Submitted: Date 01/08/2026
Accepted: Editor Use
Published: Editor Use

Review Article

[1]Department of Mathematics and Statistics, McGill University, Montréal, QC, Canada

Keywords

Biostatistics, Survival Analysis, Kaplan-Meier Estimator, Proportional Hazard Model, Log-rank Test

Email Correspondence

jiajun.zhang@mail.mcgill.ca

https://doi.org/10.26443/msurj.vxxxx

Jiajun Zhang[1]

Statistical Inference for Survival Data

Abstract

Survival analysis studies the behavior of time-to-event data, which arise naturally in modern statistical studies and are characterized by incomplete observations due to censoring and truncation. These unique features invalidate classical statistical methods and require the development of specialized inferential tools. This article presents an introduction to basic theories and methodologies used for modeling time-to-event data, which are widely used in biostatistical and epidemiological research. We begin by introducing the censoring and truncation mechanism, followed by defining basic quantities, including survival and hazard functions. We next study the key estimating techniques for survival data based on parametric, non-parametric, and semi-parametric inferences. We briefly introduce likelihood inference for parametric inference, life table methods, the Kaplan-Meier (KM) estimator and its variants for non-parametric inference, and Cox's proportional hazard (PH) model for semi-parametric inference. Finally, we introduce the log-rank test, which is commonly used for testing statistical significance in the survival data among multiple groups.

Introduction

Time-to-event data often presents characteristic features known as censoring and truncation. Censoring occurs when the observation is incompletely determined for some subjects, and truncation occurs when only a group of chosen individuals are observed and the rest are systematically excluded from the study.

There are three main types of censoring: Right censoring, left censoring, and interval censoring. Right censoring happens when we only know that the event happens after a certain time threshold; Left censoring happens when we only know that the event time happens before a certain time threshold; Interval censoring occurs when we only know that the event time happens during a certain time period. Below, we will use three examples taken from Klein and Moeschberger[1] to illustrate different types of censoring.

Example 1. During 1982-1992, 863 patients underwent kidney transplantation at the Ohio state university transplant center, where researchers investigated the survival time of those patients after transplantation. Their maximum follow-up time was 9.47 years (Right censoring).

Example 2. In 1978, Turnbull and Weiss studied the first usage of marijuana among California high school boys, and some answered with "I used it before, but I cannot remember when" (Left censoring).

Example 3. The National Longitudinal Survey of Youth investigated a random sample of females aged 14 to 21 yearly from 1979 to 1988. Females in the survey were asked about any pregnancies that have occurred since they were last interviewed (Interval censoring).

Truncation, on the other hand, can be viewed as a design issue where a group of individuals from the target population is systematically excluded from the sampling pool. This feature becomes obvious in survival analysis because due to cost we do not have access to all the individuals. For those individuals who died before entering the study, we do not have any access to their exact death time, we say those individuals were left truncated.

Basic Quantities of Survival Data

Survival Function and Hazard Function

Suppose we have drawn a random sample (that is, the life time of those individuals are mutually independent and identically distributed) of n individuals. We will use X_i as the random variable denoting the life time of the ith individual. We use $f(x)$ as the probability density function (pdf) of X_i, $F(x)$ as the cumulative distribution function (cdf) of X_i. We first define the **survival function** $S(x)$:

Definition 1. Let X be a non-negative random variable which denotes the life time, let $F(x)$ be the cumulative distribution function (cdf) of X, then the survival function is defined as

$$S(x) = \mathbb{P}(X \geq x) = 1 - F(x) = 1 - \int_0^x f(t)dt.$$

The survival function gives the probability that a subject will survive beyond the indicated time $t = x$. By the uniqueness of a distribution function, the survival function also uniquely defines a distribution. Another characteristic is called the hazard function $h(x)$, which is the instantaneous rate of death at time right after $t = x$ given that the individual is still alive at $t = x$.

