.

Among the regularizers shown in Table 3, the parameters $\lambda \geq 0$ can be tuned to adjust the influence introduced by the regularizer term. The performance of these penalized estimators depends significantly on λ, and the optimal λ_{opt} can be chosen via cross-validation.

3.2.3 CoxBoost. While there are several algorithms (such as the penalized parameter estimation) that can be applied to enable sparse survival models to cope with high-dimensional data, none is applicable if mandatory covariates must be explicitly taken into consideration in the models. The CoxBoost (Binder and Schumacher 2008) approach has therefore been proposed to incorporate mandatory covariates into the final model. The CoxBoost method also aims at estimating the coefficients β in Equation (8), as in the Cox model. It considers a flexible set of candidate variables for updating each boosting step by employing an offset-based gradient boosting approach. This is the key difference that sets it apart from the regular gradient boosting approach, which either updates only one component of β in component-wise boosting or fits the gradient by using all the covariates in each step.

3.2.4 Time-Dependent (TD) Cox Model. The Cox regression model has also been adapted to handle time-dependent covariates, which are variables whose values may change with time t for a given instance. A time-dependent variable can typically be categorized into three types (Kleinbaum and Klein 2006): internal time-dependent variables, ancillary time-dependent variables, and defined time-dependent variables. Internal time-dependent variables can change depending on one or more internal characteristics or behaviors that are specific to the individual. In contrast, an ancillary time-dependent variable's value changes primarily due to the environment, and this may affect several individuals simultaneously. A defined variable, which takes the form of the product of a time-independent variable multiplied by a function of time, is used to analyze a time-independent predictor that does not satisfy the PH assumption in the Cox model. The most commonly used layout for a dataset in the time-dependent Cox model takes the form of a Counting Process (CP) (Kleinbaum and Klein 2006).

Given a survival analysis problem that involves both time-dependent and time-independent features, we can denote the variables at time t as $X(t) = (X_{\cdot 1}(t), X_{\cdot 2}(t), \ldots, X_{\cdot P_1}(t), X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot P_2})$, where P_1 and P_2 represent the number of time-dependent and time-independent variables, respectively, and $X_{\cdot j}(t)$ and $X_{\cdot i}$ represent the jth time-dependent feature and the ith time-independent feature, respectively. Then, by incorporating the time-dependent features into the basic Cox model given in Equation (8), the time-dependent Cox model can be formulated as:

$$h(t, X(t)) = h_0(t) exp\left[\sum_{j=1}^{P_1} \delta_j X_{\cdot j}(t) + \sum_{i=1}^{P_2} \beta_i X_{\cdot i}\right], \tag{15}$$

where δ_j and β_i represent the coefficients corresponding to the jth time-dependent variable and the ith time-independent variable, respectively. For the two sets of predictors at time t: $X(t) = (X_{\cdot 1}(t), X_{\cdot 2}(t), \ldots, X_{\cdot P_1}(t), X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot P_2})$ and $X^*(t) = (X^*_{\cdot 1}(t), X^*_{\cdot 2}(t), \ldots, X^*_{\cdot P_1}(t), X^*_{\cdot 1}, X^*_{\cdot 2}, \ldots, X^*_{\cdot P_2})$, the hazard ratio for the time-dependent Cox model can be computed as follows:

$$\hat{HR}(t) = \frac{\hat{h}(t, X^*(t))}{\hat{h}(t, X(t))} = exp\left[\sum_{j=1}^{P_1} \delta_j [X^*_{\cdot j}(t) - X_{\cdot j}(t)] + \sum_{i=1}^{P_2} \beta_i [X^*_{\cdot i} - X_{\cdot i}]\right]. \tag{16}$$

Since the first component in the exponent of Equation (16) is time-dependent, we can consider the hazard ratio in the TD-Cox model as a function of time t. This means that it does not satisfy the PH assumption mentioned in the standard Cox model. Note that the coefficient δ_j is in itself not time-dependent, although it does represent the overall effect of the jth time-dependent variable

at various survival time points. The likelihood function for the time-dependent Cox model can be constructed in the same manner and optimized with the same time complexity as the Cox model.

3.3 Parametric Models

Parametric censored regression models assume that the survival times or the logarithm of the survival times of all instances in the data follow a particular theoretical distribution (Lee and Wang 2003). These models offer important alternatives to the Cox-based semi-parametric models and are also widely used in many application domains, providing a simple, efficient, and effective way to predict the time to event of interest. Parametric survival models tend to produce survival estimates that are consistent with a theoretical survival distribution. The commonly used distributions in parametric censored regression models are normal, exponential, Weibull, logistic, log-logistic, and log-normal. If the survival times of all the instances in the dataset follow these distributions, the model is referred to as a *linear regression model*. If the logarithm of the survival times of all the instances follow these distributions, the problem can be analyzed using the accelerated failure time model, in which we assume that the variable can affect the time to the event of interest of an instance by some constant factor (Lee and Wang 2003). However, if no suitable theoretical distribution is known, non-parametric methods are more efficient.

The Maximum-Likelihood Estimation (MLE) method (Lee and Wang 2003) can be used to estimate the parameters for these models. Let us assume that the number of instances is N, composed of c censored observations and $(N - c)$ uncensored observations, and use $\beta = (\beta_1, \beta_2, \ldots, \beta_P)^T$ as a general notation to denote the set of all parameters (Li et al. 2016e). Then the death density function $f(t)$ and the survival function $S(t)$ of the survival time can be represented as $f(t, \beta)$ and $S(t, \beta)$, respectively. If a given instance i is censored, the actual survival time will not be available. However, we can conclude that the instance i did not experience the event of interest before the censoring time C_i, so the value of the survival function $S(C_i, \beta)$ will be a probability close to 1. In contrast, if the event occurs for instance i at T_i, then the death density function $f(T_i, \beta)$ will have a high probability value. Thus, we can denote $\prod_{\delta_i=1} f(T_i, \beta)$ as the joint probability of all the uncensored observations and $\prod_{\delta_i=0} S(T_i, \beta)$ as the joint probability of the c censored observations (Li et al. 2016e). Therefore, we can estimate the parameters β by optimizing the likelihood function of all N instances as follows:

$$L(\beta) = \prod_{\delta_i=1} f(T_i, \beta) \prod_{\delta_i=0} S(T_i, \beta). \tag{17}$$

Table 4 shows the death density function $f(t)$ and its corresponding survival function $S(t)$ and hazard function $h(t)$ for these commonly used distributions, which are discussed in more detail below.

Exponential Distribution: Among the parametric models used for survival analysis, the exponential model is the simplest and arguably the most important since it is characterized by a single parameter, the constant hazard rate, λ. In this case, the failure or death is assumed to be a random event that is independent of time. A larger value of λ indicates a higher risk and a shorter survival time period. Based on the survival function shown in Table 4, we can have $logS(t) = -\lambda t$, in which the relationship between the logarithm of the survival function and time t is linear, with λ as the slope. It is then straightforward to determine whether the time does indeed follow an exponential distribution by plotting $log\hat{S}(t)$ against time t (Lee and Wang 2003).

Weibull Distribution: The Weibull model, which is characterized by two parameters $\lambda > 0$ and $k > 0$, is the most widely used parametric distribution for survival problems. The shape of the hazard function is determined using the shape parameter k, which provides more flexibility

Table 4. Density, Survival and Hazard Functions for the Distributions Commonly Used for Parametric Methods in Survival Analysis

Distribution	PDF $f(t)$	Survival $S(t)$	Hazard $h(t)$
Exponential	$\lambda exp(-\lambda t)$	$exp(-\lambda t)$	λ
Weibull	$\lambda k t^{k-1} exp(-\lambda t^k)$	$exp(-\lambda t^k)$	$\lambda k t^{k-1}$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{\sigma(1+e^{-(t-\mu)/\sigma})}$
Log-logistic	$\frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}$	$\frac{1}{1+\lambda t^k}$	$\frac{\lambda k t^{k-1}}{1+\lambda t^k}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(t-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{t-\mu}{\sigma})$	$\frac{1}{\sqrt{2\pi}\sigma(1-\Phi((t-\mu)/\sigma))} exp(-\frac{(t-\mu)^2}{2\sigma^2})$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t} exp(-\frac{(log(t)-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{log(t)-\mu}{\sigma})$	$\frac{\frac{1}{\sqrt{2\pi}\sigma t} exp(-(log(t)-\mu)^2/2\sigma^2)}{1-\Phi(\frac{log(t)-\mu}{\sigma})}$

compared to the exponential model. If $k = 1$, the hazard will be a constant, and, in this case, the Weibull model will become an exponential model. If $k < 1$, the hazard function will be decreasing over time. The scaling of the hazard function is determined by the scaling parameter λ.

Logistic and Log-Logistic Distribution: In contrast to the Weibull model, the hazard functions for both the logistic and log-logistic models allow for non-monotonic behavior in the hazard function, which is shown in Table 4. The survival time T and the logarithm of the survival time $log(T)$ follow a logistic distribution in logistic and log-logistic models, respectively. For the logistic model, μ is the parameter used to determine the location of the function, while σ is the scale parameter; whereas in the log-logistic model, the parameter $k > 0$ is the shape parameter. If $k \leq 1$, the hazard function is decreasing over time. However, if $k > 1$, the hazard function will initially increase over time to a maximum value and then decrease, which means that the hazard function is unimodal if $k > 1$. Thus, a log-logistic distribution may be used to describe a monotonically decreasing hazard or a hazard that first increases and then decreases (Lee and Wang 2003).

Normal and Log-Normal Distribution: If the survival time T satisfies the condition that T or $log(T)$ is normally distributed with mean μ and variance σ^2, then T is normally or log-normally distributed. This distribution is again suitable for survival patterns with an initially increasing and then decreasing hazard rate.

