AN INTRODUCTION TO SURVIVAL ANALYSIS MATH

GEORGE H. CHEN

This note is part of the CHIL 2020 and SIGMETRICS 2021 tutorial "A Tour of Survival Analysis, from Classical to Modern" and provides an introduction to some of the math used in survival analysis, covering the basic problem setup (Section 1), a few classical survival analysis methods (Sections 2 and 3), and how to measure prediction accuracy (Section 4). Throughout the note, we use "log" to denote natural log, and phrase terminology in terms of predicting time until death, where higher survival times are considered better.

1. Problem Setup

Model. We assume we have access to n training subjects' data $(x_1, y_1, \delta_1), (x_2, y_2, \delta_2), \ldots, (x_n, y_n, \delta_n)$ where the i-th subject's feature vector is $x_i \in \mathbb{R}^d$, observed time duration is $y_i \in [0, \infty)$, and event indicator is $\delta_i \in \{0, 1\}$; if $\delta_i = 1$, then y_i is a survival time, and otherwise, y_i is a censoring time. We denote X to be the random variable corresponding to a generic feature vector, T to be the random variable corresponding to the true (possibly unobserved) survival time associated with feature vector X, and C to be the random variable corresponding to the true (possibly unobserved) censoring time associated with feature vector X. The training data are modeled to be sampled i.i.d. as follows:

(1) Sample feature vector x_i from a feature distribution $\mathbb{P}_X$
(2) Sample true survival time z_i from a conditional survival time distribution $\mathbb{P}_{T|X=x_i}$
(3) Sample true censoring time c_i from a conditional censoring time distribution $\mathbb{P}_{C|X=x_i}$
(4) If $z_i \leq c_i$ (death happens before censoring): output $y_i = z_i$ and $\delta_i = 1$ (no censoring)
Otherwise: output $y_i = c_i$ and $\delta_i = 0$ (true survival time is censored)

Note that the true distributions $\mathbb{P}_X$, $\mathbb{P}_{T|X}$, and $\mathbb{P}_{C|X}$ are unknown. Also, importantly, the survival time T and censoring time C are conditionally independent given feature vector X. In our derivations below, we assume that the conditional survival time distribution $\mathbb{P}_{T|X}$ has a PDF $f(t|x)$ and CDF $F(t|x) = \int_0^t f(\tau|x)d\tau$.

Prediction task. The main prediction task in survival analysis can be stated as estimating (some variant of) the survival function

$$\begin{aligned} S(t|x) &\triangleq \mathbb{P}(\text{subject survives beyond time } t \mid \text{subject's feature vector is } x) \\ &= \mathbb{P}(T > t|X = x) = 1 - \mathbb{P}(T \leq t|X = x) = 1 - F(t|x). \end{aligned} \tag{1.1}$$

Using the n training data, we aim to construct an estimate $\widehat{S}$ of S such that for any test feature vector x (e.g., a new subject/patient), we output the predicted survival curve $\widehat{S}(\cdot|x)$. Note that we are predicting a full curve per test feature vector and *not* just a single number (survival time) per feature vector.

As equation (1.1) indicates, the true survival curve $S(\cdot|x)$ is 1 minus a CDF $F(\cdot|x)$. A few implications:

(a) $S(\cdot|x) = 1 - F(\cdot|x)$ monotonically decreases from 1 to 0 since CDF's monotonically increase from 0 to 1.
(b) Estimating the function S is equivalent to estimating the CDF F, which means that we are learning the conditional survival time distribution $\mathbb{P}_{T|X}$. (The challenge in survival analysis is that in our training data, only the non-censored data have y_i values that come from $\mathbb{P}_{T|X}$. However, the censored data still have valuable information and should not be discarded!)
(c) If we want a single number survival time estimate for feature vector x, we can back one out if we know (an estimate of) $S(\cdot|x)$. We give two ways of doing this:
   - *Median survival time.* Where a CDF crosses $1/2$ corresponds to a median of a distribution, so finding a time t for which $S(t|x) = 1 - F(t|x) = 1/2$ gives a *median* survival time of feature vector x. In practice, we only have an estimate $\widehat{S}(\cdot|x)$ of $S(\cdot|x)$ so we find t such that $\widehat{S}(t|x) \approx 1/2$.
   - *Mean survival time.* For any nonnegative random variable Z, recall that $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z > \tau)d\tau$. Thus, the mean survival time of feature vector x is $\mathbb{E}[T|X = x] = \int_0^\infty \mathbb{P}(T > \tau|X = x)d\tau = \int_0^\infty S(\tau|x)d\tau$, the area under the survival curve. In practice, we numerically integrate the survival curve estimate $\widehat{S}(\cdot|x)$.