Definition 2. Let X be a non-negative random variable that denotes the life time, then the hazard function $h(x)$ is defined as

$$h(x) = \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X \leq x + dx \big| X \geq x)}{dx}.$$

From Definition 2, we can further derive the hazard function as the ratio of

the probability density function (pdf) and the survival function:

$$\begin{aligned} h(x) &= \lim_{dx\to 0} \frac{\mathbb{P}((x \le X \le x+dx) \bigcap (X \ge x))}{\mathbb{P}(X \ge x)dx} \\ &= \lim_{dx\to 0} \frac{\mathbb{P}(x \le X \le x+dx)}{[1-F(x)]dx} \\ &= \lim_{dx\to 0} \frac{F(x+dx)-F(x)}{[1-F(x)]dx} \\ &= \frac{F'(x)}{1-F(x)} = \frac{f(x)}{S(x)}. \end{aligned}$$

In addition, if we apply the Chain rule, we have

$$h(x) = -\frac{d}{dx}\log(S(x)),$$

by integrating both sides, we have

$$H(x) = \int h(x)dx = -\log S(x), \tag{1}$$

where the integral term is denoted as the **cumulative hazard function**.

Parametric Inferences with Censoring and Truncation

After we have selected our random sample, the goal is to find the correct probability distribution for the survival function $S(x)$ or the hazard function $h(x)$, so we are able to make further inferences on it. In statistics, we have three different types of inferences on the random sample: Parametric, non-parametric, and semi-parametric. Their characteristics are as follow.

- Parametric Inference: The distribution of the sample is unknown up to finitely many parameters. For example, we know X forms a exponential distribution, but with parameter θ unknown. The methodologies are the easiest but less common in practice.
- Non-parametric Inference: We do not have any information on the distribution of the sample, and is very common in practice.
- Semi-Parametric Inference: This combines the parametric and non-parametric inference, where the distribution of the sample consists a parametric part and a non-parametric part.

In this section we will focus on parametric inference. We will discuss non-parametric and semi-parametric inferences in section 3 and 4.

Without any censoring or truncation, our task would be a straightforward parametric estimation problem where maximum likelihood estimation (MLE) method is widely used. With censoring and truncation introduced, we shall take them into consideration, and the goal of this section is to construct the corresponding inferences. We first introduce the likelihood function:

Definition 3. Suppose we have a random sample of $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} f(x,\theta)$ where f is the probability density function (pdf) of the sample, θ is the unknown parameter of interest. The **likelihood function** $\mathcal{L}(\theta)$ is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i,\theta).$$

Once we have obtained the likelihood function, the **maximum likelihood estimate** $\hat{\theta}$ of θ is obtained by maximizing $\mathcal{L}(\theta)$ as a function of θ. We first consider the presence of (right) censoring: Suppose for each individual in the sample, X_i is the true life time, C_i is the censoring time independent of X_i. We observe (T_i, δ_i), where $T_i = \min\{X_i, C_i\}$, and

$$\delta_i = \begin{cases} 1 & \text{if } X_i \le C_i \\ 0 & \text{if } X_i > C_i \end{cases}. \tag{2}$$

If $\delta_i = 1$, it means the life time is less than the censoring time, and hence we are able to observe the exact life time, and we may directly use $f(x_i,\theta)$ as the likelihood for X_i. On the other hand, if $\delta_i = 0$, it means the individual is (right) censored. In this case we know $X_i \ge C_i$ hence the likelihood is proportional to $\mathbb{P}(X_i \ge C_i)$, given by $S(C_i,\theta)$. The total likelihood using equation 2 is given by

$$\mathcal{L}(\theta) \propto \prod_{i=1}^{n} [f(x_i,\theta)]^{\delta_i} \cdot [S(C_i,\theta)]^{1-\delta_i}.$$

As for truncation, what we observe becomes the conditional probability and conditional distribution. We usually work with left truncation, meaning that we are only able to observe the samples which satisfy $X_i > H$ for the pre-determined truncation time H. In this case, the distribution of X_i is the conditional distribution of X_i given $X_i > T$, i.e.,

$$f(X_i = x_i | X_i > H) = \frac{f(x_i,\theta)}{S(H,\theta)},$$

thus the likelihood is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{f(x_i,\theta)}{S(H,\theta)}.$$

Life Table and Kaplan Meier Estimator

Non-Parametric Inference

In reality we often have non-parametric inferences where we do not know the distribution of the sample in advance: All we have is just a collection of observable data. We will try to make inference and estimate key survival features based on non-parametric inferences.