Based on the framework given in Equation (17), we can examine these commonly used parametric methods more closely.

3.3.1 Linear Regression Models. In data analysis, the linear regression model, together with the least squares estimation method, represents one of the most commonly used approaches. However, it cannot be applied directly to solve survival analysis problems since the actual event times are missing for censored instances. A number of linear models (Miller and Halpern 1982; Buckley and James 1979; Wang et al. 2008; Li et al. 2016e), including Tobit and Buckley-James (BJ) regressions, have been proposed to handle censored instances in survival analysis. Strictly speaking, although a linear regression is a specific parametric censored regression, this method is fundamental for many different types of data analysis, and hence linear regression methods for censored data will be discussed se parately here.

Tobit Regression: The Tobit model (Tobin 1958) is one of the earliest attempts to extend a linear regression with a Gaussian distribution for data analysis with censored observations. In this model, a latent variable y^* is introduced and the assumption made is that this depends linearly on X via the parameter β as $y^* = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$, where ϵ is a normally distributed error term. Then, for the ith instance, the observable variable y_i will be y_i^* if $y_i^* > 0$; otherwise, it will be 0. This

means that if the latent variable is above zero, the observed variable equals the latent variable and is zero otherwise. Based on the latent variable, the parameters in the model can be estimated using an MLE method.

Buckley-James Regression: The Buckley-James (BJ) regression (Buckley and James 1979) estimates the survival time of censored instances in terms of the response value based on the KM estimation method, and then fits a linear (AFT) model by simultaneously considering the survival times of uncensored instances and the approximate survival times of the censored instances. To handle high-dimensional survival data, Wang et al. Wang et al. (2008) applied the elastic net regularizer in the BJ regression (EN-BJ).

Penalized Regression: Penalized regression methods (Kyung et al. 2010) are best known for their useful property of simultaneous variable selection and coefficient estimation. A penalized regression method can provide better prediction results in the presence of either multicollinear covariates or high dimensionality. Recently, these methods have received a great deal of attention in survival analysis. For example, a weighted linear regression model with different regularizers for high-dimensional censored data is an efficient way to handle censored data by assigning different weights to different instances (Li et al. 2016e). In addition, a structured regularization-based linear regression algorithm (Bach et al. 2012; Vinzamuri et al. 2017) for right-censored data can be used to infer the underlying structure of the survival data.

— **Weighted Regression:** The weighted regression method (Li et al. 2016b) can be used when the constant variance assumption about the errors in the ordinary least squares regression methods is violated (i.e., heteroscedasticity), which is different from a constant variance in the errors (i.e., homoscedasticity) in ordinary least squares regression methods. Instead of minimizing the residual sum of squares, the weighted regression method minimizes the weighted sum of squares $\sum_{i=1}^{n} w_i(y_i - X_i\beta)^2$; an ordinary least squares method is a special case of this where all the weights $w_i = 1$. A weighted regression method can be solved in the same manner as an ordinary linear least squares problem with a time complexity of $O(NP)$. The weighted regression method enables us to assign higher weights to instances that we want to emphasize or those where mistakes are especially costly. If we give selected samples high weights, the model will be pulled toward matching the data for those samples. This is very helpful for survival analysis as it allows us to place more emphasis on the instances whose information may contribute most to the model.
— **Structured Regularization:** The ability to effectively infer latent knowledge through tree-based hierarchies and graph-based relationships is crucial in survival analysis. This is also supported by the effectiveness of structured sparsity-based regularization methods in regression (Bach et al. 2012). The Structured regularization based LInear REgression algorithm for right Censored data (SLIREC) (Vinzamuri et al. 2017) infers the underlying structure of the survival data directly using a Sparse Inverse Covariance Estimation (SICE) method and then applies this structural knowledge to guide the base linear regression model. The structured approach is more robust than standard statistical or Cox-based methods since it can automatically adapt to different distributions of events and censored instances.

3.3.2 Accelerated Failure Time (AFT) Model. In the parametric censored regression methods discussed previously, we assume that the survival time of all instances in the given data follows a specific distribution and that the relationship between either the survival time or the logarithm of the survival time and the features is linear. If the relationship between the logarithm of the survival time T and the covariates is linear in nature, it is also referred to as an AFT model (Kalbfleisch and Prentice 2011). Thus, we consider these regression methods to be generalized linear models.

The AFT model assumes that the relationship between the logarithm of the survival time T and the covariates is linear and can be written in the following form.

$$ln(T) = X\beta + \sigma\epsilon, \tag{18}$$

where X is the covariate matrix, β represents the coefficient vector, $\sigma(\sigma > 0)$ denotes an unknown scale parameter, and ϵ is an error variable that follows a similar distribution to $ln(T)$. Typically, we make a parametric assumption about ϵ, which can follow any of the distributions given in Table 4. In this case, the survival is dependent on both the covariate and the underlying distribution. This means that the only distinction between an AFT model and regular linear methods would be the inclusion of censored information in the survival analysis problem. The AFT model is additive with respect to $ln(T)$ but multiplicative with respect to T, and is written in the form $T = e^{X\beta}e^{\sigma\epsilon}$.

4 MACHINE LEARNING METHODS

Over the past several years, due to the advantages of machine learning techniques, such as their ability to model nonlinear relationships and the quality of their overall predictions, significant successes have been achieved in a wide range of practical domains. In survival analysis, the main challenge facing machine learning methods is the difficulty of dealing appropriately with censored information and the time estimation of the model. Machine learning is effective when there are a large number of instances in a reasonable dimensional feature space, but this is not the case for certain problems in survival analysis (Zupan et al. 2000). This section provides a comprehensive review of commonly used machine learning methods in survival analysis. Similar to Supplemental Table 1, Supplemental Table 2 summarizes the detailed characteristics of each machine learning method for survival analysis. In the table, the term K involved in the complexity column represents the number of unique time points in the survival analysis problem.

4.1 Survival Trees

Survival trees are classification and regression trees that are specifically tailored to handle censored data. The basic intuition behind tree models is that data are recursively partitioned based on a particular splitting criterion, and objects that are similar to each other based on the event of interest will be placed in the same node. Although the earliest attempt to use a tree structure for survival data was reported in Ciampi et al. (1981), Gordon and Olshen (1985) was the first paper to explicitly discuss the creation of survival trees.

The primary difference between a survival tree and the standard decision tree is in the choice of splitting criterion. The decision tree method performs recursive partitioning on the data by setting a threshold for each feature, but it can neither consider the interactions between the features nor the censored information in the model (Safavian and Landgrebe 1991). The splitting criteria used for survival trees can be grouped into two categories; namely, those that minimize within-node homogeneity and those that maximize between-node heterogeneity. The first of these approaches minimizes the loss function using the within-node homogeneity criterion. For example, Gordon and Olshen (1985) measured the homogeneity and Hellinger distances between the estimated distribution functions using the Wasserstein metric, while an exponential log-likelihood function was employed in Davis and Anderson (1989) for recursive partitioning based on the sum of residuals from the Cox model and Leblanc and Crowley LeBlanc and Crowley (1992) measured the node deviance based on the first step of a full likelihood estimation procedure. For the second type of splitting criteria, Ciampi et al. Ciampi et al. (1986) employed log-rank test statistics for between-node heterogeneity measures. The same group (Ciampi et al. 1987) then went on to propose the use of a likelihood ratio statistic to measure the dissimilarity between two nodes. Based on the Tarone-Ware class of two-sample statistics, Segal Segal (1988) introduced a procedure to measure

the between-node dissimilarity. The main advantage a survival tree has over a standard decision tree is its ability to handle censored data using the tree structure.

Another important aspect of building a survival tree is the selection of the final tree. Procedures such as backward or forward selection can be followed for choosing the optimal tree structure (Bou-Hamad et al. 2011). However, an ensemble of trees (described in Section 4.5) can avoid the problem of final tree selection and provide a better performance than is possible with a single tree.

4.2 Bayesian Methods

The Bayes theorem is one of the most fundamental principles in probability theory and mathematical statistics. It provides a link between the *posterior probability* and the *prior probability* so users can see how probability values change before and after accounting for a certain event. There are two models, namely Naïve Bayes (NB) and Bayesian networks (BN) (Friedman et al. 1997), both of which provide the probability of an event of interest as their output and are commonly used for clinical prediction (Kononenko 1993; Pepe 2003; Zupan et al. 2000). The experimental results of applying Bayesian methods to survival data show that these methods have excellent interpretability and uncertainty reasoning (Raftery et al. 1995).

NB, a well-known probabilistic method in machine learning, is one of the simplest yet most effective prediction algorithms. Its uses are legion; for example, Bellazzi and Zupan (2008) built a naïve Bayesian classifier to make predictions in clinical medicine by estimating various probabilities from the data, while Fard et al. (2016) effectively integrated Bayesian methods with an AFT model by extrapolating the prior event probability to perform early-stage prediction on survival data for future time points. One drawback of the NB method, however, is that it assumes independence between all the features, which may not be true for many problems in survival analysis.

A BN, in which the features can be related to each other at a number of different levels, can graphically represent a theoretical distribution over a set of variables. BNs can visually represent all the relationships between the variables, making it easily interpretable for the end user. It can also acquire knowledge by estimating the network structures and parameters from a given dataset. Lisboa et al. (2003) proposed a Bayesian neural network framework to perform model selection for survival data using an automatic relevance determination procedure MacKay (1995). (Raftery 1995) proposed a Bayesian model averaging procedure for Cox proportional hazards models which was then used to evaluate the Bayes factors in the problem. More recently, Fard et al. (2016) proposed a novel framework that combines the power of BN representations with the AFT model by extrapolating the prior probabilities to future time points. The time complexity of these Bayesian approaches depends primarily on the types of Bayesian method used in the models.

