Collecting Data for Specific Goals

Ashley I Naimi

August 2026

Contents

1 Data Collection

We're going to cover what data should be collected to build the information set we need to answer specific types of questions that we may want to answer in epidemiology. Concepts related to the type of information, or data, that we need to collect to construct estimates of specific types of outcomes and outcome contrasts. How operationalizing the exposure or treatment variable leads to specific types of "implied interventions." And we'll get into measurement concerns for our outcomes, exposures, and covariates of interest.

In simple terms, we're going to break down some answers to the following questions:

1) what **could** I be estimating in epidemiology and why?
2) what information or data do I need to collect if I want to estimate the target in 1?

The goal with this lecture is, in some sense, to create a **hierarchy of information**, with risk at the top, and then functions of risk below. The risk function is the most information-rich quantity we can construct. Every measure below it (rates, odds, hazards, and the contrasts we build from them) can be computed from the risk function once we have it. But, except for the hazard function, the reverse does not hold: data collected to support only a summary further down the hierarchy cannot be used to reconstruct the risk function.

Why is this important?: when building an analytic dataset out of an existing database or new study you are designing with actual participants, it's important to know what information you need to target a specific estimand, and where you can make tradeoffs to optimize expenses and effort. Keep in mind, too, where this course is headed: our main estimation tool will be regression modeling; this week establishes the estimands — the targets — that those regression models will be deployed to quantify.

2 Motivating Data Example

Let's adapt some hypothetical data from Cole et al. (2020). The example will be based on a hypothetical study of the time from AIDS diagnosis (the origin, corresponding to time zero) to death from any cause, with 10 participants followed for up to 16 years. We'll adapt the data to avoid late entry and censoring

issues for now (censoring returns later in this lecture). Alongside the event time, we also record for each participant three baseline variables that we will use later: whether they were receiving antiretroviral therapy (ART) at diagnosis (the exposure), their age at diagnosis (a confounder), and their baseline CD4 count (a confounder).

```
data_set <- data.frame(
  id  = 1:10,
  y   = c(3, 5, 6, 8, 9, 11, 13, 14, 15, 16),        # years from AIDS diagnosis to death
  art = c(0, 0, 0, 1, 0, 1, 0, 1, 1, 1),             # antiretroviral therapy at baseline (1 = yes)
  age = c(52, 45, 48, 34, 50, 38, 44, 31, 36, 29),   # age at diagnosis (years)
  cd4 = c(80, 120, 60, 210, 95, 180, 140, 320, 260, 350))  # baseline CD4 count (cells/mm^3)
```

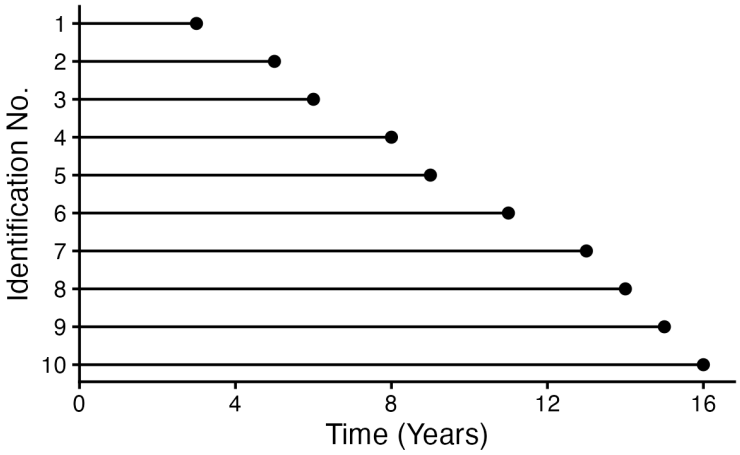

Figure 1: Follow-up time for 10 hypothetical study participants followed from time zero (e.g., AIDS diagnosis, here corresponding to time 0 for everyone) until the event of interest (e.g., death, here for everyone). Solid black lines denote time under follow-up, and solid dots denote the occurrence of the event.

3 Risk

The risk function (cumulative risk, cumulative incidence) is at the top of the hierarchy of information. Epidemiologists chiefly study transitions between states of health and disease, and are therefore concerned, in a well-defined study population, with the time between two events: an origin (here, AIDS diagnosis) and an outcome event of interest (here, death) (Cole et al., 2015).