Definition 4. Given a random sample $X_1, \cdots, X_n$, we define the **empirical cumulative distribution function (ECDF)** F_n by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}, \tag{3}$$

where

$$\mathbf{1}\{X_i \le x\} = \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{otherwise} \end{cases}.$$

$F_n(x)$ estimates the true distribution $F(x)$ by the ratio of the individuals in our observable sample that satisfies $X_i \le x$. Based on this, the survival function $S(x)$ can then be estimated by

$$\hat{S}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \ge x\}. \tag{4}$$

Using **Glivenko-Cantelli theorem**, the ecdf and empirical survival function converges to their population counterparts. However, this is based on the assumption that no censoring or truncation is present. In our setting, equation 3, 4 are not good estimates. In the rest of this section, we discuss two other popular non-parametric methods: Life table and Kaplan-Meier estimator.

Life Table

Life table is one of the earliest methods to estimate the survival distribution non-parametrically. It is a table which shows the survival data of a certain population at a certain age (or a certain time period). It represents the survivorship of people in a population, or the population's longevity. It usually has several components, where each column will represent a certain survival feature. See table 2 in the appendix for an example.

There are two types of life tables, namely **current life tables** and **cohort life tables**. A current life table measures the mortality of the population at a specific time period, while a cohort life table measures the mortality of the population at a specific age. That is, in a cohort life table, all individuals from the population is believed to have the same date (period) of birth, like table 2 in the appendix. While in a current life table, the ages of the individuals may vary, and it would not be a great representation of the mortality behavior of the individuals at a certain age. Hence throughout the discussion, we will focus on cohort life table and we will see how we can estimate key survival quantities from it.

As an example, table 2 in the appendix shows the cohort life table of Canadian populations between age 60 and 80 from year 2021 to 2023. We define l_x as the number of survivors at age x (in Table 2, l_{60} would be $91,805$). So the number of deaths in year x would be $d_x = l_x - l_{x+1}$. If we start our study from year 0, then the survival probability at age x can be estimated by

$$\begin{aligned}\hat{S}_x &= \frac{l_x}{l_0} \\ &= \frac{l_1}{l_0} \times \frac{l_2}{l_1} \times \cdots \times \frac{l_x}{l_{x-1}} \\ &= \prod_{i=0}^{x} \left(1 - \frac{d_i}{l_i}\right). \end{aligned} \tag{5}$$

Kalpan-Meier Estimator

We introduce an estimator Kaplan and Meier proposed, known as the **Kaplan-Meier (KM) Estimator**.

Theorem 1 (Kaplan-Meier[2], 1958)**.** Suppose we have observed the ordered event time $t_1, \cdots, t_m$, where d_j deaths are observed at t_j, n_j individuals are at risk (event still not occurred) at t_j^-, then the Kaplan-Meier (KM) estimator of the survival function $S(t)$ is given by

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right). \tag{6}$$

We propose a plug in approach to derive the KM estimator given in equation 6, where we split the time into sub-intervals: $[0, t_1), [t_1, t_2), \cdots, [t_m, +\infty)$. If $t \in [0, t_1)$, it means that no deaths had been observed in that interval, and in this case the survival probability is simply 1. Next if $t \in [t_1, t_2)$, by definition, we have

$$\begin{aligned} S(t) &= \mathbb{P}(X \ge t_1) \\ &= \mathbb{P}(\text{Survived after } t_1\ ) \\ &= \frac{\text{Number of Individuals survived after } t_1}{\text{Number of individuals at risk at } t_1^-} \\ &= \frac{n_1 - d_1}{n_1}. \end{aligned}$$

Now we consider $t \in [t_2, t_3)$, then we have the conditional probability given by

$$\begin{aligned} S(t) &= \mathbb{P}(X \ge t_2) \\ &= \mathbb{P}(\text{Survived after } t_2) \\ &= \mathbb{P}(\text{Survived after } t_2 \mid \text{Survived after } t_1) \cdot \mathbb{P}(\text{Survived after } t_1) \\ &= \frac{\text{Number of Individuals survived after } t_2}{\text{Number of individuals at risk at } t = t_2^-} \times \frac{n_1 - d_1}{n_1} \\ &= \frac{n_2 - d_2}{n_2} \times \frac{n_1 - d_1}{n_1}. \end{aligned}$$

