Daniel O. Scharfstein, Yuxin Zhu, Anastasios Tsiatis

Survival Analysis

0

Survival Analysis

CONTENTS

0.1 Introduction vii
0.2 ACTG 320 vii
0.3 Notation viii
0.3.1 Hazard x
0.3.2 Censoring x
0.4 Estimation of Survival Distribution xi
0.5 Hypothesis Testing xii
0.6 Cox Regression Model xv
0.7 Sample Size Calculations xviii
0.8 Informative Censoring xx
0.9 Conclusion xxiii

0.1 Introduction

In many randomized, controlled clinical trials, the primary endpoint is often time from randomization until the occurrence of an event of interest (e.g., death, relapse). The primary endpoint is commonly referred to as *time to event*, *survival time*, or *failure time*. In such trials, the major focus is on drawing inference about the distribution of time to event for competing treatments.

In most clinical trials, the time to event may not be observed as each subject is only followed over a finite time horizon and the event of the interest has not occurred before the end of that horizon. The follow-up time can vary from patient to patient. This variation can be due to staggered entry into the clinical trial, loss to follow-up or premature discontinuation of participation in the trial. Those subjects who do not have observed failure times are referred to as *right censored*. For such subjects, partial information is available about the time to event. Specifically, it is known that the failure time occurs after the follow-up time. Right censoring led to a whole new area of statistics called *Survival Analysis*.

0.2 ACTG 320

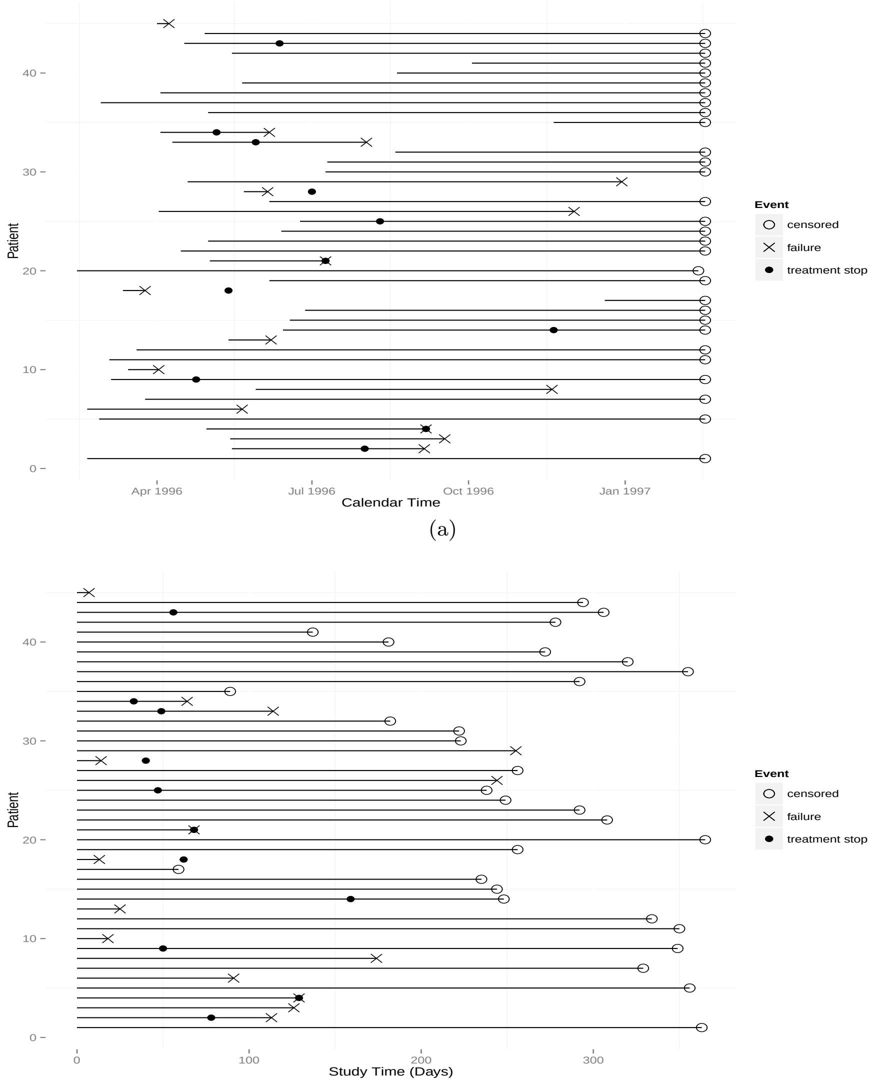
ACTG 320 was a randomized trial designed to evaluate whether indinavir sulfate is effective in treating patients with advanced HIV disease (i.e., CD4 counts less than 200) [24]. Patients were randomized to receive open-label AZT and 3TC with or without indinavir sulfate for at least 48 weeks. Randomization was stratified according to CD4 count measured at the time of screening: greater than 50 versus less than or equal to 50. Eleven hundred and fifty-six patients were randomized between January 29, 1996 and January 27, 1997. Patients who developed intolerance to AZT or had progressive disease after 24 weeks on study were allowed to substitute d4T for AZT. Patients were scheduled to be followed at weeks 4, 8, 16, 24, 32, 40, and 48 and every 8 weeks thereafter up to week 96. The primary end point was the time from randomization to the development of the AIDS or death.

Figure 1(a) presents a schematic representation of data for a random sample of 45 patients from ACTG 320. Each line represents data for an individual patient. The line starts at the calendar time of randomization. The line ends at the calendar time of end of follow-up. The symbol at the end of the line denotes the patient status on that calendar date. If it is an x, then the patient either developed AIDS or died at that time point. For these patients, the length of the line represents the failure time. If the symbol at the end of the line is a $\circ$, then follow-up has ended at that calendar time without the occurrence of AIDS or death. For these patients, the occurrence of the event of interest is known to occur after the last date of follow-up and the failure time is larger than the length of the line. Figure 1(b) presents the same data but on a study time scale (in days), i.e., time zero is the date of randomization. For the moment, ignore the treatment stop symbol on these figures; we will discuss the use of these data in Section 0.8.

0.3 Notation

Let T denote the time to event. Let $F(t) = P(T \leq t)$ and $S(t) = P(T > t)$ be the cumulative distribution function and survivor functions of the random variable T. We assume that $F(\cdot)$ can possibly have a countable number of jumps at finite times $0 \leq u_1 < u_2 < \ldots$; it is right continuous and right differentiable between the jumps. Define $u_0 = 0$ and $\tau = \sup_j u_j$. Let $dF(t) = F'(t)dt$ if t is a continuity point of $F(\cdot)$ and $dF(t) = F(t) - F(t-)$ if t is a jump point of $F(\cdot)$, where $F'(t)$ is the right derivative of $F(\cdot)$ at t [1] .

[1] $F'(t) = \lim_{dt \to 0+} \frac{F(t+dt)-F(t)}{dt} = \lim_{dt \to 0+} \frac{P[t \leq T \leq t+dt]}{dt}$



(a)

(b)

FIGURE 1
ACTG 320: Schematic representation of data for a random sample of 45 patients. (a) Calendar time scale; (b) Study time scale.