What to read: Cole et al Risk. *Am J Epidemiol* 2015;181(4):246–250

The risk function (or, more simply, risk) is the probability that the outcome event occurs during a specified period of time. It is a function of follow-up time t, and is the cumulative distribution function of the event times, written

as $F(t) = P(T \leq t)$, where T is a random variable denoting the time from the origin to the event, and where probabilities are defined as proportions in a hypothetical, arbitrarily large population from which our n observed participants are assumed to be a random sample (Cole et al., 2015).

Because risk is the CDF of event times, this function contains all of the information about the random variable (Wasserman, 2004, p20-21), and this is why risk sits at the top of the hierarchy.

Notice what we need to define the risk function here: 1) information on whether the individual experienced the event of interest, and 2) information on exactly *when* the individual experienced the event of interest, relative to some anchor start time.

The risk function is defined for every $t \geq 0$; it is bounded by 0 and 1; and it can never decrease as t increases (anyone who has died by 5 years has also died by 6 years).[1] It is also the complement of the survival function, $S(t) = 1 - F(t)$, which gives the probability of remaining event-free through time t.

[1] The CDF is a monotonically increasing, or non-decreasing function.

Risk is not a single number, so stating that "the risk of death was 50%" is uninterpretable until a time horizon is attached to it. Which values of t we choose to summarize numerically is a matter of scientific context; for example, the 5-year risk of all-cause mortality after antiretroviral therapy initiation (Cole et al., 2015).

Our data show how simple this parameter is to estimate in a "clean" dataset. The 10 participants form a closed cohort followed from a common origin: everyone enters at time zero, no one leaves, and every event time is observed exactly, and there are no ties. Under these conditions, the natural estimate of the t-year risk is a sample proportion: count the deaths at or before t, and divide by 10. By 3 years, 1 of the 10 has died, so $\widehat{F}(3) = 1/10 = 0.1$. By 9 years, 5 of 10 have, so $\widehat{F}(9) = 5/10 = 0.5$ — the 9-year risk is 50%. And because everyone's event has occurred by year 16, $\widehat{F}(16) = 10/10 = 1$. Counting at every value of t traces out the curve shown below.

This curve is sometimes referred to as the "empirical distribution function" or "empirical cumulative distribution function" or "eCDF", which represents the simple act of computing the running proportion of events over time. For our n observed event times $T_1, \ldots, T_n$, the eCDF can be implemented as:

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \le t),$$

where $I(\bullet)$ is an "indicator" function, which takes a value of 1 if the argument $\bullet$ is true, and zero otherwise.

To see exactly what this sum is doing, pick a specific time value (say $t = 9$ years) and evaluate the indicator once for each of our ten event times $T_1, \ldots, T_{10} = 3, 5, 6, 8, 9, 11, 13, 14, 15, 16$, scoring a 1 whenever the event has occurred by year 9 and a 0 otherwise:

$$\begin{aligned}
\hat{F}(9) &= \frac{1}{10}\sum_{i=1}^{10} I(T_i \le 9) \\
&= \frac{1}{10}\Big[\, I(3 \le 9) + I(5 \le 9) + I(6 \le 9) + I(8 \le 9) + I(9 \le 9) \\
&\qquad\qquad + I(11 \le 9) + I(13 \le 9) + I(14 \le 9) + I(15 \le 9) + I(16 \le 9) \,\Big] \\
&= \frac{1}{10}\Big[\, \underbrace{1+1+1+1+1}_{\text{5 event times} \le 9} + \underbrace{0+0+0+0+0}_{\text{5 event times} > 9} \,\Big] \\
&= \frac{5}{10} = 0.5.
\end{aligned}$$

Every participant contributes one term: the indicator is switched on for the five who died by year 9, and off for the five who die after year 9. The running total of these indicators over the ten observations is the height of the risk curve at $t = 9$. Moving t forward in time adds more to the risk because more indicators get switched on, eventually yielding the whole step function. This is the core concept of cumulative risk: at any time t, the height of the curve represents the proportion of the cohort whose event has occurred by t.

Here, we can estimate the risk function by simple counting because the data are ideal. When participants enter the study after the origin (left truncation) or leave it before their event occurs (right censoring), or when other events can preclude the event of interest (competing events), the numerator and denominator of this simple proportion are no longer directly countable. Purpose-built nonparametric estimators exist for each situation — the Kaplan-Meier, extended Kaplan-Meier, and Aalen-Johansen estimators, which we develop in Week 12 — and, more central to this course, regression models can be deployed to target these same quantities (Weeks 6, 12, and 13). This week's job is knowing

which *estimand* is in play, and what information your data must contain for any estimator — nonparametric or regression-based — to reach it.

