Basic Concepts in Survival Analysis

Ashley I Naimi

Spring 2022

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:

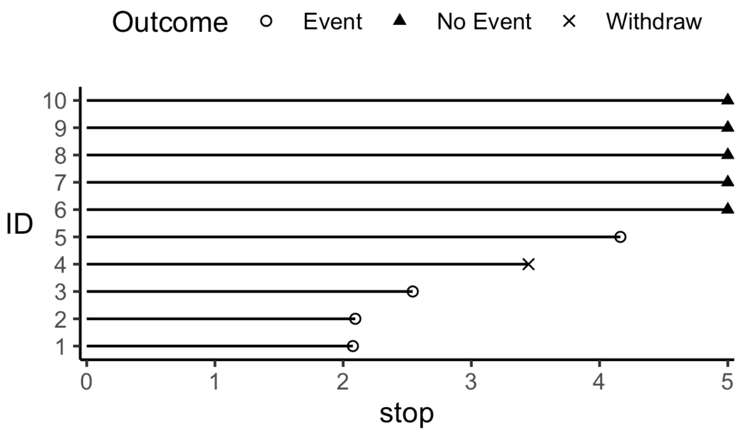

Figure 1: Observed data from a hypothetical study of 10 observations. Data are from a closed cohort with a common start time (time=0) and up to 5 time-points (e.g., weeks, months, years) of follow-up.

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

[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.

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) Naimi et al. (2021) 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) Getahun et al. (2005) 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) Huang et al. (2018) 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) Schwarzinger et al. (2018) 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) Sabia et al. (2019) 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]

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.

