Basic Concepts in Survival Analysis

Ashley I Naimi

Spring 2021

Contents

1 Cohort and Timescale

Most of the tools we use in epidemiology are either defined, or are demonstrably valid, only based upon the presence or absence of certain fundamentals or some foundation. The idea of a cohort, and a well defined timescale are two pillars of this foundation.[1]

In epidemiologic settings, a cohort is simply a group of people. Ideally, we would like to use a particular cohort to better understand features of the population from which this cohort was sampled. Cohorts can be either closed (people do not enter or leave the cohort during the study), or open (people are free to enter or leave the study at any time). In epidemiology (and particularly in this course) we deal mostly (exclusively) with closed cohorts.

That we can interpret a parameter estimate for an exposure of interest from, say, a logistic regression model as a ratio of two odds depends on the fact that we've collected data on a *cohort* with a well-defined start and stop time. Without this underlying concept of a cohort with well-defined start and stop times, all we get from logistic models are values of a parameter which maximize the likelihood function, which is not the same as an odds or risk ratio.

To completely define a cohort, we need to clearly define a start or origin time, and a stop time. In the case of a closed cohort, without a well defined start and stop time, we would not be able to decisively state whether a given person should be in or has left the cohort. Consider the following diagram:

[1] For example, you should already know about how a case-control odds ratio can be used to estimate a *cohort* risk ratio, rate ratio, or odds ratio, depending on how the controls are sampled from the original cohort. That is, the interpretation of a case-control odds ratio depends on details emanating from the cohort we have. Specifically, a case-control odds ratio quantifies a cohort risk ratio, rate ratio, and odds ratio when we use base-case sampling, incidence density sampling, and cumulative sampling, respectively.

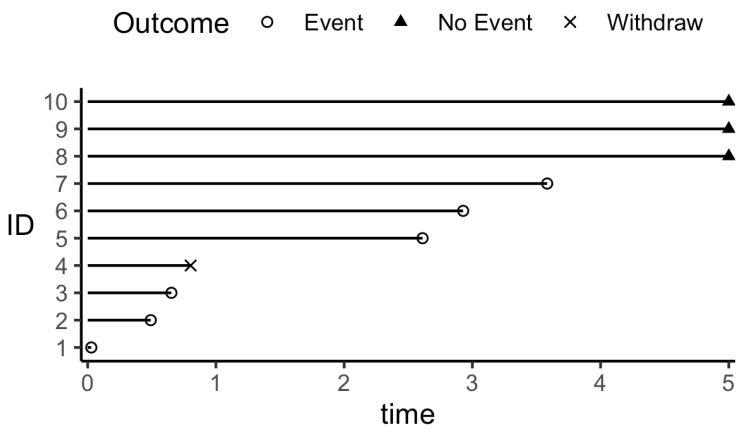

Figure 1: Observed data from a hypothetical study of 10 observations. Data are from a closed cohort with a common start time (time=0) and up to 5 time-points (e.g., weeks, months, years) of follow-up.

Figure 1 shows ten simulated observations. In this setting, time zero is our

start time. The start time should correspond to some well-defined event such as an age of interest (age as time-scale), a date of interest (calendar date as time-scale), or the timing of some important study marker (e.g., date of randomization to treatment versus placebo).

Consider the following examples from the literature with different study time-scales:

1) [!]Naimi2021 use data from a randomized trial to estimate the adherence adjusted per protocol effect of daily low-dose aspirin on pregnancy outcomes in ~1,200 women. In this study, the timescale was **weeks since randomization**, and ranged from 0 to 60 weeks.

2) [!]Getahun2005 examined stillbirth, small for gestational age, and infant mortality occurrence by the racial classification of both parents (e.g., white-white, white-black, black-white, black-black) in roughly 20 million pregnancies in the United States. In this study, the timescale was **gestational age**, starting at the 20th week of gestation.

3) [!]Huang2018 looked at the relation between different post-operative management strategies, including the use of dexamethasone versus flurbiprofen axetil on survival in 588 patients undergoing surgical lung resection for non-small-cell lung cancer. In this study, the timescale was **time since surgical resection**.

4) [!]Schwarzinger2018 looked at the relation between alcohol use and dementia risk in nearly 31 million individuals in France between 2008 and 2013. In this analysis, the timescale was age, meaning that “time 0” was the age at which the individual entered into the study, corresponding to the age at the calendar date during which the study started.