Date: July 16, 2020.

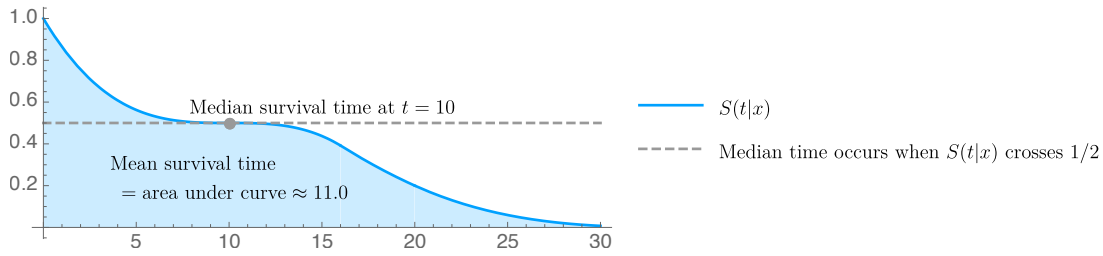

Figure 1.1. Example of a survival curve and its median and mean survival times.

See Figure 1.1 for an example survival curve and its corresponding median and mean survival times.

Keep in mind that we do not know the ground truth $S(t|x)$ even in our labeled training data. Consequently, we cannot compute an error such as $\widehat{S}(t|x) - S(t|x)$ even for the training data without imposing assumption(s) on S. We discuss how to measure prediction accuracy in Section 4.

Different survival analysis methods make different assumptions on S and often estimate transformed variants of S rather than estimating S directly. We discuss two transformed versions that are commonly used.

Cumulative hazard function. A first commonly estimated variant of S is the so-called cumulative hazard function, which we get by taking the negative log of S:

$$H(t|x) \triangleq -\log S(t|x) = \log \frac{1}{S(t|x)}.$$

This function is just another way to represent S: if we know $H(t|x)$, then we can recover $S(t|x) = \exp(-H(t|x))$. Some survival analysis methods directly estimate the cumulative hazard function such as random survival forests (Ishwaran et al., 2008). In terms of the shape of $H(\cdot|x)$, note that taking the log of a probability yields a value from $-\infty$ to 0. Taking the negative of a log probability thus yields a nonnegative value. Thus, whereas $S(\cdot|x)$ monotonically decreases from 1 to 0, $H(\cdot|x)$ monotonically increases from 0 to ∞. We remark that log probabilities are regularly used in machine learning, often due to numerical issues (e.g., taking products of many small probabilities can underflow whereas doing this calculation in log space is more numerically stable).

Hazard function. A key reason for why the cumulative hazard function H is interesting is that its derivative with respect to time has a simple interpretation in survival analysis; this derivative is called the hazard function and is the second commonly estimated variant of S that we discuss:

$$h(t|x) \triangleq \frac{\partial}{\partial t} H(t|x) = -\frac{\partial}{\partial t} \log S(t|x) = -\frac{\frac{\partial}{\partial t} S(t|x)}{S(t|x)} \overset{\text{by equation (1.1)}}{=} -\frac{\frac{\partial}{\partial t}[1 - F(t|x)]}{S(t|x)} = \frac{f(t|x)}{S(t|x)}, \tag{1.2}$$

where as a reminder, $f(\cdot|x)$ is the PDF corresponding to CDF $F(\cdot|x)$ of distribution $\mathbb{P}_{T|X=x}$. *The right-most expression reveals that the hazard function is the instantaneous rate of death conditioned on surviving up to time t for feature vector x.* Perhaps the most widely used survival model, the Cox proportional hazards model (Cox, 1972), places a structural assumption on the hazard function. Note that if we know the hazard function $h(t|x)$, then integrating it with respect to time recovers $H(t|x) = \int_0^t h(\tau|x)d\tau$ (motivating the terminology of H being the *cumulative* hazard function), from which we can recover $S(t|x) = \exp(-H(t|x))$.