0.3.1 Hazard

The (net) hazard is a useful way of characterizing the distribution of T as it describes the changing risk of failure over time among those who remain at risk. Let

$$d\Lambda(t) = \frac{dF(t)}{S(t-)},$$

where $\Lambda(t)$ is the integrated or cumulative hazard function. When t is a continuity point of $F(\cdot)$, $d\Lambda(t) = \lambda(t)dt$, where

$$\lambda(t) = \frac{F'(t)}{S(t-)} = \lim_{dt \to 0+} \frac{P[t \leq T \leq t + dt \mid T \geq t]}{dt};$$

and when t is a jump point of $F(\cdot)$, $d\Lambda(t) = P[T = t|T \geq t]$. Notice that when t is a jump point of $F(\cdot)$, $d\Lambda(t)$ is the conditional probability of experiencing an event at time t given it occurs at or after t. If t is a continuity point of $F(\cdot)$, $d\Lambda(t)$ is approximately equal, for small dt, to the conditional probability of experiencing an event in the interval $[t, t + dt]$ given it occurs at or after t. The function $\lambda(t)$ is called the hazard rate, which is the instantaneous risk of an event at time t given it occurs at or after t. The hazard rate is NOT a probability.

The survival function can be written in terms of the hazard as follows:

$$S(t) = \prod_{u_j \leq t} \{1 - d\Lambda(u_j)\} \times \exp\left\{ -\sum_{j \geq 1} \int_{u_{j-1}}^{u_j} I(s \leq t) d\Lambda(s) - \int_{\tau}^{\infty} I(s \leq t) d\Lambda(s) \right\}. \tag{0.1}$$

0.3.2 Censoring

Let C denote the follow-up time defined in the hypothetical world in which the time to event does not pre-empt its observation (e.g., time from randomization until database lock). We consider the observed outcome data for an individual as (X, Δ), where $X = \min(T, C)$ and $\Delta = I(T \leq C)$. If $\Delta = 1$, then the time to event is observed (i.e., $T = X$). If $\Delta = 0$, then the time to event is known to occur after X (i.e., $T > X$) .

The distribution of the observed data for an individual can be characterized by the following quantities: $S_X(t) = P[X > t]$ and $F^\dagger(t) = P[X \leq t, \Delta = 1]$. The latter quantity is referred to as the sub-distribution for failure. Notice that $P[\Delta = 1] = F^\dagger(\infty)$ and $P[X \leq t, \Delta = 0] = 1 - S_X(t) - F^\dagger(t)$. Another characteristic of the distribution of the observed data is the cause-specific or observed hazard for failure defined as:

$$d\Lambda^\dagger(t) = \frac{dF^\dagger(t)}{S_X(t-)}.$$

When t is a continuity point of $F^{\dagger}(\cdot)$, $d\Lambda^{\dagger}(t) = \lambda^{\dagger}(t)dt$, where

$$\lambda^{\dagger}(t) = \lim_{dt \to 0+} \frac{P[t \leq X \leq t + dt, \Delta = 1 \mid X \geq t]}{dt};$$

and when t is a jump point of $F^{\dagger}(\cdot)$, $d\Lambda^{\dagger}(t) = P[X = t, \Delta = 1 | X \geq t]$. Notice that when t is a jump point of $F^{\dagger}(\cdot)$, $d\Lambda^{\dagger}(t)$ is the conditional probability of *observing* a failure event at time t given at risk for *observing* failure at time t. If t is a continuity point of $F^{\dagger}(\cdot)$, $d\Lambda^{\dagger}(t)$ is approximately equal, for small dt, to the conditional probability of *observing* a failure in the interval $[t, t + dt]$ given at risk for *observing* failure at time t.

0.4 Estimation of Survival Distribution

Assumptions are required in order to draw inference about the marginal distribution of T based on a random sample of n independent patients (below, subscript i will denote data for the ith patient). It is typically assumed that censoring is non-informative. Mathematically, non-informative censoring corresponds to assuming, for all t,

$$d\Lambda(t) = d\Lambda^{\dagger}(t), \tag{0.2}$$

i.e., the net hazard of failure is equal to the cause-specific hazard of failure. If T and C are independent (i.e., independent censoring), then the non-informative assumption will hold. Unless there are secular trends in enrollment, censoring arising due to study termination should, in principle, be non-informative. Censoring due to premature drop-out, competing risks or treatment termination may be informative. We will discuss how to address this issue in Section 0.8. The utility of non-informative censoring is that it allows identification of $S(\cdot)$ since $d\Lambda^{\dagger}(\cdot)$ depends on the distribution of the observed data and $S(\cdot)$ can be computed from $d\Lambda(\cdot)$.

We can estimate $F^{\dagger}(t)$ by $\widehat{F}^{\dagger}(t) = N(t)/n$ and $S_X(t-)$ by $\widehat{S}_X(t-) = Y(t)/n$, where $N(t) = \sum_{i=1}^{n} I(X_i \leq t, \Delta_i = 1)$ is the called the counting process for failure and $Y(t) = \sum_{i=1}^{n} I(X_i \geq t)$ is called the "at-risk" process. Notice that $N(t)$ is a step function with jumps at the observed failure times (say, $t_1, \ldots, t_k$); the jump at a failure time t_j is $dN(t_j) = N(t_j) - N(t_j-)$. We estimate $d\Lambda(t)$, under non-informative censoring by

$$d\widehat{\Lambda}(t) = \frac{d\widehat{F}^{\dagger}(t)}{\widehat{S}_X(t-)} = \frac{dN(t)}{Y(t)}.$$

This estimator only takes positive values at the observed failure times; it is zero at all other times. Plugging this estimator for $d\Lambda(t)$ into the right hand side of (0.1), we obtain the Kaplan-Meier estimator [31]:

$$\widehat{S}(t) = \prod_{t_j \le t} \left\{ 1 - \frac{d\widehat{F}^{\dagger}(t_j)}{\widehat{S}_X(t_j-)} \right\} = \prod_{t_j \le t} \left\{ 1 - \frac{dN(t_j)}{Y(t_j)} \right\}.$$

In Figures 2(a)- 2(d), we display, for ACTG 320, the treatment-specific estimators of $F^{\dagger}(t)$, $S_X(t-)$, $\Lambda(t)$ and $S(t)$, respectively.

The estimated variance of $\widehat{S}(t)$ is given by Greenwood's formula [23]:

$$\widehat{Var}[\widehat{S}(t)] = \widehat{S}(t)^2 \prod_{t_j \le t} \left\{ \frac{dN(t_j)}{Y(t_j)(Y(t_j) - N(t_j))} \right\}.$$

This can be used to form a $(1-\alpha)\%$ point-wise confidence interval for $S(t)$. To deal with the fact that $S(t)$ is bounded between 0 and 1, it is recommended that one develop a confidence interval for $\log(-\log\{S(t)\})$ and then back-transform to a confidence interval for $S(t)$. Specifically, a confidence interval for $\log(-\log\{S(t)\})$ is of the form

$$\log(-\log\{\widehat{S}(t)\}) \pm z_{\alpha/2} \sqrt{\frac{\widehat{Var}[\widehat{S}(t)]}{(\widehat{S}(t)\log\{\widehat{S}(t)\})^2}},$$

where z_x is the $1 - x$ quantile of the standard normal distribution.