5) [!]Sabia2019 looked at the association between cardiovascular health at age 50 and the risk of subsequent dementia in ~8,000 individuals enrolled in the Whitehall II study. In this analysis, the timescale was **calendar date**, with the starting date being the date of clinical examination at age 50.

2 Censoring and Truncation

Figure 1 is an important tool, particularly for exploratory data analysis. However, for now, we will generalize this figure to depict two key concepts: **censor-**

ing and **truncation**. These concepts are illustrated in Figure 2, showing a line diagram corresponding to Figure 1, but with six distinct scenarios.

The first three observations in Figure 2 depict right, left, and interval censoring, respectively. The last three observations depict right, left, and interval truncation.[2]

[2] Interval censoring and interval truncation are often referred to as double censoring and double truncation.

Right Censoring (ID = 1 in Figure 2): occurs when an individual is enrolled in the study, but we don't know whether the individual has had an event of interest or not. This type of censoring often occurs because either an enrolled individual leaves the study (withdrawal), or the study ends (administrative censoring). This distinction is sometimes referred to as "Type I" versus "Type II" censoring. It is an important one, which will come up several times in the class. Right censoring is often said to be the most common type of censoring. Generally, when we use the word "censoring" in this class, we are referring to right censoring.

Left Censoring (ID = 2 in Figure 2): occurs when an individual is enrolled in the study, and we know has experienced an event of interest (and we know which event it is), but we have no information on *when* the event occurred. I believe this to be the most common type of censoring, due to the fact that most often, we collect data on whether an event occurred or not during the course of our study, and not on the exact timing of events. Thus, outcomes in a typical cohort study that do not have information on the timing of events are left censored.

Interval Censoring (ID = 3 in Figure 2): occurs when an individual is enrolled in the study, and we know has experienced an event of interest (and we know which event it is), but we only know that the event occurred in a bounded *interval*, with the bounds occurring after the study start date and before the study end date.