Thus, we may use this plug in approach recursively, and we will get the KM estimator:

$$\hat{S}(t) = \prod_{t_j \le t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right),$$

which is the same as equation 6. Since our survival function is based on the terms of products, this is also known as the product limit (PL) estimator. As we can see, Kaplan-Meier estimator is very similar to equation 5 obtained by using life table, however, the key difference between them is that, in KM estimator, we have observed the exact time of occurrence t, while in the life table method, the sample is interval censored.

Kaplan-Meier estimator is **almost unbiased**. The unbiasedness holds when assuming n_j is always non-zero. But despite that, the true bias is exponentially small[3]. The variance of the Kaplan-Meier estimator can be computed by **Greenwood's formula**:

Theorem 2 (Greenwood's formula)**.** Let $\hat{S}(t)$ be the KM estimator, then

$$\widehat{\mathbf{Var}}(\hat{S}(t)) = \hat{S}^2(t) \cdot \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)},$$

where d_j is the number of death observed at t_j, n_j is the number of individuals at risk at t_j^-, as in the KM estimator.

Assuming $n_j \neq 0$, we are able to derive the $100(1-\alpha)\%$ confidence interval by **Central limit theorem**:

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\mathbf{Var}[\widehat{S}(t)]}. \tag{7}$$

A log-log transformation[4] can be used to adjust the upper and lower bound of equation 7 to within $[0, 1]$.

Variants of KM Estimator

There are several variants of the KM estimator, including **Nelson-Aalen (NA) estimator** and **Aalen-Johansen (AJ) estimator**. NA estimator estimates the cumulative hazard function $H(x)$ defined through equation 1, where we have

$$H(t) = -\log(S(t)),$$

by replacing $S(x)$ with KM estimator, we have

$$\hat{H}(t) = -\sum_{t_j \le t} \log\left(1 - \frac{d_j}{n_j}\right) = \sum_{t_j \le t} \frac{d_j}{n_j} + \mathcal{O}\left(\frac{d_j^2}{n_j^2}\right).$$

When the number of events at each time is small relative to the risk set ($d_j << n_j$), we may simply drop the higher order term, and hence the cumulative hazard function can be estimated by

$$\hat{H}(t) = \sum_{t_j \le t} \frac{d_j}{n_j}. \tag{8}$$

where equation 8 is known as the NA estimator.

AJ estimator can be viewed as a generalized version of KM estimator where the model has several intermediate states rather than simply alive and death. A common model shown in figure 1 is a healthy-diseased-death model:

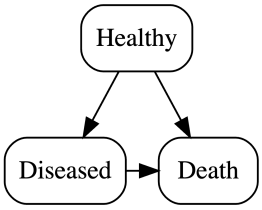

Figure 1. A multi-state model

We denote 0 as the "healthy state"; 1 as the "diseased state", and 2 as the "death state". In figure 1, there are a total of three possible transitions, and we denote $\alpha_{ij}(t)$ as the instantaneous risk of experiencing a transition from state i to j at time t. This can be viewed as a generalized hazard function between two states. We also define $\mathbb{P}_{ij}(s,t)$ as the probability that an individual in state i at time s will be in state j at time t, which can also be viewed as a generalized version of survival function between two states. For example, the probability of staying healthy from time s to t can be theoretically computed by

$$\mathbb{P}_{01}(s,t) = \exp\left\{-\int_s^t (\alpha_{01}(x)+\alpha_{02})dx\right\}.$$

In practice, we replace α_{01}, α_{02} with their corresponding NA estimator:

$$\hat{\mathbb{P}}_{01}(s,t) = \prod_{s<t_j<t}\left(1-\frac{d_{0j}}{n_{0j}}\right),$$

where d_{0j} represents the total number of transitions from state 0 at time t_j, and n_{0j} is the total number of individuals at risk at time t_j^-.