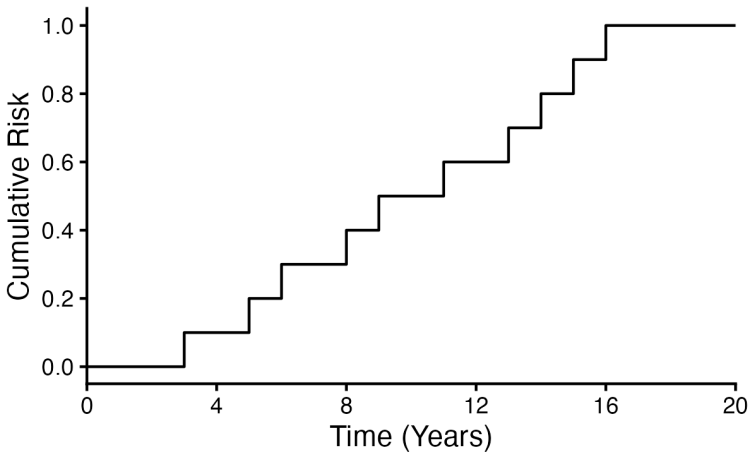

Figure 2: Cumulative risk for the 10 hypothetical study participants. Because every participant is followed from time zero until their event occurs, the risk at any time is simply the proportion of the 10 participants whose event has occurred by that time. The resulting curve is the empirical distribution function: a step function that begins at zero, increases by 1/10 at each observed event time, and reaches one at 16 years, when the last participant dies.

3.1 Risk with Right Censoring

In real settings, event times are sometimes incomplete due to loss to follow-up, or right censoring. Let's keep the same ten participants, entering together at time zero, but now suppose that three of them (at years 5, 8, and 13) leave the study before their death is observed.

Their death times are now **right-censored**: we know only that T_i is *greater than* the time at which we last saw them. Everything else is the same, including the estimand: $F(t) = P(T \leq t)$.

However, we can no longer estimate this estimand using the eCDF, which needs every event time to be observed.

```
data_cens <- data.frame(
  id = 1:10,
  y  = c(3, 5, 6, 8, 9, 11, 13, 14, 15, 16),  # observed time: death OR censoring
  d  = c(1, 0, 1, 0, 1,  1,  0,  1,  1,  1))   # 1 = death observed, 0 = right-censored
```

Why does the eCDF break? Suppose we seek to estimate the 11-year risk. The eCDF instructs us to average $I(T_i \leq 11)$ over the ten participants. But for the two participants censored at years 5 and 8 we cannot evaluate that indicator: they were alive when last seen, and we do not know whether they

died before year 11, so the terms they contribute are not observed and the sum can't be completed (the participant censored at year 13 was, by contrast, still under observation and alive at year 11, so their indicator is a known 0):

$$\hat{F}(11) = \frac{1}{10}\Big[\underbrace{1+1+1+1}_{\text{deaths at } 3,6,9,11} + \underbrace{?+?}_{\substack{\text{censored at } 5,8:\\ \text{dead by } 11?}} + \underbrace{0}_{\substack{\text{censored at } 13:\\ \text{alive at } 11}} + \underbrace{0+0+0}_{\text{deaths at } 14,15,16}\Big]$$

$$= \frac{4+?}{10} \in \{0.4,\ 0.6\}.$$

The values $\{0.4,\ 0.6\}$ are just the two extreme guesses that could arise for the "?"s: counting both censored participants as deaths ($? = 1$) returns 0.6 and, if applied throughout, will *overstate* the risk by treating everyone who left as though they died; keeping them in the denominator but never letting them contribute an event ($? = 0$) returns 0.4 and *understates* the risk.

The classical resolution is an estimator purpose-built for censored data. Kaplan and Meier (1958) were among the first to solve this problem, with an estimator that never asks any individual for an event time they did not supply; applied to these ten participants, it returns $\hat{F}(11) = 0.475$, comfortably inside the bounds the eCDF could not choose between. How it manages this — and the *redistribution to the right* of the censored observations' risk mass that its mechanics imply (Efron, 1967) — is Week 12's subject, where we develop the Kaplan-Meier estimator in full.

For this week, the lesson is about design. Censoring changed nothing about the estimand — $F(t)$ is still the target — but it broke the estimator, and the repaired pairing makes new demands on the data: to deploy an estimator that handles right censoring, we had to have collected each participant's last-seen time and an indicator of whether their follow-up ended in the event or in censoring (the `y` and `d` columns above). An estimand alone is never enough; it must be paired with an estimator that handles the processes in the data, and the data must contain the information that estimator needs.