4.3 Artificial Neural Networks

Inspired by biological neural systems, Frank Rosenblatt published the first paper (Rosenblatt 1958) on Artificial Neural Networks (ANN) in 1958. Here, the simple artificial nodes that he dubbed "neurons" are connected based on a weighted link to form a network which simulates a biological neural network. A neuron in this context is a computing element that consists of sets of adaptive weights and generates the output based on a certain kind of *activation function*. ANNs have been widely used in survival analysis. Three main methods are proposed in the literature for applying neural network methods to address survival analysis problems.

(1) Neural network survival analysis has been employed to predict the survival time of a subject directly from the inputs provided.
(2) Faraggi and Simon (1995) extended the Cox PH model to the nonlinear ANN predictor and suggested using this to fit a neural network with a linear output layer and a single logistic hidden layer. Mariani et al. (1997) used both the standard Cox model and the neural

network method proposed in Faraggi and Simon (1995) to assess prognostic factors for the recurrence of breast cancer. Although these extensions of the Cox model preserved most of the advantages of a typical PH model, they were still not optimal for modeling the baseline variation (Baesens et al. 2005).

(3) Many approaches (Liestbl et al. 1994; Biganzoli et al. 1998; Brown et al. 1997; Lisboa et al. 2003) take the survival status of a subject, which can be represented by the survival or hazard probability, as the output of the neural network. For example, Biganzoli et al. (1998) applied the Partial Logistic ANN (PLANN) method to analyze the relationship between features and survival times in order to boost the prediction ability of the model. The same group used feed-forward neural networks to obtain a more flexible nonlinear model by considering the censored information in the data using a generalization of both continuous and discrete time models (Biganzoli et al. 1998). Lisboa et al. (2003) extended PLANN to a Bayesian neural framework with covariate-specific regularization to perform model selection using automatic relevance determination (MacKay 1995).

These approaches indicate that neural networks have an excellent ability to handle the censored data in survival analysis. Recently, deep learning methods have gained considerable attention to tackle the survival analysis problems in various application domains. For example, in the healthcare domain, deep correlational survival models and Convolutional Neural Networks (CNNs) are used (Yao et al. 2017) to efficiently learn the complex interactions for multiple modalities of patient data. Several deep survival analysis approaches (Katzman et al. 2016; Ranganath et al. 2016) are also proposed to assist physicians with their clinical decisions on patients by estimating their risk of diseases and providing personalized treatment recommendations. In addition, Recurrent Neural Network (RNN)-based approaches have been successfully combined with survival analysis for studying recurring events, particularly in the context of user behavior modeling applications (Jing and Smola 2017; Yang et al. 2018).

4.4 Support Vector Machines

Support Vector Machines (SVM), a very successful supervised learning approach, are used mostly for classification and can also be modified for regression problems (Smola and Schölkopf 2004). They have also been successfully adapted for use in survival analysis problems.

A naive approach is to consider only those instances that actually have events in the Support Vector Regression (SVR), where the ϵ-insensitive loss function $f(X_i) = max(0, |f(X_i) - y_i| - \epsilon)$ is minimized with a regularizer (Smola and Schölkopf 1998). However, the main disadvantage of this approach is that the order information included in the censored instances will then be completely ignored (Shivaswamy et al. 2007). Another possible way to handle censored data is to apply support vector classification using the constraint classification approach (Har-Peled et al. 2002). This imposes constraints on the SVM formulation for two comparable instances in order to maintain the required order. However, the computational complexity for this algorithm is quadratic with respect to the number of instances, which for large datasets is clearly impractical. In addition, it only examines the ordering among the instances and ignores the actual values of the outputs.

To counter this, Khan and Zubek (2008) proposed the use of SVR for Censored Data (SVRc), as this takes advantage of the standard SVR and also adapts it for censored cases by using an updated asymmetric loss function. In this case, it considers both the uncensored and censored instances in the model. Van et al. (2007) studied a learning machine designed for the predictive modeling of independently right-censored survival data by introducing a health index, which serves as a proxy between the instance's covariates and the outcome. The same group (Van et al. 2011) then went on to introduce an SVR-based approach that combines the ranking and regression methods in the context of survival analysis. On average, the time complexity of these methods is $O(N^3)$, which follows the time complexity of the standard SVM.

The Relevance Vector Machine (RVM) proposed by others (Widodo and Yang 2011; Kiaee et al. 2016) obtains parsimonious estimates for regression and probabilistic problems using Bayesian inference utilizing the same formulation as SVM but adding a probabilistic classification. RVM adopts a Bayesian approach by considering the prior over the weights controlled by various parameters. Each of these parameters corresponds to a weight, the most probable value of which can be estimated iteratively based on the data. The Bayesian representation of the RVM can avoid these parameters in SVM as optimization methods based on cross-validation are usually used. However, it is also possible for RVMs to converge to a local minimum since an EM algorithm is used to learn the parameters. This differs from the regular Sequential Minimal Optimization (SMO) algorithm used in SVM, which guarantees convergence to a global minimum.

4.5 Advanced Machine Learning Approaches

Over the past few years, a number of advanced machine learning methods have been developed to deal with and make predictions based on censored data. These methods offer unique advantages for survival data compared to other methods described so far.

4.5.1 Ensemble Learning. Ensemble learning methods (Dietterich 2000) generate a committee of classifiers and then predict the class labels for new data points as they arrive by taking a weighted vote among the prediction results from all these classifiers. It is often possible to construct good ensembles and obtain a better approximation of the unknown function by varying the initial points, especially in the presence of insufficient data. To overcome the instability of a single method, bagging (Breiman 1996) and random forests (Breiman 2001) are generally used to perform the ensemble-based model building. Such ensemble models have been successfully adapted to survival analysis whose time complexity mainly follows that of the base-learners.

Bagging Survival Trees: Bagging is one of the oldest and most commonly used ensemble methods, typically reducing the variance of the base models that are used. In bagging survival trees, the aggregated survival function can be calculated by averaging the predictions made by a single survival tree instead of taking a majority vote (Hothorn et al. 2004). There are three main steps in this method: (i) draw B booststrap samples from the given data; (ii) for each bootstrap sample, build a survival tree and ensure that for all the terminal nodes, the number of events is greater than or equal to the threshold d; and (iii) by averaging the leaf nodes' predictions, calculate the bootstrap aggregated survival function. For each leaf node, the survival function is estimated using the KM estimator, and all the individuals within the same node are assumed to have the same survival function.

Random Survival Forests: The random forest is an ensemble method specifically developed to make predictions using tree structured models (Breiman 2001). It is based on a framework similar to bagging, with the main difference between them being that at a certain node, rather than using all the attributes, the random forest uses only a random subset of the residual attributes to select the attributes based on the splitting criterion. This randomization has been shown to reduce the correlation among the trees and thus improve the prediction performance.

The Random Survival Forest (RSF) (Ishwaran et al. 2008) extended Breiman's random forest method by using a forest of survival trees for prediction. Here there are four main steps: (i) draw B bootstrap samples randomly from the given dataset. This is also referred to as Out-Of-Bag (OOB) data because around 37% of the data is excluded in each sample. (ii) For each sample, build a survival tree by randomly selecting features and then split the node using the candidate features to maximize the survival difference between the child nodes. (iii) Build the full-size tree based on the constraint that the terminal node must be greater than or equal to a specific unique death. And (iv) using the non-parametric NA estimator, calculate the ensemble Cumulative Hazard Function

(CHF) of the OOB data by taking the average of the CHF of each tree. Ishwaran et al. (2011) also provides an effective way to apply RSF for high-dimensional survival analysis problems by regularizing forests.

Boosting: The boosting algorithm is one of the most widely used ensemble methods and is designed to combine base learners into a weighted sum that represents the final output of a strong learner. This algorithm iteratively fits a set of appropriately defined residuals based on the gradient descent algorithm (Hothorn et al. 2006; Bühlmann and Hothorn 2007). Hothorn et al. (2006) extended the gradient boosting algorithm to minimize the weighted risk function $\hat{\beta}_{\tilde{U},X} = arg\,min_{\beta} \sum_{i=1}^{N} w_i(\tilde{U}_i - h(X_i|\beta))$, where $\tilde{U}$ is a pseudo-response variable with $\tilde{U}_i = -\frac{\partial L(y_i,\phi)}{\partial \phi}|_{\phi=\hat{f}_m(X_i)}$; β is a vector of parameters; and $h(\cdot|\beta_{U,X})$ is the prediction made by regressing U using a base learner. The steps taken to optimize this problem are as follows: (i) initialize $\tilde{U}_i = y_i$ $(i = 1, \ldots, N)$, $m = 0$ and $\hat{f}_0(\cdot|\hat{\beta}_{\tilde{U},X})$; fix the number of iterations $M(M > 1)$; (ii) fit $h(\cdot|\hat{\beta}_{U,X})$ after updating the residuals $\tilde{U}_i$ $(i = 1, \ldots, N)$; (iii) iteratively update $\hat{f}_{m+1}(\cdot) = \hat{f}_m(\cdot) + \upsilon h(\cdot|\hat{\beta}_{U,X})$, where $0 < \upsilon \leq 1$ represents the step size; and (iv) repeat the procedures in steps (ii) and (iii) until $m = M$.