0.5 Hypothesis Testing

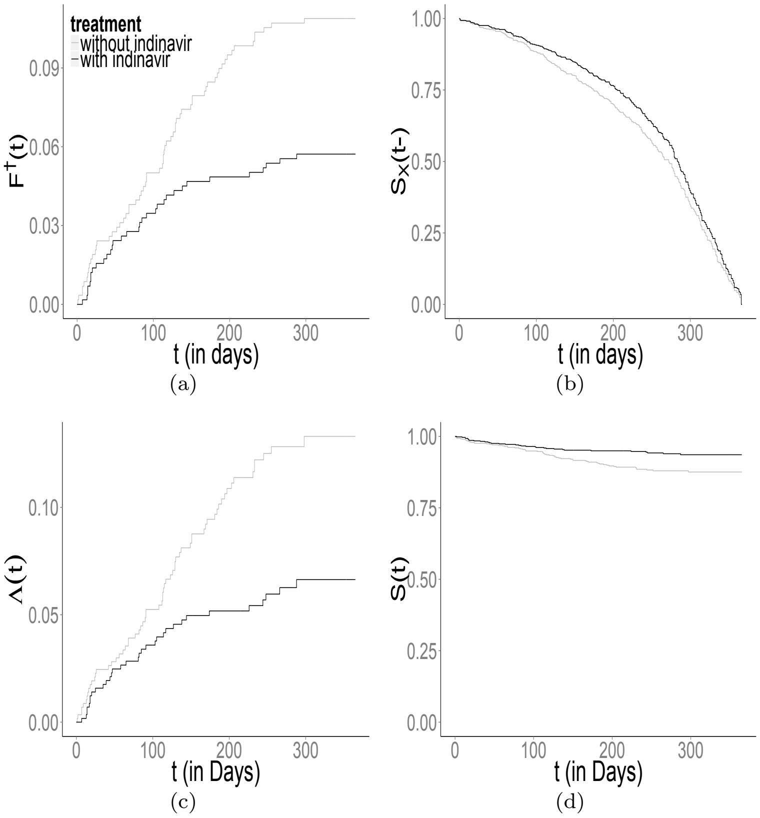
Suppose we are interested in comparing the survival curves of two randomized treatment groups. We assume non-informative censoring in both treatment groups. Let $S^{(0)}(t)$ ($\Lambda^{(0)}(t)$) and $S^{(1)}(t)$ ($\Lambda^{(1)}(t)$) denote the survival (cumulative hazard) functions for treatment groups 0 and 1, respectively. We wish to test the null hypothesis that $S^{(0)}(t) = S^{(1)}(t)$ for all t (or, equivalently $\Lambda^{(0)}(t) = \Lambda^{(1)}(t)$ for all t). Let $N^{(0)}(t)$ ($Y^{(0)}(t)$) and $N^{(1)}(t)$ ($Y^{(0)}(t)$) denote the counting process for failure (at-risk process) in groups 0 and 1, respectively. Let $N(t) = N^{(0)}(t) + N^{(1)}(t)$ and $Y(t) = Y^{(0)}(t) + Y^{(1)}(t)$. Let $t_1, \ldots, t_k$ be the observed failure times for both groups combined.

Consider the integrated weighted difference between the hazard functions, defined as

$$\beta(w) = \int w(t)\{d\Lambda^{(1)}(t) - d\Lambda^{(0)}(t)\},$$

where $w(t)$ is a non-negative weight function. Under the null hypothesis, $\beta(w)$ will be zero. If $d\Lambda^{(1)}(t) > d\Lambda^{(0)}(t)$ for all t, $\beta(w) > 0$ and if $d\Lambda^{(1)}(t) < d\Lambda^{(0)}(t)$ for all t, $\beta(w) < 0$. We can estimate $\beta(w)$ by

$$\widehat{\beta}(w) = \int w(t)\{d\widehat{\Lambda}^{(1)}(t) - d\widehat{\Lambda}^{(0)}(t)\} = \int w(t) \left\{ \frac{dN^{(1)}(t)}{Y^{(1)}(t)} - \frac{dN^{(0)}(t)}{Y^{(0)}(t)} \right\}.$$



(a) (b) (c) (d)

FIGURE 2
ACTG 320: Treatment-specific estimators of (a) $F^{\dagger}(t)$, (b) $S_X(t-)$, (c) $\Lambda(t)$ and (d) $S(t)$.

Interestingly, $\widehat{\beta}(w)$ can we re-written as:

$$\int k(t)\left\{dN^{(1)}(t)-\frac{Y^{(1)}(t)}{Y(t)}dN(t)\right\}=\sum_{t_j}k(t_j)\left\{dN^{(1)}(t_j)-\frac{Y^{(1)}(t_j)}{Y(t_j)}dN(t_j)\right\},$$

where $k(t)=w(t)\frac{Y(t)}{Y^{(1)}(t)Y^{(0)}(t)}$. Notice that the term in brackets is the typical "observed minus expected" quantity computed from a two-by-two table constructed based on the set of subjects who are at-risk for being observed to fail at or after time t_j (i.e., $\{i : X_i \geq t_j\}$), where the columns denote treatment assignment and the rows denote failure at time t_j. The two-by-two table has the following form:

Fail\Treatment	1	0	total
Yes	$dN^{(1)}(t_j)$	$dN^{(0)}(t_j)$	$dN(t_j)$
No	$Y^{(1)}(t_j)-dN^{(1)}(t_j)$	$Y^{(0)}(t_j)-dN^{(0)}(t_j)$	$Y(t_j)-dN(t_j)$
Total	$Y^{(1)}(t_j)$	$Y^{(0)}(t_j)$	$Y(t_j)$

In this table, the observed number of failures at time t_j for treatment 1 is $dN^{(1)}(t_j)$. Under the null hypothesis, the expected number of failures at time t_j for treatment 1 is $\frac{Y^{(1)}(t_j)}{Y(t_j)}dN(t_j)$, resulting in the term in brackets above. Thus, $\widehat{\beta}(w)$ is a weighted average of "observed-expected" terms from two-by-two tables constructed at each observed failure time.

The estimated variance of $\widehat{\beta}(w)$, under the null, is

$$\widehat{Var}[\widehat{\beta}(w)]=\sum_{t_j}w(t_j)^2\left\{\frac{Y(t_j)}{Y^{(1)}(t_j)Y^{(0)}(t_j)}\right\}\frac{dN(t_j)}{Y(t_j)}\frac{Y(t_j)-dN(t_j)}{Y(t_j)-1}.$$

Under the null,

$$T(w)=\frac{\widehat{\beta}(w)}{\sqrt{\widehat{Var}[\widehat{\beta}(w)]}}\approx N(0,1).$$

The null is rejected at the 0.05 level if $|T(w)|>1.96$.

With specific choices of $w(t)$, we can generate various test statistics that have been proposed for testing for treatment differences. For example, $w(t)=w_{LR}(t)=\frac{Y^{(1)}(t)Y^{(0)}(t)}{Y(t)}$ (or $k(t)=1$) yields the log-rank statistic, $w(t)=w_{GB}(t)=Y^{(1)}(t)Y^{(0)}(t)$ (or $k(t)=Y(t)$) yields the Gehan-Breslow statistic and $w(t)=w_{GW}(t)=\frac{Y^{(1)}(t)Y^{(0)}(t)}{Y(t)}\widehat{S}(t-)$ (or $k(t)=\widehat{S}(t-)$) yields the generalized Wilcoxon statistic, where $\widehat{S}$ is the Kaplan-Meier estimator of failure based on both treatment groups [18, 25, 43] .