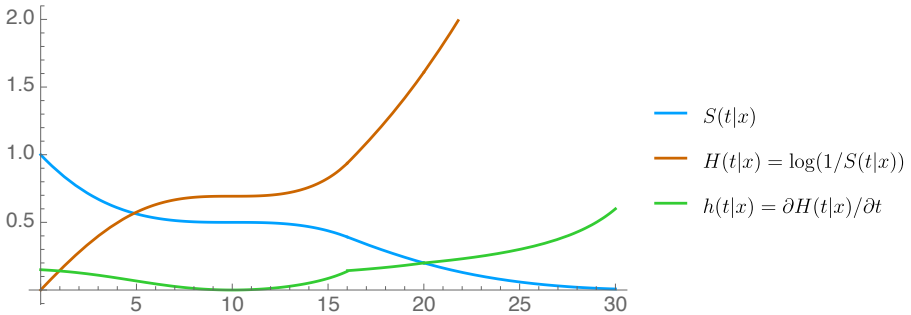

Figure 1.2. The survival curve S from Figure 1.1 along with its cumulative hazard H and hazard h functions.

An example survival function $S(\cdot|x)$ and its corresponding cumulative hazard function $H(\cdot|x)$ and hazard function $h(\cdot|x)$ are shown in Figure 1.2. Note that whereas S and H are monotonic functions, h need not be monotonic.

2. The Kaplan-Meier Estimator and Its Conditional Variants

We now discuss some classical survival analysis methods. For the moment, let's disregard feature vectors, treating our training data as $(y_1, \delta_1), (y_2, \delta_2), \ldots, (y_n, \delta_n)$, and setting our prediction goal to be the marginal survival function $S_{\text{marg}}(t) \triangleq \mathbb{P}(\text{subject survives beyond time } t) = \mathbb{P}(T > t)$. Kaplan and Meier (1958) suggested the following approach for estimating $S_{\text{marg}}(t)$. We first compute the unique times of death $t_1, t_2, \ldots, t_L$. Let d_i be the number of deaths at time t_i, and n_i be the number of subjects "at risk" (could possibly die) at time t_i:

$$d_i \triangleq \sum_{j=1}^{n} \mathbf{1}\{y_j = t_i\}\delta_j, \qquad \text{and} \qquad n_i \triangleq \sum_{j=1}^{n} \mathbf{1}\{y_j \geq t_i\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Then the Kaplan-Meier estimate for $S_{\text{marg}}(t)$ is

$$\widehat{S}_{\text{marg}}(t) \triangleq \prod_{i=1}^{L} \left(1 - \frac{d_i}{n_i}\right)^{\mathbf{1}\{t_i \leq t\}}.$$

This estimator has an intuitive explanation: it is the product of the probabilities of surviving from time 0 to t_1, t_1 to t_2, and so forth up until time t. For example:

- When $0 \leq t < t_1$, $\widehat{S}_{\text{marg}}(t) = 1$.
- When $t_1 \leq t < t_2$, $\widehat{S}_{\text{marg}}(t) = (1 - \frac{d_1}{n_1})$.
- When $t_2 \leq t < t_3$, $\widehat{S}_{\text{marg}}(t) = (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2})$.

As t increases, we multiply more and more probabilities, so $\widehat{S}_{\text{marg}}$ monotonically decreases in a piecewise constant fashion, like a descending staircase.

We can incorporate feature vectors in a straightforward manner. Suppose that our training data have feature vectors: $(x_1, y_1, \delta_1), (x_2, y_2, \delta_2), \ldots, (x_n, y_n, \delta_n)$. Then for a feature vector x that we want to predict the survival curve $S(\cdot|x)$ for, we first compute the k nearest neighbors to x among the training data according to a user-specified distance function (e.g., find the k different x_i's that are closest to x according to Euclidean distance). Then we compute the Kaplan-Meier estimator restricted to only using the y_i and δ_i values from these k nearest neighbors. Thus, the predicted survival curve now depends on the test feature vector x! This approach was proposed by Beran (1981), who also proposed a more general kernel variant. These variants are sometimes called conditional Kaplan-Meier estimators. Theory is well understood for how well conditional Kaplan-Meier estimators approximate the true survival curve $S(t|x)$; finite-sample error bounds are provided by Chen (2019).

The Kaplan-Meier estimator and its conditional variants are examples of a nonparametric methods as they don't assume survival curves to come from parametric distributions. We now examine a model with parameters.