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

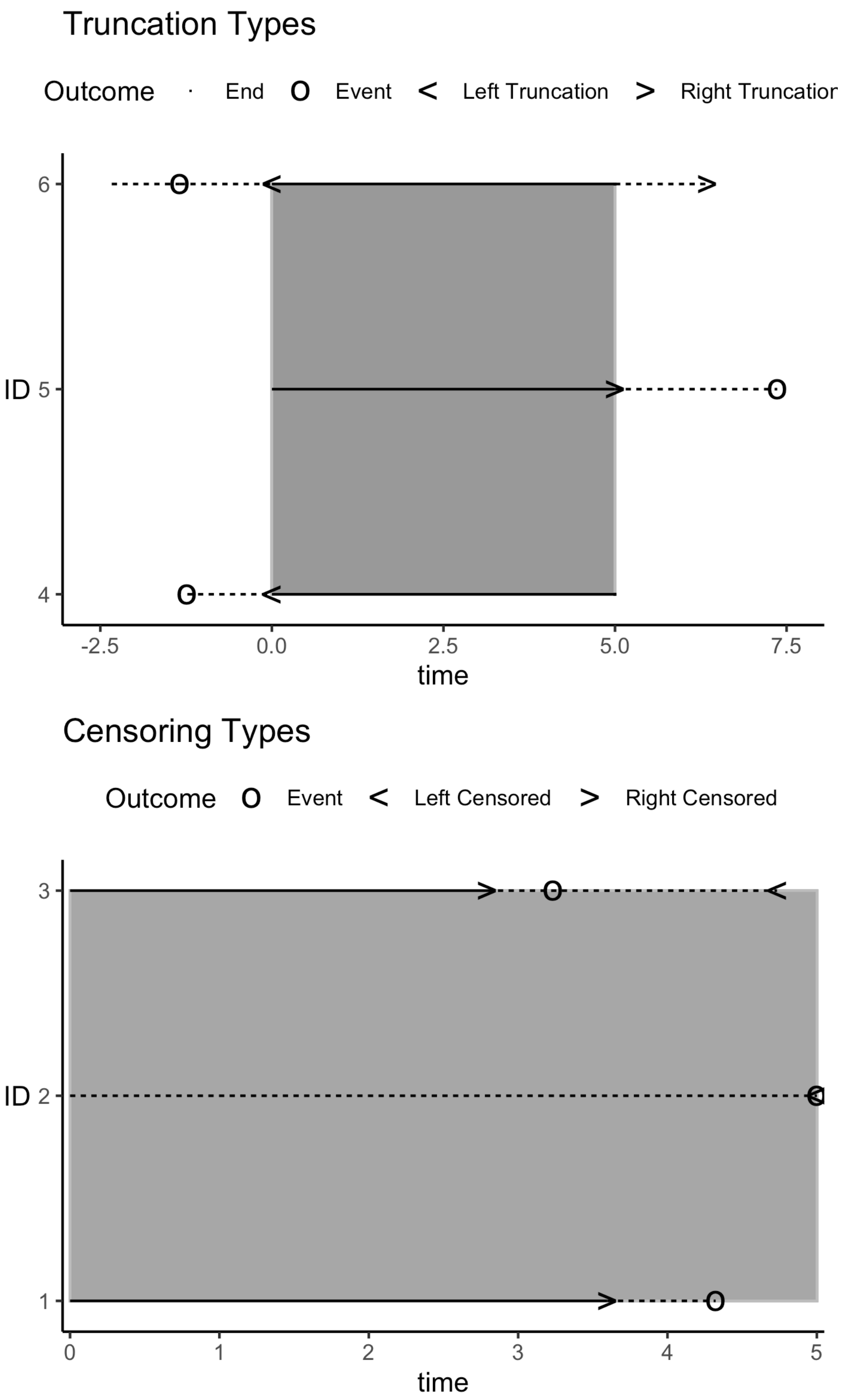

Figure 2: Six observations in a hypothetical study depicting censoring and truncation (left, right, and interval for both).

> **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 Medley et al. (1987), 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, Waller et al. (1998) 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 Rennert (2018).

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]

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.

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:

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

[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).

Table 1: Synthetic Data

ID	exposure	confounder	start_time	stop_time	outcome
1	0	1	0	2.08	Event
2	0	0	0	2.09	Event
3	0	1	0	2.54	Event
4	0	0	0	3.45	Withdrawal
5	1	0	0	4.16	Event
6	0	0	0	5.00	No Event
7	1	0	0	5.00	No Event
8	1	1	0	5.00	No Event
9	0	0	0	5.00	No Event
10	0	1	0	5.00	No Event

These are the same data displayed in Figure 1.

We are going to focus here on risk. Risk is a central parameter in cohort studies (Cole et al., 2015), and is often specified as the "probability of an event during a specified period of time." (Rothman et al., 2008)[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 (Rudolph et al., 2021).

[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 $\epsilon > 0$:
>
> $$\lim_{n\rightarrow\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0.$$
>
> 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 over follow-up 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, where the threshold t is defined over the domain of follow-up. $F(t)$ quantifies the probability of the event occurring at or before time 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 risk (or survival) cumulates over time. $S(t)$ quantifies the probability of no events occurring until after time t.

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 Wasserman (2004) (p21). Statistically speaking, there is no other function that provides more information about a random variable of interest.[7]

[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.

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 Klein and Moeschberger (2005).

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 Kahneman (2011), Gilovich et al. (2002), Taleb (2007). Measures such as the odds ratio or hazard ratio add an additional layer of complexity (Hernán, 2010, Greenland (1987), Kaufman (2010), Kaufman (2010)). 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:

Year 1: $0/10 = 0$

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

Year 3: $0/10 + 3/10 + 1/10 = .4$

Year 4: $0/10 + 3/10 + 1/10 + 1/10 = .5$

Year 5: $0/10 + 3/10 + 1/10 + 1/10 + 0/10 = .5$

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

[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.

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:

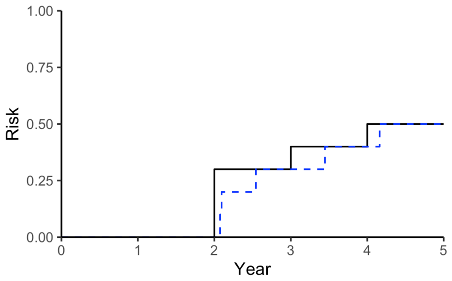

Figure 3: Basic cumulative distribution function (cumulative risk) for the synthetic data presented in Table 1. The risks in this Figure were obtained by computing basic risk quantities in each year as 'Number Events / Number At Risk', and is used only for illustrative purposes. In more realistic settings, 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 Gray (1988)). 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 subsequently.

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 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 $k = t_k$ (Cole et al., 2020). 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 "Withdrawal" to denote Type I right censoring, and the "No Event" 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")
```

```
##
## The downloaded binary packages are in
##  /var/folders/z_/cty0tpg97wz_x1d1zgdhwllr0000gs/T//RtmpJKXIey/downloaded_packages
```

```
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      2.08       1
##   2     2          0      2.09       1
##   3     3          0      2.54       1
##   4     4          0      3.45       0
##   5     5          0      4.16       1
##   6     6          0      5          0
##   7     7          0      5          0
##   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: 7 x 2
##     Year  Risk
##    <dbl> <dbl>
## 1  0     0
## 2  2.08 0.1
## 3  2.09 0.2
## 4  2.54 0.3
## 5  3.45 0.3
## 6  4.16 0.417
## 7  5    0.417
```