In ACTG 320, the log-rank, Gehan-Breslow and generalized Wilcoxon statistics are -3.23, -3.09 and -3.20, respectively. The associated p-values are all less than 0.005, indicating a statistically significant treatment effect in favor of indinavir sulfate.

0.6 Cox Regression Model

Let $Z = (Z_1, \ldots, Z_k)$ be a k-dimensional vector of baseline covariates recorded on an individual. We assume non-informative censoring within levels of Z, i.e.,

$$d\Lambda(t|z) = d\Lambda^{\dagger}(t|z) \text{ for all } z,$$

where $d\Lambda(t|z)$ and $d\Lambda^{\dagger}(t|z)$ are the net and cause-specific hazards of failure for individuals with covariates $Z = z$.

In 1972, Cox [8] proposed the following regression model

$$\frac{d\Lambda(t|z)}{1 - d\Lambda(t|z)} = \frac{d\Lambda_0(t)}{1 - d\Lambda_0(t)} \exp\{\gamma^T z\}, \tag{0.3}$$

where γ is a k-dimensional vector of unknown parameters and $d\Lambda_0(t)$ is the so-called baseline hazard function as it represents the hazard for individuals with covariates $Z = 0$. In this model, the baseline function is left completely unspecified. At continuity points t, (0.3) reduces to

$$\lambda(t|z) = \lambda_0(t) \exp\{\gamma^T z\}; \tag{0.4}$$

at jump points t, (0.3) reduces to

$$\frac{P[T = t|T \geq t, Z = z]}{P[T > t|T \geq t, Z = z]} = \frac{P[T = t|T \geq t, Z = 0]}{P[T > t|T \geq t, Z = 0]} \exp\{\gamma^T z\}.$$

Here, $\exp\{\gamma_j\}$ quantifies the relative change in the risk associated with an increase of one unit in the covariate Z_j. For this model, the relative change is assumed to be the same throughout time. The parameter value $\gamma_j = 0$ corresponds to the case where the j^{th} covariate has no effect on survival. When $\gamma_j > 0$ (< 0), the risk of failure at any point in time increases (decreases) as Z_j increases.[2]

To estimate γ, Cox [8, 9] proposed the partial likelihood technique. The partial likelihood is constructed as the product of conditional likelihoods at each observed failure time t_j. Specifically, the contribution to the partial likelihood at t_j is the conditional likelihood of observing the individuals who actually failed at t_j given information just prior to t_j *and* that there are $d_j = \sum_{i=1}^{n} dN_i(t_j)$ individuals who fail at t_j (without specification of which individuals). Let $\mathcal{Q}_j$ be all subsets of d_j individuals from the set of

[2] It is important to note that the regression model that is usually specified (i.e., (0.4) holds for all t) is for an underlying failure time that is assumed to have a continuous distribution. Because of inexact measurement (e.g., failure time measured to the level of days), the distribution of the underlying *measurable* failure time (even in the absence of censoring) is discrete. In this chapter, we specify a regression model (i.e., (0.3)) for the underlying *measurable* failure time. The impact of this distinction is on (1) the interpretation of γ and (2) how ties are handled when drawing inference about γ.

$n_j = \sum_{i=1}^{n} Y_i(t_j)$ individuals at risk at time t_j; there are $\left(\begin{array}{c} n_j \\ d_j \end{array}\right)$ subsets in $\mathcal{Q}_j$. Let S_j be the sum of the covariate vectors for patients who are observed to fail at t_j. Let $S_{j,k}$ be the sum of the covariate vectors for patients in the kth subset in $\mathcal{Q}_j$. Then the conditional likelihood at t_j can be written as

$$L_{t_j}(\gamma) = \frac{\exp(\gamma^T S_j)}{\sum_{k \in \mathcal{Q}_j} \exp(\gamma^T S_{j,k})}. \tag{0.5}$$

The overall partial likelihood is $PL(\gamma) = \prod_{t_j} L_{t_j}(\gamma)$.[3] Since using this likelihood can be computationally intensive, various likelihood approximations have been proposed by [4, 12]. [4] and [12] replace the denominator in (0.5) by $\left(\begin{array}{c} n_j \\ d_j \end{array}\right) \{\frac{1}{n_j} \sum_{l \in \mathcal{R}_j} \exp(\gamma^T Z_l)\}^{d_j}$ and by $\prod_{k=0}^{d_j - 1} \{\sum_{l \in \mathcal{R}_j} \exp(\gamma^T Z_l) - \frac{k}{d_j} \sum_{l \in \mathcal{D}_j} \exp(\gamma^T Z_l)\}$, respectively, where $\mathcal{R}_j$ is set of individuals at risk at t_j and $\mathcal{D}_j$ is the set of individuals who fail at t_j.[4] These approximated partial likelihoods are implemented in all the major software packages (i.e., R, Stata, SAS). These approximations have been shown to perform well when the ratio of d_j to n_j is small for most t_j [13].

The score function associated with $PL(\gamma)$ (or one of its approximations) is

$$\mathcal{S}(\gamma) = \sum_{t_j} \left\{ S_j - \frac{A'_j(\gamma)}{A_j(\gamma)} \right\},$$

where $A_j(\gamma)$ equals $\sum_{k \in \mathcal{Q}_j} \exp(\gamma^T S_{j,k})$ (or one of its approximations) and $A'_j(\gamma)$ is the derivative of $A_j(\gamma)$ with respect to γ. In the special case of Breslow's approximation, it can be shown that

$$\mathcal{S}(\gamma) = \sum_{i=1}^{n} \int \left\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(t) Z_j \exp(\gamma^T Z_j)}{\sum_{j=1}^{n} Y_j(t) \exp(\gamma^T Z_j)} \right\} dN_i(t). \tag{0.6}$$

The parameter γ is estimated as the maximizer, $\widehat{\gamma}$, of $PL(\gamma)$ (or one of its approximations). The maximizer is found by solving $\mathcal{S}(\gamma) = 0$. The estimator $\widehat{\gamma}$ will be approximately normal with mean γ and the inverse of the Hessian (matrix of second dervivatives) of log of $PL(\gamma)$ (or one its approximations) evaluated at $\widehat{\gamma}$. The variance of $\widehat{\gamma}_j$ is estimated by the jth diagonal component of the inverse of the aforementioned Hessian matrix. A 95% confidence interval for γ_j can be computed as $\widehat{\gamma}_j \pm 1.96\sqrt{\widehat{Var}[\widehat{\gamma}_j]}$. The null hypothesis that $\gamma_j = 0$

[3] Under the model for assuming the underlying failure time has a continuous distribution (i.e., (0.4) holds for all t), $L_{t_j}(\gamma)$ would be computed differently. Specifically, in the presence of ties at t_j, the conditional likelihood at that time needs to incorporate all the possible ways of "untying" the tied failure times.

[4] The resulting approximated partial likelihoods are identical to the approximated partial likelihoods that are used when it is assumed that the underlying failure time has a continuous distribution.

versus the alternative that $\gamma_j \neq 0$ can be tested at 0.05 type I error level by rejecting the null if the 95% confidence interval does not contain zero.