3. The Cox Proportional Hazards Model

The extremely popular Cox proportional hazards model (Cox, 1972) is essentially the linear regression analogue in survival analysis (although it's nonlinear). The Cox model assumes that the hazard function factorizes as

$$h(t|x) = h_0(t)\exp(\beta^\top x) \qquad \textit{proportional hazards assumption} \tag{3.1}$$

for a feature weighting vector $\beta \in \mathbb{R}^d$ and a baseline hazard function h_0, which takes as input a nonnegative time and outputs a nonnegative value. Note that h_0 and β are parameters that need to be estimated. Different feature vectors x change the weighting term $\exp(\beta^\top x)$ but the overall hazard function is constrained to always be proportional to $h_0(t)$. Note that the Cox model is considered semi-parametric since, whereas the influence of β has a parametric form, commonly no parametric form is assumed for h_0. As a side remark, the log of the hazard function is linear in x: $\log h(t|x) = \beta^\top x + \log h_0(t)$.

Implications for survival curves. Suppose we know what h_0 and β are for the moment. Let's look at what the proportional hazards assumption implies about the shapes of survival curves that are allowed by the model.

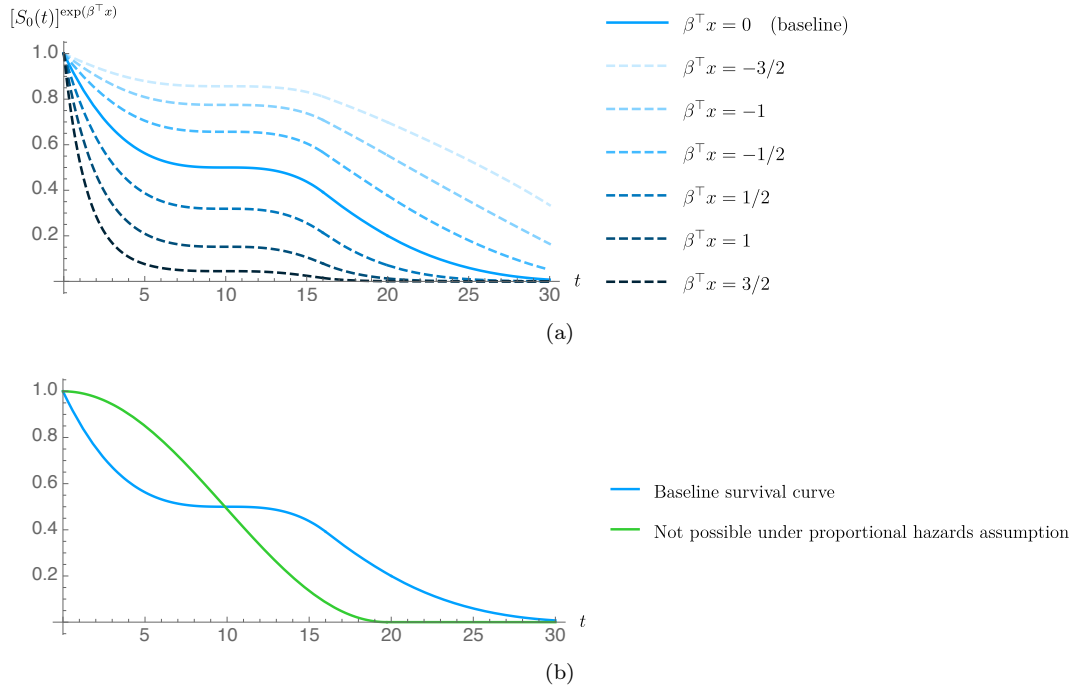

Figure 3.1. Under the proportional hazards assumption, possible survival curves are all powers of the baseline survival curve S_0 as shown in panel (a); note that we can always unambiguously order these curves based on the inner product value $\beta^\top x$. In contrast, the green curve shown in panel (b) is not possible under a proportional hazards model and is neither uniformly better nor worse than the baseline survival curve.