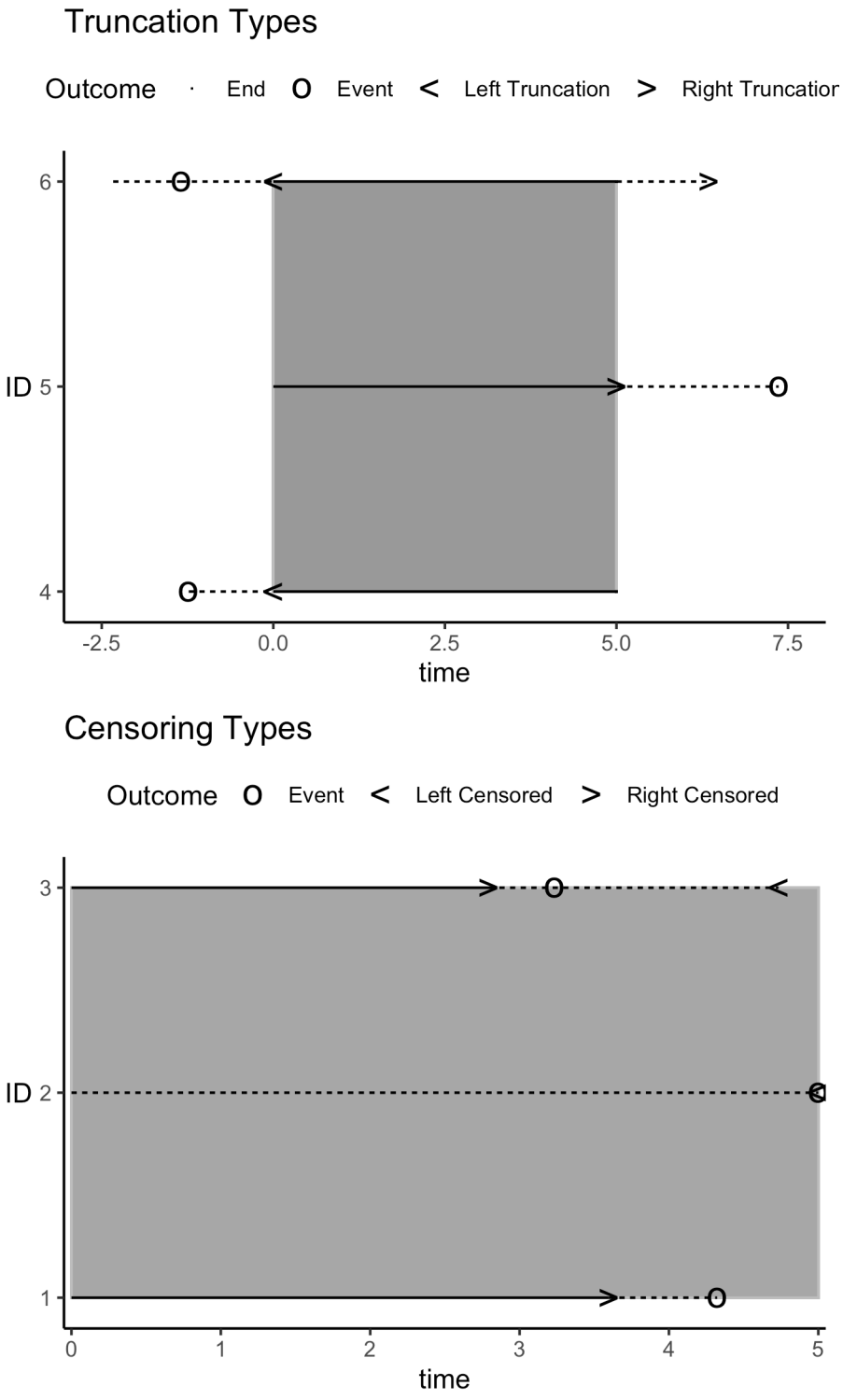

Figure 2: Six observations in a hypothetical study depicting censoring and truncation (left, right, and interval for both).

Technical Note:

In survival (a.k.a. time-to-event) analysis, survival time is typically classified as either continuous or discrete time. Simply put, in a continuous time setting, the time to the events of interest are positive real numbers ($\mathbb{R}^+$), or a quantity that can be represented as an infinite decimal expansion. In contrast, in a discrete time setting, the time to the events of interest are typically positive integer values ($\mathbb{Z}^+$), or a whole non-decimal number. In survival analysis *theory*, there are important distinctions between continuous and discrete time analyses. These distinctions are much less important for practical analyses of time-to-event data. For example, in a continuous time setting, one might have interval-censored data, since the exact timing of the event of interest might not be known. However, if the timescale of an analysis is (e.g.) week on study, and we know that the event happened in week J, this is typically enough for a discrete time analysis, and we would not have to censor the outcome.

Right Truncation (ID = 4 in Figure 2): occurs when an individual is NOT enrolled in the study because the event happened after a particular date. One example is in [!]Medley1987, who studied time from exposure to HIV contaminated blood or blood products and the development of AIDS. Data were collected retrospectively from individuals with confirmed AIDS diagnosis. The number of individuals who were exposed to HIV contaminated blood or blood products that had not yet developed AIDS was not known. In this study, only those individuals who developed AIDS by the time the study was enrolling could be identified for inclusion, which resulted in right truncated data.

Study Note:

You may have encountered various definitions of "retrospective" and "prospective" cohorts: retrospective = case-control, prospective = cohort; the investigator's perspective; and the exposure record in relation to the outcome. You may have also heard that retrospective studies are generally lower quality than prospective studies, with a range of reasons as to why. Two fundamental questions are: which of these study designs is more prone to left, right, and interval truncation?; How do the ideas of truncation and censoring relate to the quality of retrospective versus prospective studies?

Left Truncation (ID = 5 in Figure 2): occurs when an individual is NOT enrolled in the study because the event happened before a particular date. This type of truncation is common in studies of spontaneous abortion. For example, [!]Waller1998 examined the relation between prenatal exposure to trihalomethanes in drinking water (a by product of chlorination) and spontaneous

abortion. Women were recruited from prenatal care clinics. However, spontaneous abortion tends to be more common early in pregnancy (and can often be confused with normal menstruation). Thus, it is likely that many spontaneous abortions were missed because they occurred before enrollment began, resulting in left truncated data.