4.5.2 Active Learning. Active learning based on data containing censored observations can be very helpful for survival analysis since it allows the opinions of an expert in the domain to be incorporated into the models. Active learning mechanisms allow the survival model to select a subset of subjects by learning from a limited set of labeled subjects first and then querying the expert to confirm a label for the survival status before considering including new data in the training set. The feedback from the expert is particularly useful for improving the model in real-world application domains (Vinzamuri et al. 2014). The goal of active learning for survival analysis problems is to build a survival regression model by utilizing the censored instances completely, without deleting or modifying the instances. Vinzamuri et al. (2014) proposed an active regularized Cox regression (ARC) algorithm based on a discriminative gradient sampling strategy obtained by integrating the active learning method with the Cox model. The ARC framework is an iteration-based algorithm composed of three main steps: (i) build a regularized Cox regression using the training data, (ii) apply the model obtained in (i) to all the instances in the unlabeled pool, and (iii) update the training data and the unlabeled pool, then select the instance whose influence on the model is the highest and label it before running the next iteration. One of the main advantages of the ARC framework is that it can identify instances and obtain feedback about event labeling from the domain expert. The time complexity of the ARC algorithm is $O(NPK)$, where K represents the number of unique time points in the survival problem.

4.5.3 Transfer Learning. Collecting labeled information for survival problems is very time-consuming as it is necessary to wait for the event to occur in a sufficient number of training instances to build robust models. A naive solution for this insufficient data problem is to merely integrate the data from related tasks into a consolidated form and build prediction models on this integrated dataset. However, such approaches often do not perform well because the target task (for which the predictions are to be made) will be overwhelmed by auxiliary data with different distributions. In such scenarios, knowledge transfer between related tasks will usually produce much better results. Transfer learning methods have been extensively studied to solve standard regression and classification problems (Pan and Yang 2010). Recently, Li and coworkers (Li et al. 2016c) proposed the use of a regularized Cox PH model named Transfer-Cox to improve the prediction performance of the Cox model in the target domain through knowledge transfer from the source domain in the context of survival models built on multiple high-dimensional datasets. The Transfer-Cox model employs $\ell_{2,1}$-norm to penalize the sum of the loss functions (negative partial

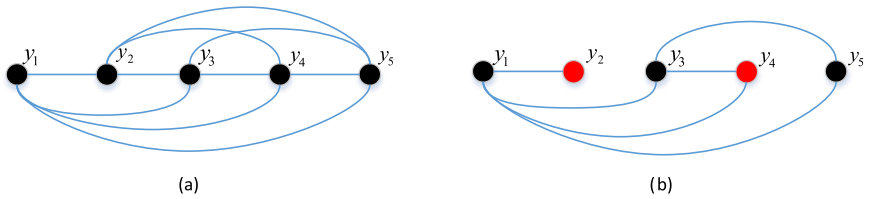

Fig. 4. Illustration of the ranking constraints in survival data for C-index calculations ($y_1 < y_2 < y_3 < y_4 < y_5$). Here, black X's indicate observed events and red circles indicate censored observations. (a) No censored data and (b) With censored data.

log-likelihood) for both source and target domains. Thus, the model, which has time complexity $O(NP)$, will not only select important features but will also learn shared representations across source and target domains to improve the model performance on the target task.

4.5.4 Multitask Learning. Li et al. (2016d) reformulated the survival time prediction problem as a multitask learning problem. In survival data, the outcome labeling matrix is necessarily incomplete since the event label of each censored instance is unavailable after its corresponding censoring time. This means that it is not possible to handle censored information using the standard multitask learning methods. To address this problem, the Multitask Learning Model for Survival Analysis (MTLSA) translates the original event labels into an $N \times K$ indicator matrix I, where $K = max(y_i)$ $(\forall i = 1, \ldots, N)$ is the maximum follow-up time for all the instances in the dataset. Element I_{ij} $(i = 1, \ldots, N; j = 1, \ldots, K)$ of the indicator matrix will be 1 if the event occurred before time y_j for instance i; otherwise it will be 0. One of the major advantages of the MTLSA approach is that it can capture the dependency between the outcomes at various time points by using a shared representation across the related tasks in the transformation, which will reduce the prediction error for each task. In addition, the model can simultaneously learn from both uncensored and censored instances based on the indicator matrix. One important characteristic of non-recurring events (namely, that once the event occurs it will not occur again) is encoded via the *non-negative non-increasing list* structure constraint. In the MTLSA algorithm, the $\ell_{2,1}$-norm penalty is employed to learn a shared representation, with time complexity $O(NPK)$, across related tasks and hence compute the relatedness between the individual models built for various unique event time points.

5 PERFORMANCE EVALUATION METRICS

Due to the presence of censoring in survival data, standard evaluation metrics for regression such as root mean squared error and R^2 are not suitable for measuring the prediction performance (Heagerty and Zheng 2005). Instead, the model performance in survival analysis needs to be measured using more specialized evaluation metrics.

5.1 C-index

In survival analysis, a common way to evaluate a model is to consider the relative risk of an event occurring for different instances rather than the absolute survival times for each instance. This can be done by computing the concordance probability or the concordance index (C-index) (Harrell et al. 1984, 1982; Pencina and D'Agostino 2004). The survival times for two instances can be ordered for two scenarios: (i) both are uncensored, and (ii) the observed event time of the uncensored instance is smaller than the censoring time of the censored instance (Steck et al. 2008). This can be visualized by the ordered graph shown in Figure 4. Figure 4(a) and Figure 4(b) are used to illustrate the possible ranking comparisons (denoted by edges between instances) for the survival data

without and with censored instances, respectively. There are $\binom{5}{2} = 10$ possible pairwise comparisons for the five instances in the survival data without censored cases shown in Figure 4(a). Due to the presence of censored instances (represented by red circles) in Figure 4(b), only 6 out of the 10 comparisons are feasible. Note that a censored instance can only be compared with an earlier uncensored instance (for example y_2&y_1). However, any censored instance cannot be compared with both censored and uncensored instances after its censored time (e.g., y_2&y_3 and y_2&y_4) since its actual event time is unknown.

Consider both the observations and prediction values for two instances, (y_1, $\hat{y}_1$) and (y_2, $\hat{y}_2$), where y_i and $\hat{y}_i$ represent the actual observation time and the predicted value, respectively. The concordance probability between them can be computed as

$$c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 \geq y_2). \tag{19}$$

Applying this definition, for the binary prediction problem, the C-index will have a similar meaning to the regular area under the ROC curve (AUC), and if y_i is binary, then the C-index is the AUC (Li et al. 2016d). As the preceding definition is not straightforward, in practice, there are multiple ways of calculating the C-index.

(1) When the output of the model is a hazard ratio (such as the outcome obtained by Cox based models), the C-index can be computed using

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I[X_i\hat{\beta} > X_j\hat{\beta}], \tag{20}$$

where $i, j \in \{1, \cdots, N\}$, num denotes the number of all comparable pairs, $I[\cdot]$ is the indicator function, and $\hat{\beta}$ represents the estimated parameters from the Cox based models.

(2) For survival methods which aim to directly learn the survival time, the C-index should be calculated as:

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)], \tag{21}$$

where $S(\cdot)$ corresponds to the estimated survival probabilities.

5.2 Brier Score

Named after its inventor, Glenn W. Brier, the Brier Score (BS) (Brier 1950) was developed to predict the inaccuracy of probabilistic weather forecasts. It can only evaluate prediction models with probabilistic outcomes; that is, the outcome must remain within the range [0,1], and the sum of all the possible outcomes for a certain individual should be 1. When we consider the binary outcome prediction for a sample of N instances and for each X_i ($i = 1, 2, \ldots, N$), the predicted outcome at t is $\hat{y}_i(t)$, and the actual outcome is $y_i(t)$. Then, the empirical definition of the BS at the specific time t can be given by

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} [\hat{y}_i(t) - y_i(t)]^2, \tag{22}$$

where the actual outcome $y_i(t)$ for each instance can only be 1 or 0.

The concept of the BS was extended in Graf et al. (1999) to serve as a performance measure for survival problems with censored information in order to evaluate prediction models where the outcome to be predicted is either binary or categorical in nature. When incorporating the censored information in the dataset, the individual contributions to the empirical BS are reweighted

according to the censored information. Then, the BS can be updated as follows:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N} w_i(t)[\hat{y}_i(t) - y_i(t)]^2. \tag{23}$$

In Equation (23), $w_i(t)$, given in Equation (24), denotes the weight for the ith instance, and it is estimated by incorporating the KM estimator of the censoring distribution G obtained for a given dataset $(X_i, y_i, 1-\delta_i), i = 1, \ldots, N$.

$$w_i(t) = \begin{cases} \delta_i / G(y_i) & \text{if } y_i \leq t \\ 1/G(y_i) & \text{if } y_i > t \end{cases}. \tag{24}$$

With this weight distribution, the weights for the instances that are censored before t will be 0. However, they still contribute indirectly to the calculation of the BS since they are used for calculating G. The weights for the instances that are uncensored at t are greater than 1 to ensure that they contribute their estimated survival probability to the calculation of the BS.

5.3 Mean Absolute Error

For survival analysis problems, the Mean Absolute Error (MAE) is defined as an average of the differences between the predicted time values and the actual observation time values. It is calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(\delta_i |y_i - \hat{y}_i|), \tag{25}$$

where y_i $(i = 1, \ldots, N)$ represents the actual observation times, and $\hat{y}_i$ $(i = 1, \ldots, N)$ denotes the predicted times. Note that only the samples for which the event occurs are being considered in this metric, since, if $\delta_i = 0$, the corresponding term will become zero. MAE can only be used for the evaluation of survival models where the event time can be provided for the predicted target value, such as AFT models.

Note that, among the three metrics, both C-index and BS can be adapted to evaluate the performance of survival models during a follow-up period. More details can be found in the supplementary material of this article.

6 RELATED TOPICS

In addition to the machine learning methods introduced in Section 4 and the traditional statistical survival methods discussed in Section 3, several other topics are also closely related to survival analysis. These are summarized in this section.