```
# 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))
```

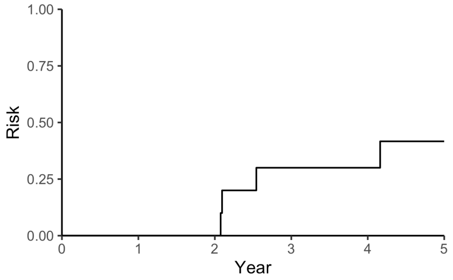

Figure 4: Cumulative distribution function obtained from the Kaplan-Meier estimator from the example data in Table 1.

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

It's important to clarify here that left truncation and right censoring are so common in time-to-event analyses, that the `survfit` function in `R` handles them by default. The way these are handled is through the coding/selection of the start time (for left truncation) and the outcome (for censoring).

When left truncation occurs, the start time for follow-up in the sample will differ across participants. Some may align with "time zero" (which could correspond to a certain age, calendar date, or other specific time), but participants will enter into the study after "time zero". This is a classic sign of left truncation, and to address it using `survfit` and `Surv`, one has to ensure that the `time` argument in the `Surv` function reflects the study entry time.

Similarly, for censored observations, one has to select a common value for the outcome for all participants who were censored.

```
surv_model <- survfit(Surv(time = entry_time, time2 = exit_time,
    event = outcome))
```

> **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 0.42, effectively stating that the overall risk in the sample of 10 observations is 0.42. But, out of the 10 individuals, only 4 of them had the event (out of 10). This suggests that the overall risk should be 0.4, and not 0.42. 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 observations. 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.
>
> A similar phenomenon occurs for left truncation. In effect, the "extended" KM estimator imputes the risks for the "ghosts" that were truncated, on the basis of the observed but delayed entries into the cohort.
>
> Both the re-distribution of censored risks and imputing of ghost risks is a "hidden imputation" that is not often recognized with the KM estimator Cole et al. (2020). Redistribution-to-the-right has particularly important implications when competing risks are present.

5 Competing Events

Thus far we've discussed estimating risk under the unique and relatively uncommon scenario where there are only two possible events that can occur in the study: 1) the single event of interest, and 2) right censoring (either due to loss-to-follow-up or administrative censoring).

This scenario is uncommon because in a typical study, there are several events that occur. For example, when studying cause specific death (e.g., death due to myocardial infarction), it is likely that death from other causes occurs in the sample. When studying preterm birth, it is likely that fetal loss (either due to early pregnancy loss, or stillbirth) occurs in the sample. When studying time to relapse among patients undergoing bone marrow transplants for leukemia, death[9] from any causes is possible.

When multiple events can occur in a study, the possibility of competing events (or competing risks) arises. **A competing event (or risk) is an event whose occurrence precludes the event of interest from occuring.**

When competing events are possible, one needs to more carefully evaluate

[9] There is a common theme: death is often an important competing event.

exactly what the risk function is quantifying. The fundamental question one must ask is: “what should I do with the events that aren’t of primary interest?” One approach, which is the simplest, is to create a **composite endpoint**. Composite endpoints can be created by coding *any* event as an “Event” with a KM or other estimator, even though the research question may be focused on a single component of the composite endpoint.

Unfortunately, while easy, creating composite endpoints to deal with competing events is often unsatisfying and potentially misleading. For example, if the exposure of interest has a large effect on a secondary endpoint, but a very small effect on the event of interest, the exposure effect on the event of interest would be overwhelmed by the effect that is not of primary interest.

To more appropriately handle competing events, we need to introduce two different *versions of risk*. The first is **cause-specific** risk. This is the risk we’d obtain if we prevented competing risks from occurring.[10] We can estimate this risk very simply by censoring the competing risks.

To explore how we obtain estimates of the cause-specific risk function, let’s use the cohort dataset (`2021_12_30-section1_cohort.csv`). This dataset has 100 observations and six variables.

```
dim(cohort)
```

```
## [1] 100   6
```

Here are the first six observations from the dataset:

```
head(cohort)
```

```
## # A tibble: 6 x 6
##      ID  stop exposure confounder outcome start
##   <dbl> <dbl>    <dbl>      <dbl>   <dbl> <dbl>
## 1     1  5           0          1       0     0
## 2     2  5           1          0       0     0
## 3     3  2.09        0          0       1     0
## 4     4  5           1          1       0     0
## 5     5  2.08        0          1       1     0
## 6     6  4.16        1          0       1     0
```

[10] Be aware that while the language here sounds very much like we’re using the potential outcomes framework, we’re not. You may have learned of this risk as the “conditional risk” in EPI 545.

Notice that the outcome has three different levels:

```
cohort %>%
    count(outcome)
```

```
## # A tibble: 3 x 2
##   outcome     n
##     <dbl> <int>
## 1       0    46
## 2       1    28
## 3       2    26
```

5.1 Cause Specific Risk

Suppose we were primarily interested in the event labeled "outcome = 2", where "outcome = 1" denotes a competing risk, and "outcome = 0" denotes right censoring. To estimate the cause specific risk function for "outcome = 2", we can again use the KM estimator. What needs to be different this time around is that we have to treat as censored those with both "outcome = 0" as well as "outcome = 1". We can do this by creating an indicator[11] variable for the outcome:

```
# note this is a numeric variable!
cohort <- cohort %>%
    mutate(cs_outcome = as.numeric(outcome == 1))
```

This new variable converts all events (i.e., censoring and competing events) to censored observations:

```
cohort %>%
    count(cs_outcome)
```

```
## # A tibble: 2 x 2
##   cs_outcome     n
##        <dbl> <int>
## 1          0    72
## 2          1    28
```

[11] Note that generally, the *indicator function* refers to a function that takes the value of 1 if the argument of the function is true, and zero otherwise. The function is usually denoted with an "$I()$" or a "$\mathbf{1}()$". In our case, if our outcome was denoted Y, the indicator variable we need to estimate the cause specific risk is $I(Y = 2)$ or $\mathbf{1}(Y = 2)$. In R, the `as.numeric()` function can serve as an indicator function.

And we can use this with the KM estimator to obtain an estimate of the cause-specific risk function:

```
surv_model <- survfit(Surv(time = stop, event = cs_outcome) ~
    1, data = cohort)

plot_dat <- tibble(time = surv_model$time, risk = 1 - surv_model$surv)

plot1 <- ggplot(plot_dat) + scale_y_continuous(expand = c(0,
    0), limits = c(0, 1)) + scale_x_continuous(expand = c(0,
    0)) + ylab("Cumulative Risk") + xlab("Time on Study") + geom_step(aes(x = time,
    y = risk))

ggsave(filename = "../figures/2021_12_29-cum_risk_section1.pdf",
    plot = plot1)
```

```
## Saving 4 x 2.5 in image
```

```
plot_dat_km <- plot_dat
```

The above code generates the `surv_model` object, creates a plot using this object with `ggplot`, and saves it to the `figures` folder using `ggsave`. We can include the saved figure in our RMarkdown document using the following code:

~~~
```{r csfigure, out.width="10cm", fig.align='center', fig.cap="Cumulative cause-specific ...", echo=T}
knitr::include_graphics("../figures/2021_12_29-cum_risk_section1.pdf")
```
~~~

which provides the following Figure:

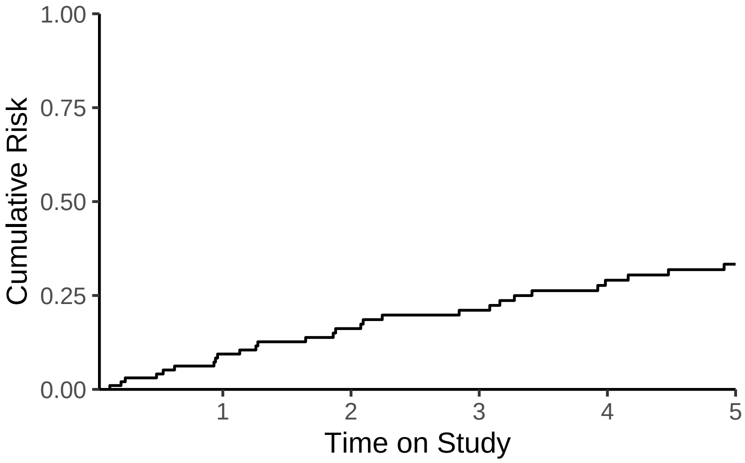

Figure 5: Cumulative cause-specific risk of the outcome in the example dataset for section 1 (outcome = 1).

One could also use the generic `plot()` function to obtain the KM curve. For example: `plot(surv_model)` (try it!). However, it is much easier to modify the figure and adapt it (e.g., to meet submission requirements) using ggplot.

It's important to understand precisely and exactly how to interpret the risk curve in Figure 5. First, we combined into a single category all individuals who were censored and who had the competing risk event (outcome = 2), and we treated this entire category as censored. Second, and perhaps more importantly, the Kaplan-Meier estimator imputes the risk of the outcome of interest for censored observations (see Technical Note above). As a result, one has to interpret the cumulative risk curve in Figure 5 as the risk that would be observed in a situation where we were able to prevent `outcome = 2` from occurring, thus keeping these individuals at risk for `outcome = 1`.[12]

This subtle point (the imputation of `outcome = 1` risk for competing events) is a controversial topic, but is unfortunately not always recognized as an issue. To place this issue into some context, imagine that in our dataset of 100 observations, `outcome = 1` denotes individuals who experience a myocardial infarction, and `outcome = 2` denotes death from any cause. In this case, estimating the cause specific risk of myocardial infarction translates directly to estimating the risk of myocardial infarction in a world where we'd prevent death from any cause from occurring.[13] For this reason, estimating cause-specific risks in the presence of competing events is my least favored approach.

[12] Note that the same is true for other methods that adjust or account for censoring, such as inverse probability of censoring weighting or imputation (e.g. , Cain and Cole, 2009, or van Buuren (2012)).

[13] Again, even though these sentences are fundamentally counterfactual in nature, these risks are not typically formulated using potential outcomes.

5.2 Sub-Distribution Risk

One alternative to cause-specific risks is **sub-distribution** risks.[14] These risks are useful even if we are interested in a primary outcome, and we have several other "nuisance" outcomes to deal with.

Suppose again our interest lies in the event labeled "outcome = 1", where "outcome = 2" is a competing risk, and "outcome = 0" is right censoring. To estimate the sub-distribution risk function for "outcome = 1", we can no longer use the KM estimator. There are several options available to us to estimate sub-distribution risks, including the Aalen-Johansen estimator or Gray's cumulative incidence function (CIF) estimator.

We'll first fit the Aalen-Johansen estimator to quantify the sub-distribution risk function for both "outcome = 1" and "outcome = 2". The code to fit the Aalen-Johansen estimator in R is nearly identical to the code we used to fit the KM estimator. The key difference lies in how the outcome is coded. To fit the Aalen-Johansen estimator, the outcome variable must be of the proper **type** and must be in a specific **order**. For example:

```r
cohort <- read_csv("../data/2021_12_30-section1_cohort.csv")

cohort
```

```
## # A tibble: 100 x 6
##       ID  stop exposure confounder outcome start
##    <dbl> <dbl>    <dbl>      <dbl>   <dbl> <dbl>
##  1     1 5            0          1       0     0
##  2     2 5            1          0       0     0
##  3     3 2.09         0          0       1     0
##  4     4 5            1          1       0     0
##  5     5 2.08         0          1       1     0
##  6     6 4.16         1          0       1     0
##  7     7 2.54         0          1       2     0
##  8     8 3.45         0          0       2     0
##  9     9 5            0          0       0     0
## 10    10 5            0          0       0     0
## # ... with 90 more rows
```

[14] You may have learned of this as an "unconditional" risk in EPI545.

```
table(cohort$outcome)
```

```
##
##  0  1  2
## 46 28 26
```

```
cohort <- cohort %>%
    mutate(outcome = factor(outcome, 0:2, labels = c("censor",
        "event", "competing risk")))  # note converted to factor

table(cohort$outcome)
```

```
##
##         censor          event  competing risk
##             46             28              26
```

In the `2021_12_30-section1_cohort.csv` data, the outcome is coded as a numeric variable. In the above code, we convert this numeric variable to a factor variable. The conversion and ordering of this factor variable is accomplished using the `factor()` function in `R`. In the arguments of this function, we tell it to convert the "outcome" variable (argument 1), we tell it that the order of the levels are `0, 1, and 2` (argument 2, coded as `0:2` in R), and we tell it that the labels for this order are "censor", "event", and "competing risk". Importantly, the censoring level has to come first, but the "event" and "competing risk" levels can come in any order.[15] Once the outcome is coded in this way, we can use the same code we used to fit the KM estimator to obtain the AJ estimator of the sub-distribution function:

```
aj_fit <- survfit(Surv(time = stop, event = outcome) ~ 1, data = cohort)

plot_dat0 <- tibble(time = aj_fit$time, risk = aj_fit$pstate[,
    2])

plot_dat1 <- tibble(time = aj_fit$time, risk = aj_fit$pstate[,
    3])
```

[15] This becomes relevant when there are several competing risks and/or events of interest.

```
p1 <- ggplot() + scale_y_continuous(expand = c(0, 0), limits = c(0,
    1)) + scale_x_continuous(expand = c(0, 0)) + ylab("Cumulative Risk") +
    xlab("Time on Study") + geom_step(data = plot_dat0, aes(x = time,
    y = risk)) + geom_step(data = plot_dat1, aes(x = time, y = risk),
    color = "#D55E00") + geom_step(data = plot_dat_km, aes(x = time,
    y = risk), linetype = 2)

ggsave(filename = "../figures/2021_12_29-subdist_risk_aj_section1.pdf",
    plot = p1)
```

Again, we can include the saved figure in our RMarkdown document using the following code:

~~~
```{r sdajfigure, out.width="10cm", fig.align='center', fig.cap="Cumulative sub-dist ...", echo=T}
knitr::include_graphics("../figures/2021_12_29-subdist_risk_aj_section1.pdf")
```
~~~

Which provides the following Figure:

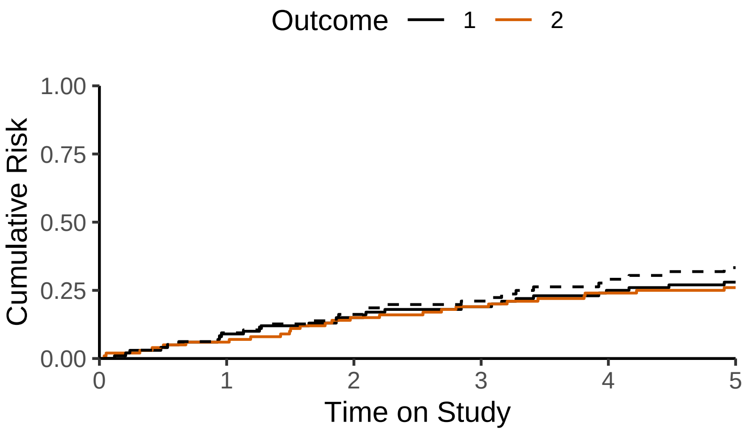

Figure 6: Cumulative sub-distribution risks of the outcomes obtained from the Aalen-Johansen estimator in the example dataset for section 1. Black line represents outcome = 1. Orange line represents outcome = 2. Dashed line represents cumulative cause-specific risks obtained with the Kaplan-Meier Estimator.

Gray's CIF estimator is available in the `cmprsk` package in R. To fit this estimator to our data, we can use the following:

```
library(cmprsk)
```

```
cohort <- read_csv("../data/2021_12_30-section1_cohort.csv")

gray_cif <- cuminc(cohort$stop, cohort$outcome, cencode = 0)  # note outcome is back to numeric!

str(gray_cif)
```

```
## List of 2
##  $ 1 1:List of 3
##   ..$ time: num [1:58] 0 0.118 0.118 0.206 0.206 ...
##   ..$ est : num [1:58] 0 0 0.01 0.01 0.02 0.02 0.03 0.03 0.04 0.04 ...
##   ..$ var : num [1:58] 0 0 0.0001 0.0001 0.000198 ...
##  $ 1 2:List of 3
##   ..$ time: num [1:54] 0 0.0375 0.0375 0.0526 0.0526 ...
##   ..$ est : num [1:54] 0 0 0.01 0.01 0.02 0.02 0.03 0.03 0.04 0.04 ...
##   ..$ var : num [1:54] 0 0 0.0001 0.0001 0.000198 ...
##  - attr(*, "class")= chr "cuminc"
```

As we can see from the `structure` command above, the `cuminc()` function provides two sets of information, each with a `time`, `est`, and `var` element. In this case, the `est` element is an estimate of the sub-distribution risk of the outcome. In the list indexed by `1 1`, `est` represents the risk of "outcome = 1". In the list indexed by `1 2`, `est` represents the risk of "outcome = 2". We can use these risks and the `time` element for each to plot the relevant cumulative incidence functions of interest:

```
plot_dat1 <- tibble(time = gray_cif$`1 1`$time, risk = gray_cif$`1 1`$est,
    Outcome = 1)  # capital for legend purposes

plot_dat2 <- tibble(time = gray_cif$`1 2`$time, risk = gray_cif$`1 2`$est,
    Outcome = 2)  # capital for legend purposes

plot_dat <- rbind(plot_dat1, plot_dat2)

plot_dat <- plot_dat %>%
    mutate(Outcome = factor(Outcome))  # convert to factor here for legend purposes
```

```
plot1 <- ggplot() + scale_y_continuous(expand = c(0, 0), limits = c(0,
    1)) + scale_x_continuous(expand = c(0, 0)) + ylab("Cumulative Risk") +
    xlab("Time on Study") + scale_color_manual(values = c("#000000",
    "#D55E00")) + geom_step(data = plot_dat, aes(x = time, y = risk,
    color = Outcome)) + geom_step(data = plot_dat_km, aes(x = time,
    y = risk), linetype = 2)

ggsave(filename = "../figures/2021_12_29-subdist_risk_cif_section1.pdf",
    plot = plot1)
```

Again, we can include the saved figure in our RMarkdown document using the following code:

~~~
```{r sdciffigure, out.width="10cm", fig.align='center', fig.cap="Cumulative sub-dist ...", echo=T}
knitr::include_graphics("../figures/2021_12_29-subdist_risk_section1.pdf")
```
~~~

Which provides the following Figure:

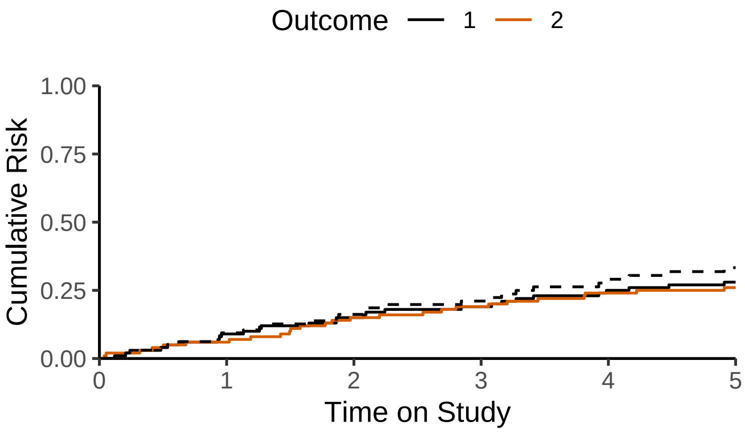

Figure 7: Cumulative sub-distribution risks of the outcomes obtained from Gray's CIF estimator in the example dataset for section 1. Black line represents outcome = 1. Orange line represents outcome = 2. Dashed line represents cumulative cause-specific risks obtained with the Kaplan-Meier Estimator.

The key now is to clarify the interpretation of the sub-distribution risk in Figure 6 or Figure 7, and the cause specific risk which is represented in both Figures with a dashed line. As we've already stated, the dashed line represents the cumulative risk of outcome = 1 we'd observe if we could prevent outcome = 2 from occurring.

On the other hand, the sub-distribution quantifies the risk for outcome 1 if outcome 2 were allowed to occur naturally. The same interpretation is obtained if outcome 2 is the primary outcome of interest. Why is this called a sub-distribution?

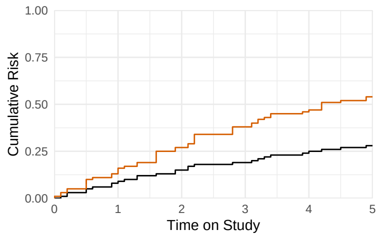

Figure 8: Stacked sub-distribution function curves. Black line represents the cumulative sub-distribution risk for outcome 1. Orange line represents cumulative sub-distribution risk for the combined outcome 1 and outcome 2 (composite outcome).

Figure 8 shows the same sub-distribution function displayed in Figure **??** except this time the risks are stacked. That is, the orange line represents the risk of the composite outcome (i.e., 1 and 2), while the black line represents the risk of outcome 1 alone. Since the risk displayed in the orange line is composed of two risks, one can think of the black line as a subset of the distribution of the composite outcome risk. Hence, sub-distribution.

Both cause-specific and sub-distribution risks have been around for a long time. They are sometimes referred to as classical statistical quantities because they were not formulated within the potential outcomes (or some other formal) causal framework. This has had one very important implication that we should clarify here. There has long been an argument in the competing risks literature that cause-specific risks are better aligned with understanding etiology, while sub-distribution risks are better aligned with generating predictive models. For example, in a great paper, Bryan Lau and colleagues have stated that "The [cause-specific relative hazard] might be more applicable for studying the etiology of diseases, whereas the [sub-distribution relative hazard] might be more appropriate for predicting an individual's risk for an outcome or resource allocation."

I believe this interpretation to be incorrect. The main reason is that in order to interpret the cause specific risk in the presence of competing risks as causal, there needs to be a clear way to prevent the competing risk from occurring. But when the competing risk is something like mortality (either all cause, or other cause), it is often not possible to prevent the competing risk from occurring. In an actual intervention aimed at improving a health outcome, we would not typically create secondary interventions to prevent competing risks from occurring, and in most settings it would be impossible to do so (can you prevent death?). For this reason, my view is that sub-distribution risks are actually easier to interpret causally than cause-specific risks (Rudolph et al., 2020).

A number of other types of quantities exist when interest lies in analyzing competing risks data (Young et al., 2020). Here, we only discussed two. Making an informed decision about whether to use cause-specific or sub-distribution estimators is the bare minimum set one should consider when handling competing risks (Cole et al., 2020).

6 Takeaways

- Well defined cohorts are fundamental to clear and accurate interpretation of any result, particularly risk. 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.
- Common threats to validity include left truncation and right censoring.
- Left truncation occurs when an individual is NOT enrolled in the study because the event happened before a particular date.
- Right censoring 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.
- With truncation, key individuals are NOT included in the cohort. With censoring, individuals are in the cohort, but information is missing on them (e.g., outcome status).
- Risk is a central measure in the empirical sciences, and is defined as "the probability of an event during a specified period of time."
- Risk, cumulated over time T, is defined 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, where the threshold t is defined over the domain of follow-up.

- Survival, cumulated over time T, is defined as $S(t) = 1 - F(t) = P(T > t)$. This equation quantifies the probability that the observed failure time T is greater than some arbitrary threshold t, where the threshold t is defined over the domain of follow-up (i.e., the probability of surviving past some time).