In the special case where Z is the indicator of treatment of assignment, γ quantifies the log relative risk of failure between the two treatments. In this setting, it is interesting to note that

- $\mathcal{S}(0)$ in (0.6) reduces to $\widehat{\beta}(w_{LR})$
- Testing whether $\gamma = 0$ using the approach above is equivalent, in large samples, to testing for a treatment difference using the logrank test statistic.

Let $h_0(t) = \frac{d\Lambda_0(t)}{1-d\Lambda_0(t)}$. A profile likelihood estimator of $h_0(t)$ puts mass only at failure times t_j. Let $h_j = h_0(t_j)$. The profile estimate for h_j is the unique non-negative solution, $\widehat{h}_j$, to the following equation:

$$\sum_{i \in \mathcal{R}_j} \frac{h_j \exp(\widehat{\gamma} Z_i)}{1 + h_j \exp(\widehat{\gamma} Z_i)} = d_j. \tag{0.7}$$

Further, an estimator of the conditional survivor function of T given $Z = z$, $S(t|z)$, is

$$\widehat{S}(t|z) = \prod_{t_j \leq t} \left\{ \frac{1}{1 + \widehat{h}_j \exp(\widehat{\gamma} z)} \right\}. \tag{0.8}$$

Above, we have focused on modeling the risk of failure as a function of covariates that do not depend on time. In many studies with time-to-event endpoints, the covariates of interest may also change with time. Such covariates are referred to as time-dependent covariates. In ACTG 320, for example, patients were clinically evaluated at multiple occasions after enrollment; at these evaluations, CD4 counts were measured. In evaluating whether CD4 is a potential surrogate marker for the development of AIDS or death (i.e., failure), it is natural to ask how the risk of failure at a given time t relates to the history of CD4 counts prior to time t.

Let $Z(t) = (Z_1(t), \ldots, Z_l(t))$ be a l-dimensional vector of covariates that is known at time t. Let $\overline{Z}(t)$ be the history of these covariates through time t, i.e., $\overline{Z}(t) = \{Z(u) : 0 \leq u \leq t\}$. A covariate that does not vary with time can be considered as a special case of a time-varying covariate. We assume non-informative censoring within covariate histories, i.e.,

$$d\Lambda(t|\overline{z}(t)) = d\Lambda^{\dagger}(t|\overline{z}(t)) \text{ for all } \overline{z}(t),$$

where $d\Lambda(t|\overline{z}(t))$ and $d\Lambda^{\dagger}(t|\overline{z}(t))$ are the net and cause-specific hazards of failure for individuals with covariate history $\overline{Z}(t) = \overline{z}(t)$.

The Cox regression model posits that

$$\frac{d\Lambda(t|\overline{z}(t))}{1 - d\Lambda(t|\overline{z}(t))} = \frac{d\Lambda_0(t)}{1 - d\Lambda_0(t)} \exp\{\gamma^T g(t, \overline{z}(t))\}, \tag{0.9}$$

where $g(t, \overline{z}(t))$ is a k-dimensional known function of t and $\overline{z}(t)$, γ is a k-dimensional vector of unknown parameters and $d\Lambda_0(t)$ is the baseline hazard function. In the ACTG 320 example, suppose $l = 1$, $Z_1(t)$ is the most recently recorded CD4 at or prior to time t. We might consider $g(t, \overline{Z}(t)) = Z_1(t)$ in which γ represents the common (over time) change in the risk of failure at time t per unit increase in the CD4 count known at that time. Or, we might consider $g(t, \overline{Z}(t)) = (Z_1(t)I(t \leq 56), Z_1(t)I(56 < t \leq 112), \ldots, Z_1(t)I(168 < t \leq 224), Z_1(t)I(t > 224))$, in which case, the effect of CD4 count is allowed to vary over time according to how the time-axis is partitioned.

Estimation of γ and $h_0(t)$ proceeds as above. It is important to emphasize that, in the case of time-independent covariates, it makes sense to estimate $S(t|z)$ as above. In the case of time-dependent covariates, it may not make sense to estimate $S(t|\overline{z}(t)) = P[T > t|\overline{Z}(t) = \overline{z}(t)]$. This is because the very fact that $\overline{Z}(t)$ is measured can imply that the patients are alive or event-free. To distinguish settings where it makes sense to estimate $S(t|\overline{z}(t))$, it is important to distinguish between internal and external covariates. An external covariate is one that can affect an individual, but can be measured even if the individual is not on study, e.g., air pollution levels. In contrast, an internal covariate is one in which the change in the covariate depends on the individual, e.g., CD4 count.

In ACTG 320, investigators were interested in evaluating whether the effect of treatment varied by baseline CD4 status. Specifically, they wanted to know whether the treatment effect was different for patients with baseline CD4 less than or equal to 50 as compared to patients with baseline CD4 between 51 and 200. For each of these CD4 strata, we can fit a Cox regression model with a single treatment-indicator covariate (1 for indinavir, 0 otherwise) and then compare the strata-specific treatment effect estimators using a Wald-type test statistic. The estimated log relative risk in patients with low and high baseline CD4 are -0.67 (standard error = 0.24) and -0.70 (standard error = 0.27), respectively. The Wald test statistic for the difference in these log relative risks is 0.099, with an associated p-value of 0.92. Thus, there is no statistically significant evidence of effect modification by baseline CD4.

0.7 Sample Size Calculations

In designing a two-arm randomized trial with a survival endpoint, it is important to define clinically meaningful alternative hypotheses. A useful way to define alternative hypotheses is through the proportional hazards assumption. Specifically, it is assumed that the underlying distribution of survival in the two treatment arms is continuous (i.e., no ties) and

$$\frac{d\Lambda^{(1)}(t)}{d\Lambda^{(0)}(t)} = \exp\{\gamma\}.$$

Hazard Ratio	Number of Failures
2.00	88
1.50	256
1.25	844
1.10	4623

TABLE 0.1
Number of failures required for various hazard ratios: 90% power, 5% two-sided type I error.

Here, $\gamma > 0$ (< 0) implies that subjects assigned to treatment 1 have worse (better) survival and $\gamma = 0$ implies the null hypothesis.

The reason statisticians focus on proportional hazards alternatives is that, in addition to having a "nice" interpretation, theory has been developed that shows that the logrank test is the most powerful nonparametric test to detect these alternatives. It can be shown, under simplifying assumptions, that

$$T(w_{LR}) \approx N(\gamma\sqrt{d \cdot p(1-p)}, 1),$$

where p is the randomization proportion and $d = \sum_j d_j$ is the total number of failures. If $p = 0.5$, then a level α two-sided test of the null hypothesis will have power $1 - \delta$ to detect the porportional hazards alternative γ_A when the number of failures equals $4\left\{\frac{(z_{\alpha/2}+z_\delta)}{\gamma_A}\right\}^2$.

Some examples of the number of failures necessary to detect an alternative where the hazard ratio equals $\exp(\gamma_A)$ with 90% power using the logrank test at the 0.05 (two-sided) level of significance is given in the Table 0.1.