6.1 Early Prediction

One of the primary challenges in the context of survival analysis, as well as in general longitudinal studies, is that a sufficient number of events in the training data can be collected only by waiting for a relatively long period. This is the most significant difference from regular supervised learning problems, where the labels for each instance can be provided by a domain expert within a reasonable time period. Therefore, a good survival model should have the ability to forecast the event occurrence at some future time based on very limited event occurrence information at an early stage of a survival analysis problem.

There are many real-world applications desperately seeking new prediction models that perform well using only data collected at an early stage. For example, in the healthcare domain, it is critical to study the effect of a new treatment in order to understand the effectiveness of a potential new treatment or drug, which would benefit from an accurate estimate as early as possible. In this case,

the patients are monitored over a certain time period and the event of interest would be the patient being admitted to hospital due to the failure of the treatment. This scenario clearly indicates the need for algorithms that can predict the event occurrence effectively using only a few events.

To address this problem, an Early Stage Prediction (ESP) approach trained on early stages of survival analysis studies was proposed to predict the time-to-event in Fard et al. (2016). Here, two algorithms based on NB and BNs were developed by estimating the posterior probability of event occurrence based on different extrapolation techniques using the Weibull, Log-logistic and Log-normal distributions discussed in Section 3. The ESP framework is a two-stage algorithm: (i) estimate the conditional probability distribution based on the training data collected up to the early stage time point (t_c) of the study, and (ii) extrapolate the prior probability of the event into the future (t_f) using an AFT model with different distributions. According to the experimental results, the ESP framework provides more accurate predictions when the prior probability at the future time is appropriately estimated using current information on event occurrence.

6.2 Data Transformation

This section discusses two promising new data transformation techniques that will be useful for data preprocessing in survival analysis. Both of these approaches transform the data into a more conducive form so that other survival-based (or sometimes even standard) algorithms can be applied effectively.

6.2.1 Uncensoring approach. In survival data, incompleteness in the event (outcome) information makes it difficult for standard machine learning methods to learn from such data. The censored observations in survival data might look similar to unlabeled samples in classification or to unknown responses in a regression problem in the sense that the status or time to event is not known for some of these observations. However, unlike unlabeled samples where the labeling information is completely missing, the censored instances actually have partial informative labeling information that indicates the possible range of the corresponding true response (survival time). Such censored data have to be handled with special care within any machine learning method in order to make good predictions. Also, in survival analysis problems, only the information before a certain time point (before censoring occurs) is available for the censored instances, and this information should be integrated into the prediction algorithm to obtain the best results.

There are typically two naive ways of handling such censored data. One is to delete the censored instances, which performs well if the number of samples is sufficiently large and the censoring instances are not censored randomly. However, this will provide a suboptimal model because it neglects the information available in the censored instances (Delen et al. 2005; Burke et al. 1997). Treating censoring as event-free is another naive and simple choice. This method performs well for data with only a few censored instances but underestimates the true performance of the model. Although both approaches offer simple ways to handle censored data, they ignore useful information present in the data. To address this issue, two other approaches have been proposed in the literature to handle censored data.

(1) Group the instances in the given data into three categorizes (Zupan et al. 2000): (i) instances which experience the event of interest during the observation will be labeled as events, (ii) instances whose censored time is later than a predefined time point are labeled as event-free, and (iii) instances whose censored time is earlier than a predefined time point will be duplicated, and one will be labeled as an event and the other will be labeled as event-free. All these instances will be weighted by a marginal probability of event occurrence estimated by the KM method.

(2) For each censored instance, estimate the probability of the event and the probability of being censored (considering censoring as a new event) using the KM estimator and assign a new class label based on these probability values (Fard et al. 2016). For each instance in the data, when the probability of the event exceeds the probability of being censored, it will be labeled as an event; otherwise, it will be labeled as event-free, which indicates that even if there is complete follow-up information for that instance, there is an extremely low chance of event occurrence by the end of the observation time period.

6.2.2 Calibration. Censoring causes missing time-to-event labels, and this effect is compounded when dealing with datasets with high amounts of censored instances. Instead of using uncensoring approaches, calibration for survival analysis can also be used to solve this problem by learning optimal time-to-event labels for the censored instances. Generally, there are mainly two reasons for performing such calibrations. First, survival analysis models are built using a given dataset where the missing time to events for the censored instances are assigned a value such as the duration of the study or the last known follow-up time. However, this approach is not suitable for handling data with many censored instances because these inappropriately labeled censored instances cannot provide much information to survival methods. Calibration methods should instead be used to overcome this missing time-to-event problem in survival analysis. Second, dependent censoring in the data, where censoring depends on covariates, may lead to some bias in standard survival estimators such as the KM method. This is the driving force behind the imputed censoring approach, which calibrates the time-to-event attribute to decrease the bias of the survival estimators.

Vinzamuri et al. (2017) proposed a calibration method which uses a regularized inverse covariance-based imputation to overcome the problems just mentioned. This has the ability to capture correlations between censored instances as well as correlations between similar features. In a calibrated survival analysis, imputing an appropriate label value for each censored instance enables a new representation of the original survival data to be learned effectively. This approach fills a gap in the current literature by estimating the calibrated time-to-event values for these censored instances based on row-wise and column-wise correlations among the censored instances.

6.3 Complex Events

Until now, the discussion in this article has focused primarily on survival problems in which each instance can experience only one event of interest. However, in many real-world domains, each instance may experience different types of events, and each event may occur more than once during the observation. For example, in the healthcare domain, one patient may be hospitalized multiple times due to different medical conditions. Since this scenario is more complex than the problems discussed earlier, we consider them to be complex events. In this section, we will briefly discuss two techniques—namely, competing risks and recurrent events—for tackling such complex events.

6.3.1 Competing Risks. In a survival problem, if several different types of events are considered, but only one of them can occur for each instance over the follow-up period, then the competing risks will be defined as the probabilities of different events. In other words, competing risks can only exist in survival problems with more than one possible event of interest, but only one event will occur at any given time. For example, in the healthcare domain, a patient may have both a heart attack and lung cancer before his death, but the cause of death can only be either lung cancer or a heart attack, not both. In this case, competing risks are events that prevent an event of interest from occurring, which is different from censoring. Note that, in the case of censoring, the event of interest still occurs at a later time, while the event of interest is impeded.

6.3.2 Recurrent Events. In many application domains, the event of interest in a survival problems may occur several times during the observation time period. This is significantly different from a single event, such as the death of a patient in the healthcare domain, which can occur only once. Here, the outcome event can occur for each instance more than once during the observation time period. In survival analysis, we refer to events that can occur more than once as *recurrent events*, which contrasts with the case of competing risks just discussed. Typically, if all the recurring events for each instance are of the same type, the Counting Process (CP) algorithm (Andersen et al. 2012) can be used to tackle this problem. However, if there are different types of events or determining the order of the events is the main goal, other methods using stratified Cox (SC) approaches are more appropriate (Ata and Sözer 2007). These methods include the stratified CP, Gap Time, and Marginal approaches. These differ not only in the way they determine the risk set, but also in the data format required.

The reader can refer to the supplementary material of this article for more details regarding both competing risks and recurrent events problems.

7 APPLICATION DOMAINS

This section examines the applications of survival analysis in various real-world domains. Table 5 summarizes the events of interest, expected goals, and the features that are typically used in each of the applications described in this section.

7.1 Healthcare

In the healthcare domain, the starting point of the observation is usually a particular medical intervention, such as a hospitalization admission, beginning to take a certain medication, or receiving a diagnosis (Klein and Moeschberger 2005; Miller Jr 2011). The event of interest might be death, hospital readmission, discharge from the hospital, or any other interesting incident that can happen during an observation period. The missing trace of an observation is also an important characteristic of the data collected in this domain. For example, during a given hospitalization, some patients may be moved to another hospital, and, in such cases, that patient will become unobserved with respect to the first hospital after that time point. In healthcare applications, survival prediction models primarily aim at estimating the failure time distribution and the prognostic evaluation of different features, including histological, biochemical, and clinical characteristics (Marubini and Valsecchi 2004).

7.2 Reliability

In the field of reliability, it is common to collect data over a period of time and record the interesting events that occur within this period. Reliability prediction focuses on developing methods which are good at accurately estimating the reliability of new products (Modarres et al. 2009; Lyu 1996). The event of interest here corresponds to the time taken for a device to fail. In such applications, it is desirable to be able to estimate which devices will fail and, if they do, when they will fail. Survival analysis methods can help to build such prediction models using the information available for these devices. These models can provide early warnings about potential failures, which is significantly important to either prevent or reduce the likelihood of failures and to identify and correct the causes of device failures.