Recall from Section 1 that by knowing the hazard function $h(t|x)$, we can recover $H(t|x) = \int_0^t h(\tau|x)d\tau$ and $S(t|x) = \exp(-H(t|x))$. We now just plug in equation (3.1). First, we have the cumulative hazard function

$$H(t|x) = \int_{0t}^{t} h(\tau|x)d\tau = \int_0^t h_0(\tau)\exp(\beta^\top x)d\tau = \exp(\beta^\top x)\underbrace{\int_0^t h_0(\tau)d\tau}_{\triangleq H_0(t)}, \tag{3.2}$$

where H_0 is referred to as the baseline cumulative hazard function. Then we recover the survival curve

$$S(t|x) = \exp(-H(t|x)) = \exp\big(-\exp(\beta^\top x)H_0(t)\big) = [\underbrace{e^{-H_0(t)}}_{\triangleq S_0(t)}]^{\exp(\beta^\top x)}. \tag{3.3}$$

Thus, all possible survival curves under the proportional hazards assumption are powers of the baseline survival curve $S_0(t)$ — see Figure 3.1(a) for an illustration. This is a strong assumption! *Survival curves that are not powers of $S_0(t)$ are not allowed by the model (e.g., the green curve in Figure 3.1(b)).* As shown in Figure 3.1(a), the allowed survival curves that are closer to the origin—which have higher $\beta^\top x$ value—are uniformly worse than ones farther away from the origin, regardless of what time t we look at. In particular, under a proportional hazards assumption, whether a subject with feature vector x has shorter or longer survival time is entirely determined by the inner product value $\beta^\top x$: higher values of $\beta^\top x$ correspond to shorter survival times. In real-world problems, different subjects' survival curves need not satisfy the proportional hazards assumption, and we could have survival curves that crisscross like the ones in Figure 3.1(b).

Parameter estimation. The Cox model parameters h_0 and β are estimated via maximum likelihood. We state how to do this where no parametric form is assumed for h_0. We first estimate β without knowing h_0 by solving the convex program

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^k} \sum_{i=1}^n \delta_i \Big[-\beta^\top x_i + \log\Big(\sum_{j=1}^n \mathbf{1}\{y_j \geq y_i\}\exp(\beta^\top x_j)\Big)\Big]. \tag{3.4}$$

This problem can be solved using, for instance, gradient descent or Newton-Raphson.

Next, we estimate a discretized version of $h_0(t)$, specifically at the time points given by the unique times of death $t_1, t_2, \ldots, t_L$. Let d_i denote the number of deaths at time t_i, and $\widehat{h}_{0,i}$ denote the estimate of $h_0(t_i)$ for each time index $i = 1, 2, \ldots, L$. We compute

$$\widehat{h}_{0,i} = \frac{d_i}{\sum_{j=1}^{n} \mathbf{1}\{Y_j \geq t_i\} e^{\widehat{\beta}^\top x_j}}. \tag{3.5}$$

Lastly, we mention how to back out estimates of the cumulative hazard function $H(t|x)$ and the survival function $S(t|x)$. Recall from equation (3.2) that $H(t|x) = \exp(\beta^\top x) H_0(t)$ where $H_0(t) = \int_0^t h_0(\tau) d\tau$ is the baseline cumulative hazard function. Since we are using a discrete approximation for $h_0(t)$, we estimate $H_0(t)$ via a finite summation:

$$\widehat{H}_0(t) = \sum_{i=1}^{L} \mathbf{1}\{t_i \leq t\} \widehat{h}_{0,i}.$$

This leads to the cumulative hazard function estimate $\widehat{H}(t|x) = \exp(\widehat{\beta}^\top x) \widehat{H}_0(t)$. Finally, recalling that $S(t|x) = \exp(-H(t|x))$, we can estimate $\widehat{S}(t|x) = \exp(-\widehat{H}(t|x))$. An alternative approach to computing this same quantity follows from equation (3.3): we first compute baseline survival curve estimate $\widehat{S}_0(t) = \exp(-\widehat{H}_0(t))$, and then compute $\widehat{S}(t|x) = [\widehat{S}_0(t)]^{\exp(\widehat{\beta}^\top x)}$.

In the procedure above, estimating β by solving optimization problem (3.4) was suggested by Cox (1972), whereas estimating h_0 with equation (3.5) is due to Breslow (1972). Why this two-step procedure maximizes a joint likelihood over β and h_0 is explained by Johansen (1983) using counting processes; see also the paper by Andersen and Gill (1982).