We may use similar methods to provide estimates for $\mathbb{P}_{11}(s,t), \mathbb{P}_{12}(s,t)$, etc., but it is wise to move one step further and consider a more general case where $\mathcal{I} := \{0,1,\cdots,k\}$ denotes $k+1$ different states. Define $\alpha_{gh}(t)$ as the instantaneous risk of transition from state g to state h ($g \neq h$) at time t. Define $\mathbb{P}_{gh}(s,t)$ to be the probability that an individual at state g at time s will be at state h at time t, and we define the **transition matrix** as

$$\mathbf{P}(s,t) = \begin{bmatrix} \mathbb{P}_{00} & \mathbb{P}_{10} & \cdots & \mathbb{P}_{k0} \\ \mathbb{P}_{01} & \mathbb{P}_{11} & \cdots & \mathbb{P}_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}_{0k} & \mathbb{P}_{1k} & \cdots & \mathbb{P}_{kk} \end{bmatrix}.$$

We denote d_{ghj} as the number of individuals with transition from state g to state h at observed time t_j, and $d_{gj} = \sum_{h\neq g} d_{ghj}$ as the total number of transitions out of state g at observed time t, n_{gj} as the number of individuals at state g at time t_j^-. Finally, we define the $(k+1)\times(k+1)$ matrix $\hat{\boldsymbol{\alpha}}_j$ with entries (g,h) by

$$\hat{\boldsymbol{\alpha}}_j(g,h) = \begin{cases} \dfrac{d_{ghj}}{n_{gj}} & \text{if } g \neq h \\ \\ -\dfrac{d_{gj}}{n_{gj}} & \text{if } g = h \end{cases},$$

then the AJ estimator[5] takes the form

$$\hat{\boldsymbol{P}}(s,t) = \prod_{s<t_j<t}(\boldsymbol{I}+\hat{\boldsymbol{\alpha}}_j),$$

where $\boldsymbol{I}$ is the $(k+1)\times(k+1)$ identity matrix.

Counting Processes

The properties of the estimators we introduced can be formally studied using counting processes and martingale theory. We will denote the number of death, $D(t)$, as a **counting process** such that $D(0)=0, D(t)-D(s)$ is the number of deaths in the interval $[s,t)$. If we denote $h(t)$ by the true hazard rate, $N(t)$ by the number of individuals at risk at t, then the predicated death at t is $N(t)\cdot h(t)$, and the total predicated death up to t is given by $\int_0^t N(s)\cdot h(s)ds$. It follows from **Doob-Meier decomposition** that the difference between total observed death and predicted death is a mean zero martingale:

$$M(t) = D(t) - \int_0^t N(s)\cdot h(s)ds \quad \text{where } \mathbb{E}[M(t)] = 0. \tag{9}$$

It follows from equation 9 that

$$\frac{dM(t)}{N(t)} = \frac{dD(t)}{N(t)} - h(t),$$

while integrating again, we have

$$H(t) := \int_0^t h(t)dt = \int_0^t \frac{dD(t)}{N(t)} - \int_0^t \frac{dM(t)}{N(t)}, \tag{10}$$

The Nelson-Aalen estimator can also be defined using a Rieman-Steiltjes integral: $\hat{H}(t) = \int_0^t \frac{dD(t)}{N(t)}$, which corresponds the one in equation 10, hence we see the unbiasedness of NA estimator:

$$\mathbb{E}[\hat{H}(t) - H(t)] = \mathbb{E}\left[\int_0^t \frac{dM(t)}{N(t)}\right] = 0,$$

since $M(t)$ is a mean zero martingale. The unbiasedness of KA estimator by a similar way. A rigorous approach can be found in Flemming and Harrington[6], and Li[7].