7.3 Crowdfunding

In recent years, the topic of crowdfunding has gained a lot of attention. Although crowdfunding platforms have been successful, the percentage of the projects that achieve their desired goal

Table 5. Summary of Various Real-world Application Domains Where Survival Analysis Has Been Successfully Used

Application	Event of interest	Estimation	Features
Healthcare (Miller Jr 2011) (Reddy and Li 2015)	Rehospitalization Disease recurrence Cancer survival	Likelihood of hospitalization within t days of discharge.	**Demographics**: age, gender, race. **Measurements**: height, weight, disease history, disease type, treatment, comorbidities, laboratory, procedures, medications.
Reliability (Lyu 1996) (Modarres et al. 2009)	Device failure	Likelihood of a device failing within t days.	**Product**: model, years after production, product performance history. **Manufacturer**: location, no. of products, average failure rate for all the products, annual sales of the product, total sales of the product. **User**: user reviews of the product.
Crowdfunding (Rakesh et al. 2016) (Li et al. 2016a)	Project success	Likelihood of a project being successful within t days.	**Projects**: duration, goal amount, category. **Creators**: past success, location, no. of projects. **Twitter**: no. of promotions, backers, communities. **Temporal**: no. of backers, funding, no. of retweets.
Bioinformatics (Li et al. 2016d) (Beer et al. 2002)	Cancer survival	Likelihood of cancer within time t.	**Clinical**: demographics, labs, procedures, medications. **Genomics**: gene expression measurements.
Student Retention (Murtaugh et al. 1999) (Ameri et al. 2016)	Student dropout	Likelihood of a student dropping out within t days.	**Demographics**: age, gender, race. **Financial**: cash amount, income, scholarships. **Pre-enrollment**: high-school GPA, ACT scores, graduation age. **Enrollment**: transfer credits, college, major. **Semester performance**: semester GPA, % passed credits, % dropped credits.
Customer Lifetime Value (Zeithaml et al. 2001) (Berger and Nasr 1998)	Purchase behavior	Likelihood of a customer purchasing from a given service supplier within t days.	**Customer**: age, gender, occupation, income, education, interests, purchase history. **Store/Online store**: location, customer reviews, customer service, price, quality, shipping fees and time, discount.
Click Through Rate (Yin et al. 2013) (Barbieri et al. 2016)	User clicking	Likelihood of a user clicking the advertisement within time t.	**User**: gender, age, occupation, interests, users click history. **Advertisement (ad)**: time of the ad, location of the ad on the website, topics of the ad, ad format, total click times of the ad. **Website**: no. of users of the website, page view each day of the website, no. of websites linking to the website.
Unemployment Duration in Economics (Kiefer 1988)	Getting a job	Likelihood of a person finding a new job within t days.	**People**: age, gender, major, education, occupation, work experience, city, expected salary. **Economics**: job openings, annual unemployment rates.

amount remains below 50% (Rakesh et al. 2015). Moreover, many of the prominent crowdfunding platforms follow the "all-or-nothing" policy. In other words, only if the goal is achieved before the predetermined time period can the pledged funding be collected. Therefore, in the crowdfunding domain, one of the most important challenges is to estimate the success probability of each project. The need to estimate project success probability is driving the development of new prediction approaches that integrate the advantages of both regression (for estimating the time for success) and classification (for considering both successful and failed projects simultaneously in the model) (Li et al. 2016). For the successful projects, the time to success can be collected easily. However, for the projects that fail, it is not possible to collect information about the length of the time required for project success. The only information that can be collected is the funding amount that was raised before the pre-determined project end date. Li et al. (2016) considered both the failed and successful projects simultaneously by using censored regression methods to fit the probability of project success using log-logistic and logistic distributions and predict the time needed for a project to have a realistic chance of becoming successful.

7.4 Bioinformatics

One of the most popular applications of survival analysis in the domain of bioinformatics is gene expression. Gene expression is the process of synthesizing a functional gene product based on its gene information and can be quantified by measuring either messenger RNA (mRNA) or proteins. Gene expression profiling is rapidly becoming a powerful technique for studying the cell transcriptome. In recent years, a number of studies (Li et al. 2016d; Beer et al. 2002) have correlated gene expression with survival outcomes for cancer applications at a genome-wide scale. Survival analysis methods are helpful in assessing the effect of a single gene on the survival prognosis and then identifying the most relevant genes to serve as biomarkers for patients. In this scenario, the event of interest is the specific type of cancer (or other disease of interest), and the goal is to estimate the likelihood of it occurring based on the gene expression measurement values. Generally, survival prediction based on gene expression data is a high-dimensional problem since each cell contains tens of thousands of mRNA molecules. The authors in Antonov et al. (2014) developed a statistical tool for biomedical researchers to enable them to define the clinical relevance of individual genes via their effect on patient survival outcomes. Survival analysis has been shown to be an effective way of predicting gene expression for a range of cancer data containing censored information.

7.5 Student Retention

In higher education, student retention rates can be evaluated by the percentage of students who return to the same university for the following semester after completing a semester of study. In the United States and around the world, one of the long-term goals of a university is to improve student retention. Higher student retention rates make it more probable that the university will be ranked higher, secure more government funding, and have an easier path to program accreditation. For all of these reasons, directors and administrators in higher education constantly try to implement new strategies to increase student retention. Survival analysis has shown some success in improving student retention (Murtaugh et al. 1999; Ameri et al. 2016). The goal of survival analysis is to estimate the time of event occurrence, which is critical for addressing student retention problems because correctly identifying whether a student will drop out and estimating when that drop out will happen are both important. In such cases, the ability to reliably estimate the drop out risk at an early stage of student education using both pre-enrollment and post-enrollment information would be immensely helpful.

7.6 Customer Lifetime Value

Customer Lifetime Value (LTV) (Berger and Nasr 1998; Zeithaml et al. 2001) refers to the profit that a customer brings to a retail outlet based on his or her purchase history. In the marketing domain, customer LTV is used to evaluate the relationships between customers and stores. It is important for a store to improve the LTV of its customers in order to maintain or increase its profits over the long term since it is often quite expensive to acquire new customers. In this case, the main goal of this problem is to identify the purchase patterns of customers who have a high LTV and provide recommendations for a relatively new user who has similar interests. Identifying loyal customers using LTV estimation has been studied by various researchers (Rosset et al. 2003; Mani et al. 1999) using survival analysis methods and data mining approaches, both of which are helpful in identifying purchase patterns. In this context, LTV can be defined using a survival function, which can then be used to estimate the time of purchase for every customer from a given store based on store information and also the customer demographic information available in the store's database including, for example, gender, income, and age.

7.7 Click-Through Rate

Nowadays, many free web services, including online news portals, search engines, and social networks present users with advertisements (Barbieri et al. 2016). Both the topics and the display orders of the ads will affect user clicking probability (Richardson et al. 2007). Studies have mostly focused on predicting the Click-Through Rate (CTR), which indicates the percentage of users who click on a given ad. This can be calculated as the ratio of the clicking times and the corresponding presentation times (no. of ad impressions). The CTR value indicates the attraction effect of the ad to users (Barbieri et al. 2016). The goal is to predict how likely the user will be to click on specific ads based on the available information about the website, users, and ads. The time taken to click the ad is considered the event time. Those users who did not click on the ads are considered to be censored observations.

7.8 Duration Modeling in Economics

Traditionally, duration data, which measure how long individuals remain in a certain state, is analyzed in biometrics and medical statistics using survival analysis methods. Actually, duration data also appear in a wide variety of situations in economics such as unemployment duration, marital instability, and time-to-transaction in the stock market. Among these, the unemployment duration problem, which is the most widely studied, analyzes the time people spend without a job (Gamerman and West 1987). Generally, in the domain of economics, the time spent being unemployed is extremely important since the length of unemployment plays a critical role in economics theories of job searches (Kiefer 1988). For this problem, the data contain information on the time duration of unemployment for each individual in the sample. The event of interest here is getting a new job for each person, and the objective is to predict the likelihood of getting a new job within a specific time period. It is desirable to understand how the re-employment probability changes over the period spent unemployed and to know more about the effect of unemployment benefits on these probabilities.

8 SOFTWARE RESOURCES

This section provides a list of the various software implementations developed for statistical methods and machine-learning algorithms for survival analysis. Table 6 summarizes the basic information for the software packages available for each survival method. Most of the existing survival analysis methods are implemented in R.

Table 6. Summary of Software Packages for Various Survival Analysis Methods

Algorithm	Software	Language	Link
Kaplan-Meier	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Nelson-Aalen			
Life-Table			
Basic Cox	survival	R	https://cran.r-project.org/web/packages/survival/index.html
TD-Cox			
Lasso-Cox	fastcox	R	https://cran.r-project.org/web/packages/fastcox/index.html
Ridge-Cox			
EN-Cox			
Oscar-Cox	RegCox	R	https://github.com/MLSurvival/RegCox
CoxBoost	CoxBoost	R	https://cran.r-project.org/web/packages/CoxBoost/
Tobit	survival	R	https://cran.r-project.org/web/packages/survival/index.html
BJ	bujar	R	https://cran.r-project.org/web/packages/bujar/index.html
AFT	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Baysian Methods	BMA	R	https://cran.r-project.org/web/packages/BMA/index.html
RSF	randomForestSRC	R	https://cran.r-project.org/web/packages/randomForestSRC/
BST	ipred	R	https://cran.r-project.org/web/packages/ipred/index.html
Boosting	mboost	R	https://cran.r-project.org/web/packages/mboost/
Active Learning	RegCox	R	https://github.com/MLSurvival/RegCox
Transfer Learning	TransferCox	C++	https://github.com/MLSurvival/TransferCox
Multi-Task Learning	MTLSA	Matlab	https://github.com/MLSurvival/MTLSA
Early Prediction	ESP	R	https://github.com/MLSurvival/ESP
Uncensoring			
Calibration	survutils	R	https://github.com/MLSurvival/survutils
Competing Risks	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Recurrent Events	survrec	R	https://cran.r-project.org/web/packages/survrec/

(1) *Non-parametric methods*: All three non-parametric survival analysis methods can be implemented by employing the function *coxph* and *survfit* in the *survival* package in R.
(2) *Semi-parametric methods*: Both the basic Cox model and the time-dependent Cox model can be trained using the *coxph* function in the *survival* package in R. Among the regularized Cox methods, the Lasso-Cox, Ridge-Cox, and EN-Cox methods can be trained using the *cocktail* function in the *fastcox* package. The *RegCox* package can be used to implement the OSCAR-Cox method, and the *CoxBoost* function in the *CoxBoost* package can fit a Cox model using a likelihood-based boosting algorithm.
(3) *Parametric methods*: The Tobit regression can be trained using the *survreg* function in the *survival* package. Buckley-James Regression can be fitted using the *bujar* package. The parametric AFT models can be trained using the *survreg* function with various distributions.
(4) *Machine learning methods*: The *BMA* package can be used to train a Bayesian model by averaging for Cox models. Bagging survival tree methods can be implemented using the *bagging* function in the R package *ipred,* and the random survival forest is implemented in the *rfsrc* function in the package *randomForestSRC*. The *mboost* function in the package *mboost* can be used to implement the boosting algorithm. The *arc* function in *RegCox*

package can be used to train the active learning survival model. Transfer-Cox model is written in C++ language. The multitask learning survival method is implemented using the *MTLSA* package in MATLAB.