3.2 Cause-Specific Risk

In the presence of a competing risk, we can choose between the cause-specific or sub-distribution risks.

Referring to our data above, the **cause-specific risk** of AIDS death is the

risk we would see if the competing event could be prevented, holding everyone at risk for AIDS. If, among our ten participants, seven die of AIDS and three of other causes, the cause-specific risk imagines a world in which the other cause(s) of death is(are) switched off, and those three people are kept at risk of dying from AIDS. It answers the question, *what if only the cause of interest operated?* by imagining a world where we could prevent competing events from occurring.

In practice, estimating the cause-specific risk amounts to treating the competing events as though they were censored observations, inside an estimator built to handle censoring: in our example, code the three other-cause deaths as censored, and hand the data to the Kaplan-Meier estimator (Week 12). Doing so re-distributes the risk mass of the competing events onto the event of interest, which is exactly what the switched-off world demands: carried to the end of follow-up, the estimated cause-specific risk of AIDS death reaches 1 — as though, with other causes eliminated, all ten participants would eventually have died of AIDS.

3.3 Sub-Distribution Risk

The **sub-distribution risk** is the probability of dying of AIDS by time t in the real world, where other causes of death are allowed to happen. This can be written as $F(t, k) = P(T \leq t, J = k)$ (Cole et al., 2015; Lau et al., 2009), where $k \in \{1, 2, ...\}$ represents the type of event in the study. In the competing-events literature, this event-specific function is what "the cumulative incidence function" conventionally refers to — the same phrase we used at the top of this lecture as a loose synonym for the all-cause risk function. The two usages coincide when only one type of event is possible; once competing events are present, an unqualified "cumulative incidence" is ambiguous, which is why we prefer *sub-distribution risk* here (see also the terminology note in Week 1).

To make this concrete, take the censored observations from the example in the previous section and re-envision them. Imagine that the three participants who were *censored* at years 5, 8, and 13 are instead **competing** other-cause deaths ($J = 2$), while the other seven remain AIDS deaths ($J = 1$). No one is censored, every participant now has an observed event of one type or the other.

```r
data_ce <- data.frame(
  id    = 1:10,
  y     = c(3, 5, 6, 8, 9, 11, 13, 14, 15, 16),
  cause = factor(c(1, 2, 1, 2, 1, 1, 2, 1, 1, 1), levels = c(0, 1, 2),
                 labels = c("censored", "AIDS", "other")))  # yrs 5,8,13 now competing
```

With every participant's event type observed, the sub-distribution risk at the end of follow-up is directly countable: seven of the ten died of AIDS, so the AIDS sub-distribution risk at sixteen years is $7/10 = 0.7$, and the other-cause sub-distribution risk is $3/10 = 0.3$. No redistribution is involved — each person's risk mass stays with the event they actually experienced. Contrast this with the cause-specific risk of the previous section, which reached 1 by imputing an AIDS death for each of the three people who died of other causes.

Tracing the sub-distribution risk *function* over follow-up time — not just its end-of-study value — takes more care, particularly once censoring is also present. That is the job of the **Aalen–Johansen** estimator (Aalen and Johansen, 1978), the competing-events generalization of the Kaplan–Meier approach, which we develop alongside it in Week 12.

A defining property of sub-distribution risks is additivity: the risks for each specific event at the end of follow-up sum to the all-cause risk, so at the end of follow-up the AIDS and other-cause sub-distribution risks add up to the total risk of death.

Each is a *piece* of the whole distribution, hence "sub-distribution."

3.4 Cause Specific, Sub-Distribution, and Causality

There is a long tradition that assigns the cause-specific risk to studying *etiology* (causality), the sub-distribution risk for developing *predictions* and resource allocation decisions (Lau et al., 2009).

But this is backwards. Reading the cause-specific risk causally requires accepting that an intervention that prevents the competing event from occurring altogether is possible. When that event is death, no such intervention exists, so the cause-specific risk describes a world we can never bring about.

The sub-distribution risk is the risk in the world we actually inhabit, and thus the quantity that a patient or a policymaker actually faces. For that reason the

sub-distribution risk is the more causal of the two (Young et al., 2020; Rudolph et al., 2020). Which one you want is a question about the *target of inference*, decided by the scientific question and by what interventions are conceivable.