Cox's Proportional Hazard (PH) Model

Model Set Up

Cox's proportional Hazard Model[8] is a semi-parametric model which utilizes **covariates**, for example the individual's gender, age, habits, etc. We usually denote $\boldsymbol{z} \in \mathbb{R}^p$ as a p-dimensional covariate vector, and the survival probability becomes the conditional probability given the covariates $\boldsymbol{z}$. Cox's PH model suggests that the hazard function given the covariates $\boldsymbol{z}$ takes the form

$$h(x|\boldsymbol{z}) = h_0(x)\cdot\exp\left(\boldsymbol{z}^\top\boldsymbol{\beta}\right), \tag{11}$$

where $h_0(x)$ is the **baseline hazard**, and $\boldsymbol{\beta}$ is the **coefficient vector**. To estimate the coefficients vector $\boldsymbol{\beta}$, we assume for simplicity that there are no tied events (which means at each t_j there is exactly one death observed). Suppose at time t_0 we have m individuals with hazard function $h_1(t_0),\cdots,h_m(t_0)$ respectively, and we know there is exactly one death, then the probability that individual j has the event is given by

$$\mathbb{P}(\text{individual } j \text{ has the event}|t_0) = \frac{h_j(t_0)}{h_1(t_0)+\cdots+h_m(t_0)}.$$

If we replace each $h_i(t_0)$ by equation 11, we have

$$\mathbb{P}(\text{individual } j \text{ has the event}|t_0) = \frac{\exp\left(\boldsymbol{z}_j^\top\boldsymbol{\beta}\right)}{\sum_{i=1}^m \exp\left(\boldsymbol{z}_i^\top\boldsymbol{\beta}\right)},$$

where the baseline hazard is not required. We see that this probability is independent of the event time, and only depends on the order in which events occur. Based on this observation we specify a model where we assume individual i fails at time t_i exactly. Then by multiplying the probability of each individual, we get the **partial likelihood** as follows:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\boldsymbol{z}_i^\top \boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{z}_j^\top \boldsymbol{\beta}\right)}, \tag{12}$$

where $R(t_i)$ is the set of individuals at risk at time t_i. Likelihood inference can be based on equation 12, and one may estimate the coefficient vector β by maximizing the function given by equation 12.

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) = \sup_{\boldsymbol{\beta}} \prod_{i=1}^{m} \frac{\exp\left(\boldsymbol{z}_i^\top \boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{z}_j^\top \boldsymbol{\beta}\right)}. \tag{13}$$

Relative Hazard

One advantage of Cox's PH model is that the exponential part is independent of time, and thus remains a constant under the ratio. If we choose two individuals with different covariates $\boldsymbol{z}_1, \boldsymbol{z}_2$, the ratio of the two hazard rates is a constant given by

$$\frac{h(x|\boldsymbol{z}_1)}{h(x|\boldsymbol{z}_2)} = \exp((\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top \boldsymbol{\beta}).$$

Hence despite not knowing the baseline hazard $h_0(x)$, we may compare the ratio of different covariates, and see the relative hazard rate among reference group and non-reference group.

We next present a study of 863 kidney transplant patients[1] where two different covariates were introduced: The patients' gender (male and female) and race (white and black). Then we have 4 groups, namely black male; black female; white male; white female. We set $z_1 = 1$ if we have a black male and 0 otherwise; $z_2 = 1$ if we have a white male and 0 otherwise; $z_3 = 1$ if we have a black female and 0 otherwise. The white female is considered as the reference group and its covariate vector is simply $\boldsymbol{z} = \mathbf{0}$, and its hazard function in Cox model is equal to the base line hazard $h_0(t)$, and we have

$$h(t|\boldsymbol{z}) = h_0(t) \exp\Big(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3\Big).$$

By solving equation 13 numerically, we have $\beta_1 = 0.1596, \beta_2 = 0.2484, \beta_3 = 0.6567$. Now we can see the performances of non-reference group compared to the reference group:

$$\frac{h(t|z_1)}{h_0(t)} = 1.17, \frac{h(t|z_2)}{h_0(t)} = 1.28, \frac{h(t|z_3)}{h_0(t)} = 1.93.$$

Thus the relative hazard risks for black male, white male and black female compared to white female are given by $1.17, 1.28, 1.93$ respectively. One can also use this method to test if a certain treatment is effective by defining the group with placebo as the reference group, and see the relative hazard risk for treatment group. If the hazard rate is greater than 1, it means the treatment is somehow not effective, if less than 1 it means the treatment is somehow effective.