- The Kaplan-Meier estimator is a useful tool for quantifying cumulative risk or survival when there is a single outcome, with observations potentially subject to right censoring or left truncation. When there is more than one outcome of interest (e.g., competing events) other estimators should be considered.

- A competing risk is an event or outcome that is not of primary interest, and that prevents the event or outcome of primary interest from occurring.

- Cause specific risks can be estimated in the presence of competing events using a Kaplan-Meier estimator by censoring the competing risks. They are interpreted as the risk of the outcome of interest if we completely prevented the competing event(s) from occurring.

- Subdistribution risks can be estimated with the Aalen-Johansen estimator or Gray's cumulative incidence function estimator. They provide a measure of risk for the event of interest if the competing risk(s) were allowed to occur naturally.

- Statisticians and scientists have long espoused the view that cause-specific risks are better for etiologic questions, while subdistribution risks are better for predictive questions, or for the allocation of resources. However, because cause-specific risks require we prevent competing events from occurring, and subdistribution risks incorporate the impact of competing events, the latter are better suited to answering real-world causal questions.

References

L. E. Cain and S. R. Cole. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident aids or death. *Stat Med*, 28(12): 1725–38, 2009.

Stephen R. Cole, Michael G. Hudgens, M. Alan Brookhart, and Daniel Westreich. Risk. *Am J Epidemiol*, 181(4):246–250, 02 2015.

Stephen R Cole, Jessie K Edwards, Ashley I Naimi, and Alvaro Muñoz. Hidden imputations and the kaplan-meier estimator. *Am J Epidemiol*, 189(11):1408–1411, 2020.

Darios Getahun, Cande V. Ananth, Nandini Selvam, and Kitaw Demissie. Adverse perinatal outcomes among interracial couples in the united states. *Obstetrics & Gynecology*, 106(1), 2005.

T Gilovich, D Griffin, and D Kahneman, editors. *Heuristics and biases: The psychology of intuitive judgment.* Cambridge University Press, New York, 2002.

Robert J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3):1141–1154, 1988.

Sander Greenland. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*, 125(5):761–768, 1987.

M. A. Hernán. The hazards of hazard ratios. *Epidemiol*, 21:13–5, 2010.

Wen-Wen Huang, Wen-Zhi Zhu, Dong-Liang Mu, Xin-Qiang Ji, Xiao-Lu Nie, Xue-Ying Li, Dong-Xin Wang, and Daqing Ma. Perioperative management may improve long-term survival in patients after lung cancer surgery: A retrospective cohort study. *Anesth Analg*, 126(5):1666–1674, 2018.

D Kahneman. *Thinking: Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

Jay S Kaufman. Marginalia: Comparing adjusted effect measures. *Epidemiol*, 21(4):490–493, 2010.

John P. Klein and Melvin L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer, New York, 2005.

G. F. Medley, R. M. Anderson, D. R. Cox, and L. Billard. Incubation period of aids in patients infected via blood transfusion. *Nature*, 328(6132):719–721, 1987.

Ashley I Naimi, Neil J Perkins, Lindsey A Sjaarda, Sunni L Mumford, Robert W Platt, Robert M Silver, and Enrique F Schisterman. The effect of preconception-initiated low-dose aspirin on human chorionic gonadotropin-detected pregnancy, pregnancy loss, and live birth : Per protocol analysis of a randomized trial. *Ann Intern Med*, 174(5):595–601, 2021.

Lior Rennert. *Statistical Methods For Truncated Survival Data*. PhD thesis, University of Pennsylvania, 2018.

K. J. Rothman, S. Greenland, and T.L. Lash. *Modern Epidemiology*. Wolters Kluwer, Philadelphia, PA, 3rd edition, 2008.

Jacqueline E Rudolph, Catherine R Lesko, and Ashley I Naimi. Causal inference in the face of competing events. *Current epidemiology reports*, 7(3):125–131, 2020.

Jacqueline E Rudolph, Abigail Cartus, Lisa M Bodnar, Enrique F Schisterman, and Ashley I Naimi. The role of the natural course in causal analysis. *American Journal of Epidemiology*, (kwab248), 2021.

Séverine Sabia, Aurore Fayosse, Julien Dumurgier, Alexis Schnitzler, Jean-Philippe Empana, Klaus P Ebmeier, Aline Dugravot, Mika Kivimäki, and Archana Singh-Manoux. Association of ideal cardiovascular health at age 50 with incidence of dementia: 25 year follow-up of whitehall ii cohort study. *BMJ*, 366:l4414, 2019.

Michaël Schwarzinger, Bruce G Pollock, Omer S M Hasan, Carole Dufouil, and Jürgen Rehm. Contribution of alcohol use disorders to the burden of dementia in france 2008-13: a nationwide retrospective cohort study. *Lancet Public Health*, 3(3):e124–e132, 2018.

Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable (The Incerto Collection)*. Random House and Penguin, New York, 2007.

S van Buuren. *Flexible imputation of missing data*. Chapman & Hall/CRC, 2nd edition, 2012.

Kirsten Waller, Shanna H. Swan, Gerald DeLorenze, and Barbara Hopkins. Trihalomethanes in drinking water and spontaneous abortion. *Epidemiology*, 9(2), 1998.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, New York, 2004.

Jessica G Young, Mats J Stensrud, Eric J Tchetgen Tchetgen, and Miguel A Hernán. A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in medicine*, 39(8):1199–1236, 2020.