Note: Censoring, Competing Risks, Kaplan-Meier, and Aalen-Johansen

The goal here is not to convince you that you must fit KM under censoring or AJ under competing events. Rather, it's to be able to understand that these are processes that must be dealt with when constructing an analytic dataset from source. If competing risks are present in your source data, and you don't collect any information on them, you are subsetting your analysis to only those who stayed alive long enough to experience your event. If you collect information on the competing risks in your data, but then decide to censor them using IPW (because you're not quite sure what else to do with them) you are estimating a cause-specific type of metric (risk difference, risk ratio, odds ratio, etc).

Most of the time, you will not be explicitly interested in the cumulative risk function from a Kaplan-Meier or Aalen-Johansen estimator (though it is nice to know they exist, and how to use them). However, the concepts here of competing risks and censoring must still be dealt with, even if you are simply estmating an odds ratio from a logistic regression model. This should be a key takeaway.

4 Functions of Cumulative Risk

The risk functions that we saw above sit at the top of a hierarchy of measures of occurrence because all other measures of occurrence can be derived from them. There are important mathematical relationships between the probability density function, the cumulative distribution function, survival, hazards, risk, rate, and odds (e.g, Klein and Moeschberger, 2005, Sections 2.2 and 2.3). Here, we start with the cumulative risk and discuss how it contains information on the hazard, rate, odds, and risk.

Importantly, in the presence of competing risks, each of these measures of occurrence can be classified as cause-specific or sub-distribution, depending on how these competing risks are handled. Each of these measures of occurrence is also subject to bias that results from left truncation and right censoring. Finally, in this section we note how and why one should choose between these measures of occurrence: Why avoid hazard and odds? Why focus on risk?

4.1 Hazard

We start with the **hazard**, denoted $h(t)$, and defined formally as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t \mid T \geq t]}{\Delta t}$$

In English, this quantity can be interpreted as the "instantaneous risk" of an outcome at a specific time t among those who actually survived past that time. The hazard is functionally related to the cumulative risk. Specifically, we can rewrite the hazard function as:

$$h(t) = f(t)/S(t)$$

noting that $f(t) = dF(t)/dt$ and $S(t) = 1 - F(t)$.[2]

As a measure of occurrence, the hazard, though commonly used (in the context of Cox proportional hazards regression modeling), is not ideal for a number of reasons. The first is that, as a measure of occurrence, it is somewhat complicated to understand, due in part to the calculus used to define it, but more so to the fact that the numerator conditions on observations whose survival time is greater than the indexing time (which we explain shortly). The second reason is that it is not a good causal estimand. This second reason requires some explanation.

[2] To see more, and why, refer to the video on defining the hazard.

Note: Causal Inference with Statistical Functionals

Causal inference starts by defining an estimand as a summary of a contrast of potential outcomes. For a binary exposure with a time-to-event outcome (and no censoring or competing risks), each individual has two potential outcomes, say $\{T(0), T(1)\}$ – parenthetical notation for the same objects Week 1 wrote with superscripts (as in Y^a); both conventions are common in the literature. These potential outcomes can be thought of as random variables, so they have a joint distribution with two margins: $F_1(t) = P\{T(1) \leq t\}$ and $F_0(t) = P\{T(0) \leq t\}$. These **counterfactual marginals** are the event time distributions, represented as cumulative risks, in the two worlds we compare (i.e., where everybody is exposed/unexposed).

A **statistical functional** is a map from a distribution to a number, $\psi : \mathcal{P} \mapsto \mathbb{R}$ (Kennedy, 2018, 2022). This means that we can use statistical functionals to define a causal effect by choosing the functional and applying it to the counterfactual marginal distributions. For instance, the ten-year risk difference in the context of the time-to-event outcomes can be defined as $\psi = F_1(10) - F_0(10)$. Several similar estimands can be defined as functionals: risk ratios, odds ratios, mean restricted survival time differences, hazard ratios, and many others. However, each can be though of as a *different* functional of the *same* counterfactual marginal distributions, where a *functional* is a function of a function.

Thinking of causal estimands this way is mathematically useful: a causal parameter can be thought of simply as a functional of the counterfactual marginals. For example, in technical point 1.1 of the book by Hernán and Robins (2020) (page 6), they note that "a population causal effect may also be defined as a contrast of functionals (including the median, variance, hazard, or cdf) of counterfactual outcomes." Once F_1 and F_0 are specified, simply contrasting summaries of them allows us to compare two counterfactual worlds, and hence define causal effects.