Hypothesis Testing

In this section, we will introduce the **log-rank test**, which is a non-parametric hypothesis testing methods commonly used. It can be used to compare the survival data from 2 or more groups (most commonly, treatment group & placebo group), and conclude their statistical similarity. This usually involves a null hypothesis $\mathcal{H}_0$: They are statistically equivalent at a certain significance level, and an alternative hypothesis $\mathcal{H}_1$: They are not statistically equivalent at a certain significance level.

We will consider a study done by Freireich et. al.[9], where they studied the effectiveness of 6-mercaptopurine (6-MP) among children with leukemia. In their experiment, 21 children were given 6-MP treatment (group 1), and 21 children were given a placebo (group 2). The data collected is shown in Table 3 in the appendix.

We use KM and NA estimators to estimate the survivor and hazard function for these data and generate the following plots:

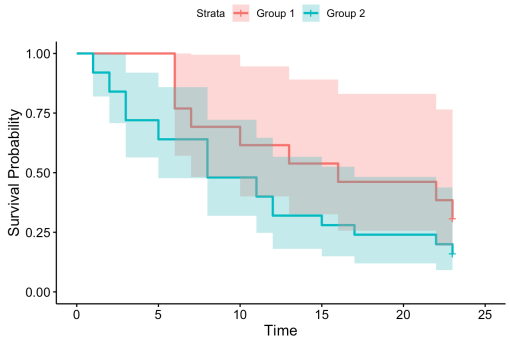

Figure 2. Kaplan–Meier survival curve with 95% confidence interval.

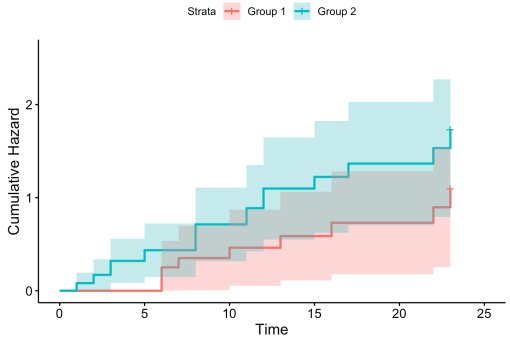

Figure 3. Nelson-Aalen hazard curve with 95% confidence interval.

From the plots we can see that group 1 (6-MP) has a lower hazard rate and higher survival probability compared to group 2 (placebo), and hence we may conclude that 6-MP is important in leukemia remission. In order to conclude our hypothesis, a rigorous testing is required. We first define the expected number of failures e_{if} for group $i = 1, 2$ at each observed failure time by

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right) \times \Big(d_{1f} + d_{2f}\Big),$$

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times \Big(d_{1f} + d_{2f}\Big).$$

That is, e_{if} is the proportion of individuals at risk in group i multiplied by the total number of failures over both groups. Let $n_f = n_{1f} + n_{2f}$ and $d_f = d_{1f} + d_{2f}$, at time $t_{(f)}$, we obtain the following table:

	Group 1	Group 2	Total
Fails	d_{1f}	d_{2f}	d_f
Survivals	$n_{1f} - d_{1f}$	$n_{2f} - d_{2f}$	$n_f - d_f$
Total	n_{1f}	n_{2f}	n_f

Table 1. contingency table for all subjects in the risk set at time $t_{(f)}$.

Under the null hypothesis $\mathcal{H}_0$, where the survival curve of group 1,2 are statistically equivalent, the random variable d_{if} forms a hyper-geometric distribution, with expected value and variance given by

$$\mathbb{E}[d_{if}] = e_{if} = \frac{n_{if}}{n_f} \cdot d_f \quad \textbf{Var}[d_{if}] = v_{if} = n_{if} \cdot \frac{d_f}{n_f} \cdot \frac{n_f - d_f}{n_f} \cdot \frac{n_f - n_{if}}{n_f - 1}.$$

We further define the total difference between observed fails and the expected fails in each group by

$$D_i - E_i = \sum_{f=1}^{n} \Big(d_{if} - e_{if} \Big), i = 1, 2,$$

then the **log-rank statistic** is defined by

$$Z^2 = \frac{(D_i - E_i)^2}{\textbf{Var}(D_i)},$$

for any group $i = 1, 2$. Under $\mathcal{H}_0$, the log-rank statistic Z^2 forms a χ-squared distribution with one degree of freedom. From table 3 we are able to compute the log-rank statistic directly: $Z^2 = 10.37$. The p-value is given by 0.0013, which provides strong evidence against the null hypothesis. Hence there is a significant difference in leukemia remission between 6-MP group and placebo group.