During the design stage it must be ensured that a sufficient number of patients are entered into the trial and followed long enough so that the requisite number of events are attained. Arbitrarily picking a number of patients and waiting for the requisite number of events to occur will not be adequate for the proper planning of the trial. The design of a clinical trial with a time to event endpoint requires the following elements:

- number of patients (n)
- accrual period (A): calendar period that patients are entering the study
- follow-up time (F): calendar period after accrual has ended and the final analysis is conducted

Consider the situation where the treatment-specific hazard functions are constant. Specifically, assume (1) $d\Lambda_0(t) = \lambda_0 dt$ and (2) $d\Lambda_1(t) = \lambda_1 dt$ with $\lambda_1 = \lambda_0 \exp(\gamma_A)$. Further, assume (3) censoring results only from staggered entry into the trial, (4) a constant enrollment rate a per year and (5) 1:1

randomization ratio. Under these assumptions, it can be shown that the expected number of patients enrolled is Aa and the expected number of observed failures in treatment group j is

$$\frac{a}{2}\left[A - \frac{\exp(-\lambda_j L)}{\lambda_j}\{\exp(\lambda_j A) - 1\}\right],$$

where $L = A + F$.

Suppose that the probability of survival in treatment 0 at 18 months is 67%; then $\lambda_0 = 0.022$. Further, suppose it is desired to detect an increase in survival in treatment 1 at 18 months to 75% with 90% power using a two-sided logrank test at the 0.05 type I error level. Then, $\lambda_1 = 0.016$ and $\gamma_A = \log(0.72)$. Using the above results, the total number of required failures is 384. To achieve this number of failures, a, A and F must be chosen so that

$$aA - \frac{a}{2}\frac{\exp(-\lambda_0 L)}{\lambda_0}\{\exp(\lambda_0 A) - 1\} - \frac{a}{2}\frac{\exp(-\lambda_1 L)}{\lambda_1}\{\exp(\lambda_1 A) - 1\} = 384.$$

Suppose that the trial is scheduled to have an accrual period of 36 weeks and that the last enrolled patient will be followed for 48 weeks, i.e., $A = 36$ weeks and $F = 48$ weeks. Using the above formula, the accrual rate must be 40 patients per week, yielding a total enrollment of 1440 patients.

Other factors that may affect power include premature loss to follow-up, competing risks and non-compliance. An excellent account on how to deal with these issues during the design stage is given by [33].

0.8 Informative Censoring

The methods described above for estimation of $S(t)$ and testing for differences in survival between treatment groups rely on the assumption of non-informative censoring, i.e., Equation (0.2) holds for each treatment group. It can be shown that Assumption (0.2) is equivalent to assuming, for t and $t' > t$,

$$d\Lambda_C^{\dagger}(t|T = t') = d\Lambda_C^{\dagger}(t|T > t), \qquad (0.10)$$

where $d\Lambda_C^{\dagger}(t|T = t')$ is the cause-specific hazard of censoring at t given information that failure occurs at time t' and $d\Lambda_C^{\dagger}(t|T > t)$ is the cause-specific hazard of censoring at t given that failure occurs after t (not when it occurs).

Robins and Finkelstein [38] developed a method that relaxes this assumption. Specifically, they assume that

$$d\Lambda_C^{\dagger}(t|\overline{z}(t), T = t') = d\Lambda_C^{\dagger}(t|\overline{z}(t), T > t), \qquad (0.11)$$

where $d\Lambda_C^{\dagger}(t|\overline{z}(t), T = t')$ is the cause-specific hazard of censoring at t

given information on $\overline{Z}(t) = \overline{z}(t)$ and that failure occurs at time t' and $d\Lambda^{\dagger}_{C}(t|\overline{z}(t), T > t)$ is the cause-specific hazard of censoring at t given information on $\overline{Z}(t) = \overline{z}(t)$ and that failure occurs after t (not when it occurs). They further assume a dimension-reduction model for $d\Lambda^{\dagger}_{C}(t|\overline{z}(t), T > t)$. Here, we consider a model of the form:

$$\frac{d\Lambda^{\dagger}_{C}(t|\overline{z}(t), T > t)}{1 - d\Lambda^{\dagger}_{C}(t|\overline{z}(t), T > t)} = \frac{d\Lambda^{\dagger}_{C,0}(t)}{1 - d\Lambda^{\dagger}_{C,0}(t)} \exp\{\eta^T g(t, \overline{z}(t))\}, \tag{0.12}$$

where $g(t, \overline{z}(t))$ is a k-dimensional known function of t and $\overline{z}(t)$, η is a k-dimensional vector of unknown parameters and $d\Lambda^{\dagger}_{C,0}(t)$ is the baseline hazard function. To estimate η and $h^{\dagger}_{C,0}(t) = \frac{d\Lambda^{\dagger}_{C,0}(t)}{1-d\Lambda^{\dagger}_{C,0}(t)}$, use the techniques described in Section 0.6 above with the following modifications: (1) reverse the roles of T and C and (2) break ties between failure and censoring times by assuming that failure preceeds censoring. Let

$$\tilde{K}_i(t; \tilde{\eta}) = \prod_{c_k \le t} \left\{ \frac{1}{1 + \tilde{h}^{\dagger}_{C,0}(c_k; \tilde{\eta}) \exp\{\tilde{\eta}^T g(t_k, \overline{Z}_i(c_k))\}} \right\},$$

where c_k are the unique ordered censoring times, $\tilde{\eta}$ is an estimator of η in Model (0.12), $\tilde{h}^{\dagger}_{C,0}(t; \tilde{\eta})$ is the corresponding profile estimator of $h^{\dagger}_{C,0}(t)$, and $\overline{Z}_i(t)$ is the time varying covariate vector associated with subject i. Here, $\tilde{K}_i(t; \tilde{\eta})$ is an estimator of the probability that subject i is uncensored at time t. The adjusted survival curve of [38] is of the form:

$$\tilde{S}(t) = \prod_{t_j \le t} \left\{ 1 - \frac{\sum_{i=1}^{n} \left(dN_i(t_j) / \tilde{K}_i(t_j; \tilde{\eta}) \right)}{\sum_{i=1}^{n} \left(Y_i(t_j) / \tilde{K}_i(t_j; \tilde{\eta}) \right)} \right\},$$

where t_j are the unique ordered failure times.

The intuition of the adjusted estimator is as follows. The numerator and denominator of the ratio inside the curly brackets estimates, in the absence of censoring, the number of subjects who are expected to fail at t_j and the number of subjects expected to be at risk for failure at t_j, respectively. Why? In the numerator (denominator), each subject who fails (is at risk) at t_j is inverse weighted by the probability of being uncensored at that time. Inverse weighting (a technique derived from the survey sampling literature) serves to upweight the contribution of subjects with observed data to account for themselves and others like them who were unobserved. For example, if the probability of being uncensored at time t_j for a subject is 0.25, then he/she accounts for him/her-self plus three other similar subjects who were unobserved at that time. In the absence of censoring, the ratio in the curly brackets estimates the hazard of failure at t_j and one minus the ratio estimates the probability of being event free at t_j given at risk at t_j. Therefore, the product of the terms in curly brackets through time t estimates the probability of being event free at time t. It is important to note that if $\tilde{\gamma} = 0$ (i.e., $\overline{Z}(t)$ is not prognostic for censoring), the survival curve estimator of [38] reduces to the Kaplan-Meier estimator.