(5) *Related topics*: The *ESP* package performs the early-stage prediction for survival analysis problems and incorporates uncensoring functions in the data pre-processing part. The *survutils* package in R can be used to implement the calibration for survival datasets. The *survfit* function in the *survival* package can also be used to train the model for competing risks. The function *Survr* in the package *survrec* can also be used to train the survival models for recurrent event data.

9 CONCLUSION

The primary goal of survival analysis is to predict the occurrence of specific events of interest at future time points. Due to the widespread availability of survival data from various domains, combined with the recent developments in various machine learning methods, there is an increasing demand for methods that can help us understand and improve the way we handle and interpret survival data. In this survey article, we provided a comprehensive review of conventional survival analysis methods and various machine learning methods for survival analysis, and we described various related topics, along with appropriate evaluation metrics. We began by introducing the basic notations and concepts related to survival analysis, including the structure of survival data and the common functions used in survival analysis. We then introduced the well-studied statistical survival methods and the representative machine learning-based survival methods. Related topics in survival analysis, including data transformation, early prediction, and complex events, were also discussed. Implementation details were provided for these survival methods, and we also described the commonly used performance evaluation metrics for these models. Besides the traditional applications in healthcare and biomedicine, survival analysis has also been successfully applied in various real-world problems, such as product reliability, student retention, and user behavior modeling. We hope that this survey article provides a coherent understanding of this important research topic and creates unique opportunities for both basic and applied researchers to pursue future research in this area.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation grants IIS-1619028, IIS-1707498 and IIS-1838730.

REFERENCES

Odd Aalen. 1978. Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6, 4 (1978), 701–726.

Paul D Allison. 2010. *Survival Analysis Using SAS: A Practical Guide*. Sas Institute.

Sattar Ameri, Mahtab J Fard, Ratna B Chinnam, and Chandan K Reddy. 2016. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of ACM International Conference on Conference on Information and Knowledge Management*. ACM, IN, 903–912.

Per Kragh Andersen, Ornulf Borgan, Richard D. Gill, and Niels Keiding. 2012. *Statistical Models Based on Counting Processes*. Springer Science & Business Media.

A. V. Antonov, M. Krestyaninova, R. A. Knight, I. Rodchenkov, G. Melino, and N. A. Barlev. 2014. PPISURV: A novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. *Oncogene* 33, 13 (2014), 1621–1628.

Nihal Ata and M. Tekin Sözer. 2007. Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics* 36, 2 (2007), 157–167.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. 2012. Structured sparsity through convex optimization. *Statistical Sciences* 27, 4 (2012), 450–468.

Bart Baesens, Tony Van Gestel, Maria Stepanova, Dirk Van den Poel, and Jan Vanthienen. 2005. Neural network survival analysis for personal loan data. *Journal of the Operational Research Society* 56, 9 (2005), 1089–1098.

Nicola Barbieri, Fabrizio Silvestri, and Mounia Lalmas. 2016. Improving post-click user engagement on native ads via survival analysis. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conference Committee, Montreal, 761–770.

David G. Beer, Sharon L. R. Kardia, Chiang-Ching Huang, Thomas J. Giordano, Albert M. Levin, David E. Misek, Lin Lin, Guoan Chen, Tarek G. Gharib, Dafydd G. Thomas, et al. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 8, 8 (2002), 816–824.

Riccardo Bellazzi and Blaz Zupan. 2008. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics* 77, 2 (2008), 81–97.

Paul D. Berger and Nada I. Nasr. 1998. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing* 12, 1 (1998), 17–30.

Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. 1998. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine* 17, 10 (1998), 1169–1186.

Harald Binder and Martin Schumacher. 2008. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* 9, 1 (2008), 1–10.

Imad Bou-Hamad, Denis Larocque, Hatem Ben-Ameur, et al. 2011. A review of survival trees. *Statistics Surveys* 5 (2011), 44–71.

Leo Breiman. 1996. Bagging predictors. *Machine Learning* 24, 2 (1996), 123–140.

Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

Norman E. Breslow. 1972. Discussion of the paper by D. R. Cox. *Journalof the Royal Statistical Society B* 34 (1972), 216–217.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1 (1950), 1–3.

Stephen F. Brown, Alan J. Branford, and William Moran. 1997. On the use of artificial neural networks for the analysis of survival data. *IEEE Transactions on Neural Networks* 8, 5 (1997), 1071–1077.

Jonathan Buckley and Ian James. 1979. Linear regression with censored data. *Biometrika* 66, 3 (1979), 429–436.

Peter Bühlmann and Torsten Hothorn. 2007. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Sciences* 22, 4 (2007), 477–505.

Harry B. Burke, Philip H. Goodman, David B. Rosen, Donald E. Henson, John N. Weinstein, Frank E. Harrell, Jeffrey R. Marks, David P. Winchester, and David G. Bostwick. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 79, 4 (1997), 857–862.

Ching-Fan Chung, Peter Schmidt, and Ana D. Witte. 1991. Survival analysis: A survey. *Journal of Quantitative Criminology* 7, 1 (1991), 59–98.

A. Ciampi, R. S. Bush, M. Gospodarowicz, and J. E. Till. 1981. An approach to classifying prognostic factors related to survival experience for non-Hodgkin's lymphoma patients: Based on a series of 982 patients: 1967–1975. *Cancer* 47, 3 (1981), 621–627.

A. Ciampi, C-H. Chang, S. Hogg, and S. McKinney. 1987. Recursive partition: A versatile method for exploratory-data analysis in biostatistics. In *Biostatistics*, I. B. MacNeill, G. J. Umphrey, A. Donner, and V. K. Jandhyala (Eds.). Springer, 23–50.

Antonio Ciampi, Johanne Thiffault, Jean-Pierre Nakache, and Bernard Asselain. 1986. Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis* 4, 3 (1986), 185–204.

Joseph A. Cruz and David S. Wishart. 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics* 2 (2006).

Sidney J. Cutler and Fred Ederer. 1958. Maximum utilization of the life table method in analyzing survival. *Journal of Chronic Diseases* 8, 6 (1958), 699–712.

Cox R. David. 1972. Regression models and life tables. *Journal of the Royal Statistical Society* 34, 2 (1972), 187–220.

Cox R. David. 1975. Partial likelihood. *Biometrika* 62, 2 (1975), 269–276.

Roger B. Davis and James R. Anderson. 1989. Exponential survival trees. *Statistics in Medicine* 8, 8 (1989), 947–961.

Dursun Delen, Glenn Walker, and Amit Kadam. 2005. Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine* 34, 2 (2005), 113–127.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Proceedings of the International Workshop on Multiple Classifier Systems*. Springer, 1–15.

Olive J. Dunn and Virginia A. Clark. 2009. *Basic Statistics: A Primer for the Biomedical Sciences*. John Wiley & Sons.

David Faraggi and Richard Simon. 1995. A neural network model for survival data. *Statistics in Medicine* 14, 1 (1995), 73–82.

Mahtab J. Fard, Ping Wang, Sanjay Chawla, and Chandan K. Reddy. 2016. A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (2016), 3126–3139.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Vol. 1. Springer series in statistics. Springer.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 3 (2008), 432–441.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning* 29, 2 (1997), 131–163.

Dani Gamerman and Mike West. 1987. An application of dynamic survival models in unemployment studies. *The Statistician* (1987), 269–274.

Louis Gordon and Richard A. Olshen. 1985. Tree-structured survival analysis. *Cancer Treatment Reports* 69, 10 (1985), 1065–1069.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18, 17-18 (1999), 2529–2545.

Sariel Har-Peled, Dan Roth, and Dav Zimak. 2002. Constraint classification: A new approach to multiclass classification. In *Algorithmic Learning Theory*, N. Cesa-Bianchi, M. Numao, and R. Reischuk (Eds.). Springer, 365–379.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247, 18 (1982), 2543–2546.

Frank E. Harrell, Kerry L. Lee, Robert M. Califf, David B. Pryor, and Robert A. Rosati. 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 3, 2 (1984), 143–152.

Patrick J. Heagerty and Yingye Zheng. 2005. Survival model predictive accuracy and ROC curves. *Biometrics* 61, 1 (2005), 92–105.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.

Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7, 3 (2006), 355–373.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. 2004. Bagging survival trees. *Statistics in Medicine* 23, 1 (2004), 77–91.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. Random survival forests. *The Annals of Applied Statistics* 2, 3 (2008), 841–860.

Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, and Andy J. Minn. 2011. Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining* 4, 1 (2011), 115–132.

How Jing and Alexander J. Smola. 2017. Neural survival recommender. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. ACM, Cambridge, 515–524.

John D. Kalbfleisch and Ross L. Prentice. 2011. *The Statistical Analysis of Failure Time Data*. Vol. 360. John Wiley & Sons.

Edward L. Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 282 (1958), 457–481.

Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. 2016. Deep survival: A deep Cox proportional hazards network. *arXiv preprint arXiv:1606.00931* (2016).

Carl T. Kelley. 1999. *Iterative Methods for Optimization*. Vol. 18. SIAM.