4. Assessing Prediction Accuracy: Concordance Index

One of the most common accuracy metrics used in survival analysis is Harrell's concordance index (Harrell et al., 1982), often abbreviated as "c-index". The c-index is the fraction of pairs of subjects that are correctly ranked by a prediction procedure among pairs of subjects that can be ranked unambiguously, which as we will see shortly neatly handles censored data. As it is a fraction, the c-index ranges from 0 to 1, where 1 is the best performance. C-indices can be computed for a validation or test set and not just the training data.

Let's work out a simple example. Consider if we have three patients Alice, Bob, and Charlie. Suppose that Alice dies after 2 days, Bob dies after 10 days, and Charlie's survival time is unknown but we know that he was still alive after 6 days. In other words, their respective observed times (y_i values) and event indicators (δ_i values) are $(2, 1)$ for Alice, $(10, 1)$ for Bob, and $(6, 0)$ for Charlie. We know for sure that Alice died before Bob and that Alice died before Charlie. However, we do not know which of Bob or Charlie has a shorter survival time. Thus, in this example, only two pairs (Alice & Bob, Alice & Charlie) have subjects that can be ranked unambiguously. If a prediction procedure ranks Alice as having shorter survival time than Bob, and Alice as having a longer survival time than Charlie, then only one of the pairs is correctly predicted so the c-index is $1/2$.

Computing c-indices requires translating a survival analysis method's prediction output into some way to rank subjects. This can easily be done for the Cox proportional hazards model. As mentioned previously, under the proportional hazards assumption, how different feature vectors' survival curves differ is entirely determined by the inner product $\beta^\top x$. Thus, after we estimate $\widehat{\beta}$, for any pair of feature vectors x and x', we have a ranking for which feature vector has a lower survival curve (and thus shorter survival time) — whichever of $\widehat{\beta}^\top x$ and $\widehat{\beta}^\top x'$ is larger corresponds to the feature vector with shorter survival time.

For other survival models, especially ones that do not operate under a proportional hazards assumption (such as the nearest-neighbor Kaplan-Meier estimator from Section 2), translating a prediction output into a ranking could be less straightforward. One approach is to use predicted median or mean survival times to rank subjects. Alternatively, random survival forests use a "mortality" metric for ranking subjects (Ishwaran et al., 2008). Here's how this works (slightly simplified for ease of exposition). Let the unique times of death in the training data be $t_1, t_2, \ldots, t_L$. Let $\widehat{H}$ be a predicted cumulative hazard function (random survival forests directly predict cumulative hazards functions). Then the mortality approach says that a subject with feature vector x has a shorter survival time than a subject with feature vector x' if

$$\sum_{i=1}^{L} \widehat{H}(t_i|x) > \sum_{i=1}^{L} \widehat{H}(t_i|x'). \tag{4.1}$$

For the Cox model, this mortality approach to ranking actually is the same as using the inner product since

$$\sum_{i=1}^{L} \widehat{H}(t_i|x) = \sum_{i=1}^{L} \exp(\widehat{\beta}^\top x)\widehat{H}_0(t_i) = \exp(\widehat{\beta}^\top x) \sum_{i=1}^{L} \widehat{H}_0(t_i),$$

i.e., inequality (4.1) happens precisely when

$$\exp(\widehat{\beta}^\top x) \sum_{i=1}^{L} \widehat{H}_0(t_i) > \exp(\widehat{\beta}^\top x') \sum_{i=1}^{L} \widehat{H}_0(t_i) \qquad \Leftrightarrow \qquad \widehat{\beta}^\top x > \widehat{\beta}^\top x'.$$

Other accuracy metrics are possible as well. There is the time-dependent concordance index by Antolini et al. (2005). Separately, there are accuracy metrics such as the Brier score and its integrated variant (Graf et al., 1999; Gerds and Schumacher, 2006) that are quite different from concordance index and focus directly on how well the survival curve S is estimated.

References

Per Kragh Andersen and Richard D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120, 1982.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

Rudolf Beran. Nonparametric regression with randomly censored survival data. *Technical report, University of California, Berkeley*, 1981.

Norman Breslow. Discussion of the paper by D. R. Cox (1972) cited below. *Journal of the Royal Statistical Society, Series B*, 34(2):216–217, 1972.

George H. Chen. Nearest neighbor and kernel survival analysis: Nonasymptotic error bounds and strong consistency rates. In *International Conference on Machine Learning*, pages 1001–1010, 2019.

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

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

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

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

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

Søren Johansen. An extension of Cox's regression model. *International Statistical Review*, pages 165–174, 1983.

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