The hazard is such a function, $h_a(t) = f_a(t)/\{1 - F_a(t)\}$, so from this perspective the hazard ratio is a perfectly good causal estimand. However, as we'll see, many have noted that the hazard ratio incorporates a problem, in that the denominators are different in the exposed and unexposed groups. This issue has led to refer to the hazard ratio as fundamentally "biased" (Hernán, 2010). This leads to an additional set consideration needed when we define what causal effect is. Not only do we need (A) a comparison of counterfactual marginal distributions, but **also** (B) a contrast made *within a common set of units*. As we will see, the hazard ratio does not possess quality B, whereas other measures do.

Suppose we have 5,000 observation's, and, for each unit i we observe their **pair** of potential event times $\{T_i(0), T_i(1)\}$ (drawn **independently** with different marginals).[3] Note that in this illustration, we see **both** potential outcomes for every unit. Thus, we *are not* concerned about the properties of different estimators here; we aren't concerned with threats to validity such as censoring, confounding, or other such issues. We here have the actual potential outcomes

[3] This example was adapted from a social media post by Arman Oganisian at Brown University

time to event values and are focusing on definitional issues (issues with the estimand).

We can plot these potential outcomes for each of the 5,000 units in a scatterplot. We'd never be able to see this in any realistic setting, but, again, we're focusing here on definitional issues that would arise even if we had impossibly perfect data:

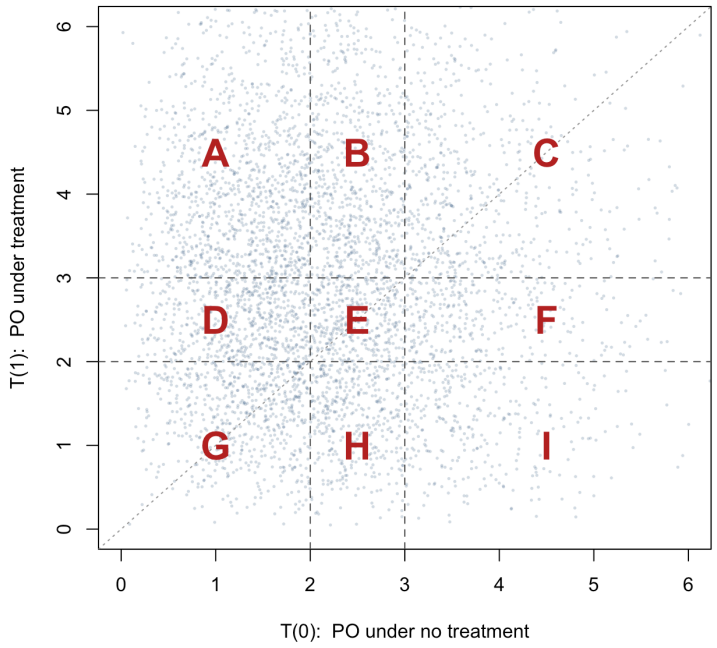

Figure 3: The 5,000 potential outcome pairs, partitioned at $t_1 = 2$ and $t_2 = 3$. Each point is one unit: its horizontal position is the event time that would be observed under control, its vertical position the event time that would be observed under treatment. The dashed lines cut both axes at the two times, producing nine regions A–I, and the dotted diagonal marks equal potential outcomes, so points above it are units the treatment helps. Survival proportions at t_1 take their denominator from the entire square; the two hazards at t_1 take theirs from a horizontal band (treated) and a vertical band (control) – two different sets of points.

We can also divide this plot into different sectors, which will help us convey the main points. We can use this plot to construct a number of estimands, such as the risk of death by two years that would be observed if everyone was exposed, versus if everyone was unexposed. This estimand can be constructed as the complement of the proportion surviving past $t_1 = 2$ under each intervention:

$$r_1 = 1 - \frac{A + B + C + D + E + F}{n}, \qquad r_0 = 1 - \frac{B + C + E + F + H + I}{n}$$

Note that if we took the difference $r_1 - r_0$, we obtain the causal risk difference. This measure is causal because (A) it is a comparison of counterfactual

marginal distributions (namely, a simple function of the marginal distributions of $T(0)$ and $T(1)$), AND (B) this contrast is made *within a common set of units* (namely, among all individuals n). To emphasize the latter point, the contrast is defined for all individuals in the population, because the denominator of the risk estimate is the same in both r_0 and r_1: n.