Faisal M. Khan and Valentina B. Zubek. 2008. Support vector regression for censored data (SVRc): A novel tool for survival analysis. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, Pisa, 863–868.

Farkhondeh Kiaee, Hamid Sheikhzadeh, and Samaneh Eftekhari Mahabadi. 2016. Relevance vector machine for survival analysis. *IEEE Transactions on Neural Networks and Learning Systems* 27, 3 (2016), 648–660.

Nicholas M. Kiefer. 1988. Economic duration data and hazard functions. *Journal of Economic Literature* 26, 2 (1988), 646–679.

John P. Klein and Melvin L. Moeschberger. 2005. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media.

David G. Kleinbaum and Mitchel Klein. 2006. *Survival Analysis: A Self-learning Text*. Springer Science & Business Media.

Igor Kononenko. 1993. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal* 7, 4 (1993), 317–337.

Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13 (2015), 8–17.

Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, et al. 2010. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* 5, 2 (2010), 369–411.

Michael LeBlanc and John Crowley. 1992. Relative risk trees for censored survival data. *Biometrics* 48, 2 (1992), 411–425.

Elisa T. Lee and John Wang. 2003. *Statistical Methods for Survival Data Analysis*. Vol. 476. John Wiley & Sons.

Yan Li, Vineeth Rakesh, and Chandan K. Reddy. 2016a. Project success prediction in crowdfunding environments. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*. ACM, San Francisco, California, 247–256.

Yan Li, Bhanukiran Vinzamuri, and Chandan K. Reddy. 2016b. Regularized weighted linear regression for high-dimensional censored data. In *Proceedings of SIAM International Conference on Data Mining*. SIAM, Miami, FL, 45–53.

Yan Li, Jie Wang, Jieping Ye, and Chandan K. Reddy. 2016d. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, CA, 1715–1724.

Yan Li, Lu Wang, Jie Wang, Jieping Ye, and Chandan K. Reddy. 2016c. Transfer learning for survival analysis via efficient L2,1-norm regularized Cox regression. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, Barcelona, 231–240.

Yan Li, Kevin S. Xu, and Chandan K. Reddy. 2016e. Regularized parametric regression for high-dimensional survival analysis. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, Miami, FL, 765–773.

Knut Liestbl, Per Kragh Andersen, and Ulrich Andersen. 1994. Survival analysis and neural nets. *Statistics in Medicine* 13, 12 (1994), 1189–1200.

Paulo J. G. Lisboa, H. Wong, P. Harris, and Ric Swindell. 2003. A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine* 28, 1 (2003), 1–25.

Michael R. Lyu. 1996. *Handbook of Software Reliability Engineering*. Vol. 222. IEEE computer society press CA.

David J. C. MacKay. 1995. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* 6, 3 (1995), 469–505.

D. R. Mani, James Drew, Andrew Betz, and Piew Datta. 1999. Statistics and data mining techniques for lifetime value modeling. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Diego, CA, 94–103.

L. Mariani, D. Coradini, E. Biganzoli, P. Boracchi, E. Marubini, S. Pilotti, B. Salvadori, R. Silvestrini, U. Veronesi, R. Zucali, et al. 1997. Prognostic factors for metachronous contralateral breast cancer: A comparison of the linear Cox regression model and its artificial neural network extension. *Breast Cancer Research and Treatment* 44, 2 (1997), 167–178.

Ettore Marubini and Maria Grazia Valsecchi. 2004. *Analysing Survival Data from Clinical Trials and Observational Studies*. Vol. 15. John Wiley & Sons.

Rupert Miller and Jerry Halpern. 1982. Regression with Censored Data. *Biometrika* 69, 3 (1982), 521–531.

Rupert G. Miller Jr. 2011. *Survival Analysis*. Vol. 66. John Wiley & Sons.

Mohammad Modarres, Mark P. Kaminskiy, and Vasiliy Krivtsov. 2009. *Reliability Engineering and Risk Analysis: A Practical Guide*. CRC press.

Paul A. Murtaugh, Leslie D. Burns, and Jill Schuster. 1999. Predicting the retention of university students. *Research in Higher Education* 40, 3 (1999), 355–371.

Wayne Nelson. 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics* 14, 4 (1972), 945–966.

Sinno J. Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.

Michael J. Pencina and Ralph B. D'Agostino. 2004. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine* 23, 13 (2004), 2109–2123.

Margaret S. Pepe. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

James L. Powell. 1994. Estimation of Semiparametric Models. *Handbook of Econometrics* 4 (1994), 2443–2521.

Adrian Raftery, David Madigan, and Chris T. Volinsky. 1995. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* 5, 323–349.

Adrian E. Raftery. 1995. Bayesian model selection in social research. *Sociological Methodology* 25 (1995), 111–163.

Vineeth Rakesh, Jaegul Choo, and Chandan K. Reddy. 2015. Project recommendation using heterogeneous traits in crowdfunding. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*. AAAI Press, Oxford, 337–346.

Vineeth Rakesh, Wang-Chien Lee, and Chandan K. Reddy. 2016. Probabilistic group recommendation model for crowdfunding domains. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*. ACM, San Francisco, California, 257–266.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. 2016. Deep survival analysis. *arXiv preprint arXiv:1608.02158* (2016).

Chandan K. Reddy and Yan Li. 2015. A review of clinical prediction models. *Healthcare Data Analytics* 36 (2015), 343–378.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, Banff, Alberta, 521–530.

Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain.*Psychological Review* 65, 6 (1958), 386–408.

Saharon Rosset, Einat Neumann, Uri Eick, and Nurit Vatnik. 2003. Customer lifetime value models for decision support. *Data mining and knowledge discovery* 7, 3 (2003), 321–339.

S. Rasoul Safavian and David Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics* 21, 3 (1991), 660–674.

Mark R. Segal. 1988. Regression trees for censored data. *Biometrics* 44, 1 (1988), 35–47.

Pannagadatta K. Shivaswamy, Wei Chu, and Martin Jansche. 2007. A support vector approach to censored targets. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, Omaha, Nebraska, 655–660.

Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani, et al. 2011. Regularization paths for Coxs proportional hazards model via coordinate descent. *Journal of Statistical Software* 39, 5 (2011), 1–13.

Alex J. Smola and Bernhard Schölkopf. 1998. *Learning with Kernels*. https://mitpress.mit.edu/books/learning-kernels.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 3 (2004), 199–222.

Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. 2008. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems*. Whistler, British Columbia, 1209–1216.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 1 (1996), 267–288.

Robert Tibshirani. 1997. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 16, 4 (1997), 385–395.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 1 (2005), 91–108.

James Tobin. 1958. Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society* 26, 1 (1958), 24–36.

Belle Van, Kristiaan Pelckmans, Huffel S. Van, and Johan A. K. Suykens. 2011. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine* 53, 2 (2011), 107–118.

Belle V. Van, Kristiaan Pelckmans, Johan A. K. Suykens, and S. Van Huffel. 2007. Support vector machines for survival analysis. In *Proceedings of the 3rd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED'07)*. Plymouth, 1–8.

Hans C. van Houwelingen and Hein Putter. 2011. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

Pierre J. M. Verweij and Hans C. Van Houwelingen. 1994. Penalized likelihood in Cox regression. *Statistics in Medicine* 13, 23-24 (1994), 2427–2436.

Bhanukiran Vinzamuri, Yan Li, and Chandan K. Reddy. 2014. Active learning based survival regression for censored data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, Shanghai, 241–250.

Bhanukiran Vinzamuri, Yan Li, and Chandan K. Reddy. 2017. Pre-processing censored survival data using inverse covariance matrix based calibration. *Transactions on Knowledge and Data Engineering* 29, 10 (2017), 2111–2124.

Bhanukiran Vinzamuri and Chandan K. Reddy. 2013. Cox regression with correlation based regularization for electronic health records. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, Dallas, TX, 757–766.

Sijian Wang, Bin Nan, Ji Zhu, and David G. Beer. 2008. Doubly penalized Buckley–James method for survival data with high-dimensional covariates. *Biometrics* 64, 1 (2008), 132–140.

Achmad Widodo and Bo-Suk Yang. 2011. Application of relevance vector machine and survival probability to machine degradation assessment. *Expert Systems with Applications* 38, 3 (2011), 2592–2599.

Guolei Yang, Ying Cai, and Chandan K. Reddy. 2018. Spatio-temporal check-in time prediction with recurrent neural network based survival analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. ACM, Stockholm, 2203–2211.

Sen Yang, Lei Yuan, Ying-Cheng Lai, Xiaotong Shen, Peter Wonka, and Jieping Ye. 2012. Feature grouping and selection over an undirected graph. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Beijing, 922–930.

Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang. 2017. Deep correlational learning for survival prediction from multi-modality data. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'17)*, Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne (Eds.). Springer International Publishing, 406–414.

Jieping Ye and Jun Liu. 2012. Sparse methods for biomedical data. *ACM SIGKDD Explorations Newsletter* 14, 1 (2012), 4–15.

Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. Silence is also evidence: Interpreting dwell time for recommendation from psychological perspective. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Chicago, IL, 989–997.

Valarie A. Zeithaml, Katherine N. Lemon, and Roland T. Rust. 2001. *Driving Customer Equity: How Customer Lifetime Value Is Reshaping Corporate Strategy*. Simon and Schuster.

Hao H. Zhang and Wenbin Lu. 2007. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 94, 3 (2007), 691–703.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2 (2005), 301–320.

Blaž Zupan, Janez Demšar, Michael W. Kattan, Robert J. Beck, and Ivan Bratko. 2000. Machine learning for survival analysis: A case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine* 20, 1 (2000), 59–75.

Received November 2016; revised December 2017; accepted April 2018