Acknowledgments

The original research was supervised by Professor Masoud Asgharian at the department of Mathematics and Statistics at McGill University. This article summaries my original research report *Analyzing Incident and Prevalent Cohort Survival Data* written in Summer 2025. I would like to thank Professor Asgharian for his professional guidance, without whose support this article would not have been possible.

References

1. Klein, J. P. & Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data, 2003* (Springer, 2003).
2. Kaplan, E. L. & Meier, P. *Non-parametric Estimation from Incomplete Observations* (Journal of the American Statistical Association, 1958).
3. Zhou, M. *Two-Sided Bias Bound of the Kaplan–Meier Estimator* (Massachusetts Institute of Technology, 1988).
4. Breheny, P. *Survival Data Analysis, BIOS:7210/STAT:7570/IGPI:7210 Lecture Notes.* (University of Iowa., 2019).
5. Borgan, O. *Aalen–Johansen Estimator.* (University of Oslo, 2005).
6. Fleming, T. R. & Harrington, D. P. *Counting Processes and Survival Analysis.* (Wiley Series in Probability and Statistics, 1991).
7. Li, Y. *Lecture Notes on Survival Analysis: A Counting Processes Approach.* (University of Michigan.).
8. Cox, D. R. *Regression Models and Life Tables.* (Journal of the Royal Statistical Society. Series B (Methodological), Imperial College London, 1972).
9. Kleinbaum, D. G. & Klein, M. *Survival Analysis, A Self-Learning Text.* (Springer, 2012).

Appendix

	A	B	C	D
Age	2021 to 2023	2021 to 2023	2021 to 2023	2021 to 2023
60	91,805	552	0.00601	24.95
61	91,253	598	0.00655	24.10
62	90,655	648	0.00715	23.26
63	90,007	703	0.00781	22.42
64	89,304	763	0.00854	21.59
65	88,541	828	0.00935	20.77
66	87,713	898	0.01024	19.96
67	86,815	975	0.01123	19.17
68	85,840	1,058	0.01233	18.38
69	84,782	1,149	0.01355	17.60
70	83,634	1,246	0.01490	16.84
71	82,387	1,352	0.01641	16.08
72	81,035	1,466	0.01809	15.34
73	79,570	1,588	0.01996	14.62
74	77,982	1,719	0.02204	13.90
75	76,263	1,858	0.02437	13.21
76	74,405	2,007	0.02697	12.52
77	72,398	2,163	0.02988	11.86
78	70,235	2,327	0.03313	11.21
79	67,908	2,498	0.03678	10.57
80	65,410	2,674	0.04087	9.96

Table 2. The Life Table of Canadian Population Between Age 60 to 80 from year 2021 to 2023. The elements in the first rows represent: (A) The number of survivors at age x; (B) The number of deaths between age x and $x+1$; (C) Death probability between age x and $x+1$; (D) Life expectancy (in years) at age x. *Source: Statistics Canada. Table 13-10-0114-01 Life expectancy and other elements of the complete life table, three-year estimates, Canada, all provinces except Prince Edward Island.*

$t_{(f)}$	d_{1f}	d_{2f}	n_{1f}	n_{2f}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
3	0	2	21	16
5	0	2	21	14
6	3	0	21	12
7	1	0	17	12
8	0	4	16	12
10	1	0	15	8
11	0	2	13	8
12	0	2	12	6
13	1	0	12	4
15	0	1	11	4
16	1	0	11	3
17	0	1	10	3
22	1	1	7	2
23	1	1	6	1

Table 3. The remission data among the sample of 42 children. For each ordered time $t_{(f)}$ in the sample, we use d_{if} to denote the number of children failed (whether leukemia relapse or death) at that time, separately by group i; we use n_{if} as the number of children at risk at that time.[9]