Consider instead a simplified (discrete time) hazard estimated at time t_1. Here, we want to use the figure to compute the proportion who fail in $(2, 3]$ among those still at risk at 2. We use a discrete bin here $(2, 3]$, to avoid having to take the limit.[4]

$$\text{haz}_1 = \frac{D+E+F}{A+B+C+D+E+F}, \qquad \text{haz}_0 = \frac{B+E+H}{B+C+E+F+H+I}$$

Note here how the denominators of each hazard represent different sets of people. This is an example illustrating why the hazard ratio is not an ideal causal estimand. It fails to adhere to property (B): that the contrast is taken within a common set of units. Though this is not guaranteed to be a problem in every setting, it creates uncertainty about who makes it into the denominator, and who doesn't. This issue is why Hern'{a}n refers to the hazard ratio as possessing a "built-in selection bias" (Hernán, 2010, p 13). Region **H** represents units who would survive past $t = 2$ under control, but not under treatment. Region **H** is in the denominator of haz_0, but is not in the denominator of haz_1.

[4] Note that the numerical value of the discrete-bin hazard ratio here is not the same number as the numerical value of the continuous time instantaneous hazards ratio for this data generating mechanism. Binning over $(2, 3]$ is a coarse approximation to the derivative. However, the conceptual point about denominators is true for both the discrete and continuous time hazard ratio, while the discrete time representation makes it easier to see practically.

Note: A Note on the "Bias" of the Hazard Ratio

An **estimand** is a precisely defined feature of a population or of one or more counterfactual distributions. For example, the two-year causal risk difference, $P\{T(1) \leq 2\} - P\{T(0) \leq 2\}$, is an estimand. An **estimator** is a function that, when applied to data, returns a number that is intended to quantify the estimand. For example, if we define a regression model that encodes the two-year causal risk difference as a parameter of the model, we can use (e.g.) maximum likelihood estimation to estimate this parameter, and thus the estimand.

Technically, epidemiologic bias is often expressed as a problem of **statistical inconsistency**. Given a chosen target estmand ψ, an estimator $\hat{\psi}_n$ is consistent when it converges to ψ as the sample size grows under the relevant data-generating process. If confounding, selection, measurement error, model misspecification, or inappropriate handling of censoring means that $\hat{\psi}_n$ instead converges to some other value, then the estimator is inconsistent for the target (or epidemiologically biased). This is the large-sample sense in which we say the estimate is biased. It differs from the narrower finite-sample definition of bias, $E(\hat{\psi}_n) - \psi$: an estimator can have finite-sample bias and still be consistent, or be approximately unbiased in a particular finite sample yet inconsistent in the long run.

This terminology matters for the hazard ratio. A population hazard ratio is an **estimand**: it is a functional of the two event-time distributions, such as $h_1(t)/h_0(t)$ at a time t (or a model-based summary of those hazards). Therefore, technially speaking, a hazard ratio cannot itself biased or unbiased; it has no sampling distribution and cannot converge to the wrong value. Bias and consistency are properties of an estimator *relative to an estimand*. Thus, it is not precise to say that the hazard ratio itself is "biased." A Cox-model hazard-ratio estimator may be consistent or inconsistent for its stated hazard-ratio estimand under its assumptions.

The substantive concern raised by Hernán (2010) is therefore about the quality of the hazard ratio as a measure of causal effect. Because we take that a good causal estimand possesses the two qualities we've referred to earlier: (A) a contrast of counterfactual marginals, and (B) obtained in the same population, the hazard ratio is not a good causal target because B is violated. But, technically, it is not "biased".

4.2 Risk, Rate, Odds, and Other Measures

We've just connected the hazard function to the survival or cumulative distribution (risk) function. We can actually do the same for other measures of occurrence (without calculus!), including the risk, the rate, the odds, as well as a host of other measures. Most of these are much easier to compute from the cumulative risk $F(t) = P(T \leq t)$.

The first is **risk**, which is simply the cumulative risk function at some specified time t. To compute it you need, for each person, both *whether* and *when*

the event occurred.

A second measure is the **rate** (incidence rate), which divides the number of events by the *person-time* at risk. It measures how fast events accrue per unit of time spent at risk. Our ten participants each contribute person-time equal to their follow-up, $3 + 5 + \cdots + 16 = 100$ person-years, and all ten die, so the overall death rate is $10/100 = 0.1$ per person-year.