Interval Truncation (ID = 6 in Figure 2): occurs when an individual is NOT enrolled in the study because the event happened between two dates. Interval truncation occurs in studies of, for example, autopsy confirmed neurodegenerative diseases (ND). On the one hand, diagnosing ND is difficult, and studies tend to focus on the occurrence of disease in older populations. Thus, individuals who experience ND early tend not to be included in these studies. On the other hand, because autopsy confirmation is required for inclusion in the study, individuals who survive past the study start date are also not included. This example, as well as methods to address interval truncation, are discussed in [!]Rennert2018.

There are some important takeaways from these definitions and examples:

First, with censoring, the individuals are included in our study but we do not see when their events occur. With truncation, we do not see the individuals, and thus cannot include them in our study.[3]

[3] Linguistically, we say that *individuals* are censored, but *data* are truncated.

Second, it's important to connect the idea of censoring and truncation back to the idea of cohort and timescale, and our ability to validly interpret regression model parameters as risk differences, risk ratios, or odds ratios.[4] Clearly, censoring and truncation matter because they determine whose outcome is observed or who is in cohort. Without carefully considering how to handle censored or truncated data, we can obtain biased (i.e., inconsistent) results.

[4] Validity here depends on more than just the presence or absence of censoring and truncation. But appropriate handling of censoring and truncation are essential (i.e., necessary, but not sufficient).

3 Risk (Functions)

Let's say we did a study of the effect of some exposure on an outcome of interest, which yielded the following dataset:

These are the same data displayed in Figure 1.

Table 1: Synthetic Data

ID	exposure	confounder	start_time	stop_time	outcome
1	1	1	0	0.03	Event
2	1	1	0	0.49	Event
3	1	1	0	0.65	Event
4	1	1	0	0.80	Withdrawal
5	0	1	0	2.61	Event
6	1	1	0	2.93	Event
7	1	1	0	3.58	Event
8	0	1	0	5.00	No Event
9	1	0	0	5.00	No Event
10	0	1	0	5.00	No Event

We are going to focus here on risk. Risk is a central parameter in cohort studies [^!^Cole2015], and is often specified as the "probability of an event during a specified period of time." [^!^Rothman2008][5] For now, let's evaluate the risk without looking at the role that the exposure plays in influencing the outcome. This is akin to a "no intervention" or "no treatment" scenario, by which we mean that we want to compute the risk of the outcome that we actually observed–i.e., the risk under the natural settings in the study. Importantly, this is **not** the risk if everyone's exposure were set to zero. It's the risk that would be observed if we did nothing. This is sometimes referred to as the **natural course** risk [^!^Rudolph2021].

[5] It's useful to separate the linguistic connotations of the word "risk" from its mathematical definition, which can sometimes lead to confusion. For example, one might define the "risk of live birth". Linguistically, "risk" connotes something bad, whereas in scenarios in reproductive epidemiology successful live birth is good. Here, we will be using the word "risk" in its strictly mathematical sense. In practice, I will often use "probability" instead of risk to avoid this potential dissonance.

Technical Note:

Often when we use the word "bias" in epidemiology, we actually mean "inconsistent" in the statistical sense. Technically, an estimator $\hat{\theta}$ is consistent if, for some arbitrarily small $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0.$$

When epidemiologic bias is present (confounding, selection, information), the estimator will not converge to the truth no matter how large a sample we have. In contrast, we say that an estimator is biased (in finite samples) if:

$$E(\hat{\theta} - \theta) \neq 0.$$

That is, we can have zero confounding (i.e., a consistent estimator), but still have a biased estimator because of how poorly it performs at using the data to estimate the effect at a given sample size. One example of this is the partial likelihood estimator used to quantify parameters of a Cox regression model (see Johnson1982). Usually, this statistical bias will disappear as the sample size increases.

Mathematically, we define the risk of an outcome as

$$F(t) = P(T \leq t)$$

This equation quantifies the probability (or risk) that the observed failure time T is less than or equal to some arbitrary threshold t.

Relatedly, we could define survival as:

$$S(t) = 1 - F(t) = P(T > t)$$

The risk and survival functions are complements to one another. Both equations are a compact way of asking how the risk cumulates over time.

The risk function[6] is a fundamental function in epidemiologic analyses specificaly, and data science more generally, for several reasons:

1) It is the most complete summary available of a random variable of interest [!]Wasserman2004 (p21). Statistically speaking, there is no other function that provides more information about a random variable of interest.[7]

2) All other measures of effect or occurrence can be defined as a function of the CDF. The risk, rate, odds, and hazards, which are commonly used to analyse epidemiologic data, can all be derived from the CDF [!]Klein2003.