Robins and Finkelstein [38] also showed how to construct an adjusted estimator of γ in model (0.3) with treatment assignment indicator as the sole covariate by modifying $\mathcal{S}(\gamma)$ in (0.6) to ensure that it has mean zero (in large samples) when (0.11) and (0.12) are assumed for each treatment group *and* reducs to (0.6) when the covariates in model (0.12) are not prognostic for censoring. Towards this end, [38] defined the modified score function:

$$\tilde{\mathcal{S}}(\gamma) = \sum_{i=1}^{n} \int \left\{ Z_i - \frac{\sum_{j=1}^{n} \tilde{W}_j(t) Y_j(t) Z_j \exp(\gamma^T Z_j)}{\sum_{j=1}^{n} \tilde{W}_j(t) Y_j(t) \exp(\gamma^T Z_j)} \right\} \tilde{W}_i(t) dN_i(t),$$

where

$$\tilde{W}_i(t) = \frac{Z_i \tilde{K}_i^{(1)}(t;0) + (1 - Z_i)\tilde{K}_i^{(0)}(t;0)}{Z_i \tilde{K}_i^{(1)}(t;\tilde{\eta}^{(1)}) + (1 - Z_i)\tilde{K}_i^{(0)}(t;\tilde{\eta}^{(0)})},$$

the superscripts are used to reference treatment groups, and $\tilde{K}_i^{(z)}(t;0)$ are the treatment-specific Kaplan-Meier estimators of the survival curve for censoring (with ties broken as above). The adjusted estimator $\tilde{\gamma}$ of γ is the solution to $\tilde{\mathcal{S}}(\gamma) = 0$. [38] showed that $\tilde{\gamma}$ will be normally distributed in large samples. While it is possible to compute an analytic expression for the standard error of $\tilde{\gamma}$, it is easier to use bootstrapping procedures to estimate standard errors for $\hat{\gamma}$ and construct confidence intervals for γ.

To illustrate this method, we use data from ACTG 320 where we further censor patients at discontinuation of their assigned treatment if it occurs prior to failure. That is, we use the earlier of the open and solid circles in Figure 1. In the analyses discussed above, there were 96 failures and 1054 censored observations. Incorporating censoring at treatment stop reduces the number of failures to 66 (43 without indinavir, 23 with indinavir) and increases the number of censored observations to 1084. Specifically, 30 of the 96 failures occurred after treatment stop - 20 from the without indinavir arm and 10 from the indinavor arm. In addition, 207 of the original 1054 censored observations are moved to an earlier censoring time - 147 from the without indinavir arm and 60 from the indinavor arm.

It is plausible that such censoring may be informative as patients who are sicker or experiencing side effects may be more likely to discontinue their assigned therapy. In this analysis, we seek to estimate the effect of treatment in a world without non-compliance. This contrasts with the aforementioned analyses which was focused on estimating what is often referred to as the intention-to-treat effect.

For each treatment group, we fit Model (0.12) with Karnofsky score and hemoglobin at baseline as time-independent covariates and CD4 as a time-varying covariate. We used Breslow's approximation to the partial likelihood;

	Without Indinavir		With Indinavir	
Covariate	Effect	95% CI	Effect	95% CI
CD4	0.9997	[0.9987, 1.001]	0.9984	[0.9976, 0.9992]
Karnofsky Score	0.9926	[0.9820, 1.003]	1.0075	[0.9961, 1.0190]
Hemoglobin	1.0290	[0.7388, 1.433]	1.4818	[0.9706, 2.2622]

TABLE 0.2
Treatment-specific censoring model fits: exponentiated regression coefficients and associated 95% confidence intervals.

the results based on Efron's approximation were similar. Table 0.2 displays the exponentiated regression coefficients and associated 95% confidence intervals from fitting these models. The table shows that patients with higher CD4 (i.e., healthier) are less likely to be censored, although the effect is only statistically significant (at the 5% level) for patients assigned to the indinavir arm. The effects of baseline Karnofsky score and hemoglobin are not statistically significant. This result suggests that a non-informative censoring analysis may be optimistic. Figure 0.8 shows the unadjusted (i.e., Kaplan-Meier) and adjusted estimated survival curves. As expected, the figure shows that the adjusted curves tend to be lower than the unadjusted curves, although the shifts are negligible. The unadjusted and adjusted estimates of the relative risk are 0.49 (95% CI: 0.30-0.82) and 0.46 (95% CI: 0.25-0.77), respectively. In contrast, the estimate the relative risk in the previous analysis (i.e., not censoring at treatment stop) was 0.51 (95% CI: 0.33-0.77). The estimate of the effect of indinavir under full compliance is slightly larger, although it does not appear to be clinically significant.

Scharfstein and Robins [40] developed methods for evaluating the sensitivity of survival curve estimation to deviations from Assumption (0.11). Rotnitzky *et al.*[39] extended the ideas of [40] to address competing causes of censoring (e.g., end-of-study censoring vs. censoring due to treatment stop). Zhang *et al.*[48] used the ideas of [38] to draw inference, in the presence of non-compliance, about the distribution of a time-to-event for treatment regimes with specified treatment stops.

0.9 Conclusion

In this chapter, we discussed the most commonly used survival analysis methods (survival curve estimation, treatment group comparsions, Cox regression, sample size calculation) in clinical trials. We also reviewed a method for adjusting for informative censoring that we think should be more widely utilized. We emphasized the discrete Cox regression model because it has been our ex-

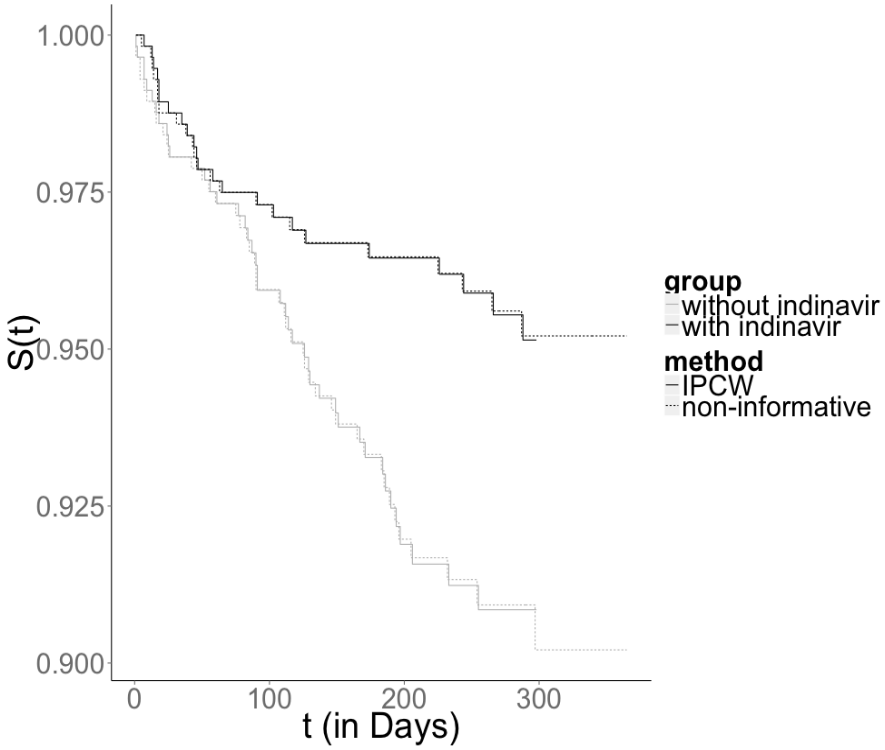


FIGURE 3
Treatment-specific unadjusted (Kaplan-Meier) and adjusted (Robins and Finkestein) estimated survival curves.

perience that event times are usually measured inexactly (e.g., at the level of days rather than hours or seconds). Such inexactness renders the events times to have discrete support. Importantly, the partial likelihood approximations under a discrete Cox model are identical to the approximations for handling ties in the continuous time Cox model. Thus, the estimators resulting from maximizing these approximated partial likelihoods can be considered as estimating the regression parameter in either the discrete or continuous time Cox model.