[6] The cumulative risk function, or cumulative distribution function, i.e., CDF

[7] In the context of this class, and most epidemiologic analyses, the random variable of interest will be a time-to-event outcome, but this need not be the case. One can define a CDF for any continuous random variable of interest.

3) It is among the most intuitive measures of occurrence available. There is a lot of literature now on how poorly humans reason with quantitive or probabalistic summaries [!Kahneman2011 (thinking fast and slow), [!Gilovich2002 (heuristics and biases), [!Taleb2001 (fooled by randomness). Measures such as the odds ratio or hazard ratio add an additional layer of complexity when reasoning with quantitative summaries [!Hernan2010 (hazards of hazard ratios) [!Greenland1987 (interpretation and choice) [!Kaufman2010 (marginalia) [!Kaufman2010 (decomposing with alot of supposing). Thus, focusing on risk has benefits in terms of keeping things simple.

For these reasons, we focus on risk extensively in this course.

We can compute the cumulative risk function $F(t)$ in several ways. Consider the synthetic data in Table 1, but imagine that instead of "outcome = Withdrawal" for ID 4, they had "outcome = Event". If this were the case, one could simply compute the risk function by calculating the average number of events in the first, second, third, fourth, and fifth years on study.[8] The denominator for this risk is everyone in the sample. For example, using the ten observations from the synthetic data in Table 1, we have:

[8] This **only** works in a simple setting where there is only a single event type, no censoring, and no left truncation. Because this is very unlikely the approach we are using here is only for demonstration.

Year 1: $4/10 = .4$

Year 2: $4/10 + 0/10 = .4$

Year 3: $4/10 + 0/10 + 2/10 = .6$

Year 4: $4/10 + 0/10 + 2/10 + 1/10 = .7$

Year 5: $4/10 + 0/10 + 2/10 + 1/10 + 0/10 = .7$

This simple approach is sometimes referred to as the empirical distribution function (ECDF) estimator, but (again) doesn't usually work in survival data (becuase of censoring and truncation).

If we plot these risks using a step-function with Year as the x-axis and risk as the y-axis, we might get the following:

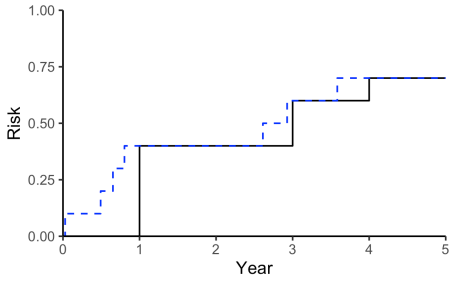

Figure 3: Basic cumulative distribution function (cumulative risk) for the synthetic data presented in Table 1. The risks in this Figure were obtained by computing basic risk quantities in each year as 'Number Events / Number At Risk', and is used only for illustrative purposes. In more realistic settings, alternative approches (which will be presented later) should be used. Blue dashed line is CDF estimated via Kaplan-Meier, discussed next.

The approach we just used to compute the cumulative distribution function above was used to simply illustrate the core idea behind the risk function $F(t)$. This is not the approach one would use in typical settings, because we often have to deal with issues such as right censoring and left truncation.

The next section will be about **how** to estimate the cumulative distribution function for a time-to-event outcome. We will be introduced to two different approaches, first the Kaplan-Meier estimator, and then later (when we cover competing risks) the cumulative incidence function estimator (from !^!Gray1988). We'll also discuss the factors that should lead you to decide choosing one or the other, and go over how to use them in the R programming language.

It's important to note here that the KM and CIF estimators will estimate the same thing and yield the same results when there are no competing risks present. We will cover what competing risks are, and what happens to these estimators when competing risks are present in later sections.

4 Kaplan-Meier Estimator

The first estimator is the Kaplan-Meier (KM) approach. This approach should be used in a setting where you have a single time-to-event outcome of interest (e.g., all cause mortality). It can also handle right censoring, and with a slight modification is able to handle left-truncation.