There is a wide body of survival analysis methods that we have not discussed. While we focused on time to event data that may be right censored, there are methods that handle event times that may also be interval censored (i.e., only known to fall into a finite time interval) [45, 16, 17, 21, 34, 20, 15, 14, 42, 46]. There is also a great deal of work on alternative regression models, including the accelerated failure time model [47, 37, 3, 29] and the semi-parametric transformation model [6, 5].

The issue of competing risks, whereby subjects are at-risk for multiple *pre-emptive* causes of failure, is particularly challenging. If an individual is observed to fail from one cause then he/she is no longer at risk for failure from another cause. Thus, when analyzing failure due to a given cause, it is not appropriate to simply consider individuals who failed due to another cause as censored observations. This is why many analysts often work with a composite endpoint which is the time of the first failure regardless of cause. Alternatively, some analysts report cause-specific hazards [36] or cause-specific sub-distribution functions [22].

Another important area is multivariate survival analysis, where multiple failure events are to be recorded on each subject (either in series or in parallel) or a single failure event on subjects who are themselves clustered into groups. Methods are available that are similar in spirit to the marginal, copula and random effects models using in longitudinal data analysis, with the exception that in survival analysis random effects models are often referred to as frailty models. [26] provides a detailed review of methods for analyzing multivariate survival data.

Methods are also available for the design and analysis of clinical trials in which time-to-event data are to be analyzed at interim time points at which a decision can be made to prematurely stop the trial for efficacy or futility. Scharfstein, Tsiatis and Robins [41] and Jennison and Turnbull[27] developed a general framework, based on the concept of statistical information, for designing and monitoring such trials in which a type I error spending function (developed by [11]) along with a stopping boundary is utilized to preserve the overall operating characteristics of the trial. Jennison and Turnbull [28] is a great resource to learn more about what is often called group sequential clinical trials.

Survival analysis is a very well researched field. There are great reference books available, including but not limited to [32, 19, 44, 30, 2, 10]. There are also great software routines available in SAS [1], R [35] and STATA [7] for analyzing survival data. There is a great deal of online material demonstrating how to use these routines.

Bibliography

[1] Paul D Allison. *Survival analysis using SAS: a practical guide.* Sas Institute, 2010.

[2] Per Kragh Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes.* Springer Science & Business Media, 2012.

[3] Rebecca A Betensky, Daniel Rabinowitz, and Anastasios A Tsiatis. Computationally simple accelerated failure time regression for interval censored data. *Biometrika*, 88(3):703–711, 2001.

[4] Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

[5] Kani Chen, Zhezhen Jin, and Zhiliang Ying. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668, 2002.

[6] SC Cheng, LJ Wei, and Z Ying. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995.

[7] Mario Cleves. *An introduction to survival analysis using Stata.* Stata Press, 2008.

[8] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

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

[10] David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

[11] David L Demets and KK Lan. Interim analysis: the alpha spending function approach. *Statistics in medicine*, 13(13-14):1341–1352, 1994.

[12] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

[13] VT Farewell and Ross L Prentice. The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, 67(2):273–278, 1980.

[14] Michael P Fay. Comparing several score tests for interval censored data. *Statistics in Medicine*, 18(3):273–285, 1999.

[15] Michael P Fay and Pamela A Shaw. Exact and asymptotic weighted logrank tests for interval censored data: the interval r package. *Journal of Statistical Software*, 36(2), 2010.

[16] Dianne M Finkelstein. A proportional hazards model for interval-censored failure time data. *Biometrics*, pages 845–854, 1986.

[17] Dianne M Finkelstein and Robert A Wolfe. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, pages 933–945, 1985.

[18] Thomas R Fleming and David P Harrington. A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics-Theory and Methods*, 10(8):763–794, 1981.

[19] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

[20] Els Goetghebeur and Louise Ryan. Semiparametric regression analysis of interval-censored data. *Biometrics*, 56(4):1139–1144, 2000.

[21] William B Goggins, Dianne M Finkelstein, David A Schoenfeld, and Alan M Zaslavsky. A markov chain monte carlo em algorithm for analyzing interval-censored data under the cox proportional hazards model. *Biometrics*, pages 1498–1507, 1998.

[22] Robert J Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, pages 1141–1154, 1988.

[23] Major Greenwood et al. A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects. Ministry of Health*, (33), 1926.

[24] Scott M Hammer, Kathleen E Squires, Michael D Hughes, Janet M Grimes, Lisa M Demeter, Judith S Currier, Joseph J Eron Jr, Judith E Feinberg, Henry H Balfour Jr, Lawrence R Deyton, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11):725–733, 1997.

[25] David P Harrington and Thomas R Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

[26] Philip Hougaard. *Analysis of multivariate survival data.* Springer Science & Business Media, 2012.

[27] Christopher Jennison and Bruce W Turnbull. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440):1330–1341, 1997.

[28] Christopher Jennison and Bruce W Turnbull. *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

[29] Zhezhen Jin, DY Lin, LJ Wei, and Zhiliang Ying. Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353, 2003.

[30] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[31] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[32] David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 1996.

[33] Edward Lakatos. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, pages 229–241, 1988.

[34] Jane C Lindsey and Louise M Ryan. Methods for interval-censored data. *Statistics in medicine*, 17(2):219–238, 1998.

[35] Melinda Mills. *Introducing survival and event history analysis*. Sage Publications, 2011.

[36] Ross L Prentice, John D Kalbfleisch, Arthur V Peterson Jr, Nancy Flournoy, VT Farewell, and NE Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554, 1978.

[37] James Robins and Anastasios A Tsiatis. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*, 79(2):311–319, 1992.

[38] James M Robins and Dianne M Finkelstein. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, pages 779–788, 2000.

[39] Andrea Rotnitzky, Andres Farall, Andrea Bergesio, and Daniel Scharfstein. Analysis of failure time data under competing censoring mechanisms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):307–327, 2007.

[40] Daniel O Scharfstein and James M Robins. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634, 2002.

[41] Daniel O Scharfstein, Anastasios A Tsiatis, and James M Robins. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association*, 92(440):1342–1350, 1997.

[42] Jianguo Sun. *The statistical analysis of interval-censored failure time data*. Springer Science & Business Media, 2007.

[43] Robert E Tarone and James Ware. On distribution-free tests for equality of survival distributions. *Biometrika*, 64(1):156–160, 1977.

[44] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2000.

[45] Bruce W Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

[46] Lianming Wang, Christopher S McMahan, Michael G Hudgens, and Zaina P Qureshi. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, 2015.

[47] LJ Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

[48] Min Zhang, Anastasios A Tsiatis, Marie Davidian, Karen S Pieper, and Kenneth W Mahaffey. Inference on treatment effects from a randomized clinical trial in the presence of premature treatment discontinuation: the synergy trial. *Biostatistics*, 12(2):258–269, 2011.