The KM estimator for the survival curve is the product, taken over the ordered set of distinct event times, of the complement of the number of events divided by the number at risk:

$$\hat{S}(t) = \prod_{k \in t_k \leq t} (1 - d_k/n_k)$$

where d_k is the number of events, and n_k is the number at risk, both at time $k = t_k$ [^!^Cole2020]. Here, $n_k = \sum_{i=1}^{n} I(t_k \leq T_i)$, which is the number of individuals in the risk-set at time t_k. Taking the complement of this estimator gives us a KM estimator for the cumulative distribution function.

To implement the KM estimator in R, we need to use the `survival` package, which includes the `Surv()` and the `survfit()` functions. We will use the data in Table 1, and we have to set up the data in so the `Surv()` and `survfit()` functions work as we want them to.

The key issue we need to address is the following: in Table 1, we use the number "2" to denote Type I right censoring, and the number "0" to denote Type II right censoring. However, the functions in `R` do not distinguish between Type I and Type II censoring. We need set all these observation's (ID = 4, 8, 9, 10) outcome to the same value. We'll pick the number "0":

```
install.packages("survival", repos = "http://cran.us.r-project.org")
```

```
## Installing package into '/usr/local/lib/R/4.1/site-library'
## (as 'lib' is unspecified)
```

```
library(survival)

# modify the data: 'surv_dat' was used to create table 1
surv_dat <- surv_dat %>%
    mutate(outcome = if_else(outcome %in% c(0, 2), 0, outcome))

# examine
surv_dat %>%
    select(ID, start_time, stop_time, outcome) %>%
    arrange(ID)
```

```
## # A tibble: 10 x 4
##       ID start_time stop_time outcome
##    <int>      <dbl>     <dbl>   <dbl>
```

```
##  1     1          0    0.0282       1
##  2     2          0    0.492        1
##  3     3          0    0.653        1
##  4     4          0    0.803        0
##  5     5          0    2.61         1
##  6     6          0    2.93         1
##  7     7          0    3.58         1
##  8     8          0    5            0
##  9     9          0    5            0
## 10    10          0    5            0
```

```
# fit KM curve
example_surv <- survfit(Surv(time = start_time, time2 = stop_time,
    event = outcome) ~ 1, data = surv_dat)

# create dataset for plotting
plot_dat <- tibble(Year = c(0, example_surv$time), Risk = c(0,
    1 - example_surv$surv))

# examine dataset
plot_dat
```

```
## # A tibble: 9 x 2
##     Year  Risk
##    <dbl> <dbl>
## 1 0      0
## 2 0.0282 0.1
## 3 0.492  0.2
## 4 0.653  0.3
## 5 0.803  0.3
## 6 2.61   0.417
## 7 2.93   0.533
## 8 3.58   0.65
## 9 5      0.65
```

```
# plot KM curve
km_plot <- ggplot() + geom_step(data = plot_dat, aes(x = Year,
    y = Risk), direction = "hv") + scale_x_continuous(expand = c
    0)) + scale_y_continuous(expand = c(0, 0), limits = c(0,
    1))
```

The CDF curve generated by the above code is presented in Figure 4 in the margin.

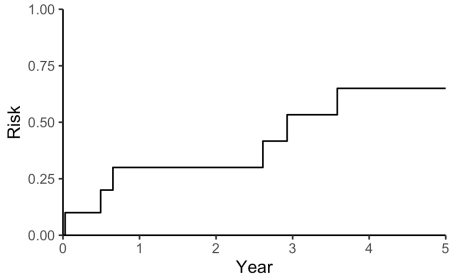

Figure 4: Cumulative distribution function

Technical Note:

Consider the Table (the `tibble`) in the R output above that includes the Year and the Risk plotted in the margin figure. Notice that the last number is 0.65, effectively stating that the overall risk in the sample of 10 observations is 0.65. But, out of the 10 individuals, only six of them had the event. This suggests that the overall risk should be 0.60, and not 0.65. Why the discrepancy? Is the KM estimator wrong?

The explanation for this higher than expected risk is the censored observation (ID=4), and the fact that, built into the KM estimator is the "redistribution to the right" algorithm. This algorithm spreads the risk that would have resulted from any censored observations had they not been censored, and redistributes it proportionally to the events that occur after the censoring for this observation takes place. In effect, this algorithm redistributes the risk from censored observations to remaining ovservations. As a result, the end of study risk estimated with a KM estimator is usually higher in the presence of censoring than the empirical risk function. This re-distribution is a "hidden imputation" that is not often recognized with the KM estimator [1]Cole2020.

5 Takeaways