It might seem that the rate requires more than the risk (an additional quantity, person-time at risk, for the denominator). However, person-time is already encoded in the survival function. A participant followed to time t contributes $\min(T_i, t)$ person-time, and averaging that quantity over the cohort is the same operation as accumulating the survival curve:

$$\frac{1}{n}\sum_{i=1}^{n}\min(T_i, t) \;=\; \int_0^t \hat{S}(u)\,du.$$

That is, in words, the area under the survival curve up to t *is* the average time at risk per participant. Because in our example $\hat{S}$ is a simple step function, the area under it is just:

$$\int_0^{16} \hat{S}(u)\,du = \underbrace{1.0\times 3}_{[0,3)} + \underbrace{0.9\times 2}_{[3,5)} + \underbrace{0.8\times 1}_{[5,6)} + \cdots + \underbrace{0.1\times 1}_{[15,16)} = 10 \text{ years},$$

which, multiplied by the $n = 10$ participants, returns the 100 person-years. Since the number of events by time t is likewise $n\hat{F}(t)$, the n's cancel, and the rate turns out to be a ratio of two features of the same risk function:

$$\widehat{\text{rate}}(t) = \frac{n\hat{F}(t)}{n\int_0^t \hat{S}(u)\,du} = \frac{\hat{F}(t)}{\int_0^t \hat{S}(u)\,du} = \frac{\hat{F}(t)}{\int_0^t \{1-\hat{F}(u)\}\,du}.$$

Evaluated at sixteen years, this returns $\hat{F}(16)/10 = 1/10 = 0.1$ per person-year. By nine years, five deaths have accrued over 76 person-years, and the formula gives $\hat{F}(9)/\int_0^9 \hat{S}(u)du = 0.5/7.6 = 0.066$ per person-year. It is the risk function divided by its own accumulated complement.

However, the rate *does* require the **whole** risk function over the interval. Unlike the risk at one time point, the rate's denominator needs S, and therefore F, at every $u \le t$. A rate is thus a summary of the entire path the risk function traced on its way to t, compressed into a single number per unit of time. This

is also the sense in which the hazard of the previous section is the rate's instantaneous limit: writing the numerator as $F(t) = \int_0^t h(u)S(u)\,du$ reveals the rate to be a survival-weighted average of the hazard over the window, collapsing to the hazard itself when the hazard is constant. Rates can be estimated using a number of different methods, for example Poisson regression models (covered in Week 6).

The **odds** of the event by time t is the risk divided by its complement, $F(t)/[1 - F(t)] = F(t)/S(t)$. At nine years our risk is 0.5, so the odds are $0.5/0.5 = 1$: one death for every survivor. Odds carries the same information as risk (each is a deterministic function of the other), so no extra data are needed to move between them. However, they are on completely different scales (the scale on which logistic regression operates, covered in Week 6). The price we pay when using odds as a measure of occurrence is interpretability: an odds, and especially an odds *ratio*, is harder to reason about than a risk. This is both because the odds is (by some accounts, at least) less intuitive than the risk, and is also subject to non-collapsibility, which complicates interpretation, and which we cover later in the course.

Many other summaries are available, and all are functions of the same $F(t)$. The **restricted mean survival time** averages survival over a fixed window and is measured in the natural units of time; over our sixteen-year window it is the area under $S(t)$, which here works out to 10 years — equivalently, the average time to death. **Quantiles** of the event-time distribution answer "by when has a given fraction had the event?"; the median survival for our cohort is 9 years, the time by which half have died. The lesson of the hierarchy is that the choice among these is a choice of *summary*, made after the risk function is in hand — and, crucially, that a study which collected only a coarse summary (say, ever/never events with no timing) cannot be made to yield a rate, a hazard, or a median after the fact.

5 Risk Function Contrasts: The Exposure and the Implied Intervention

We are often interested in how measures of occurrence compare between individuals defined according to some treatment or exposure status. Binary exposures or treatments are often considered to be the simplest. However, care

must still be exercised in constructing and using binary treatment variables. Consider, two examples.

The first is an example from the NuMoM2b-Heart Health Study, in which 3,281 women provided data on periceptional diet. In this sample of women, usual dietary intake in the 3 mo before conception was assessed at between 6 and 13 weeks of gestation. The tool used to collect these data was a self-administered modified Block 2005 food frequency questionnaire (FFQ). This tool assessed 59 nutrients from ≈120 food and beverage items. The questionnaire was modified to reflect a 3-mo period, and questionnaires were sent to Block Dietary Data Systems (Berkeley, CA) for scanning and nutrient analysis.

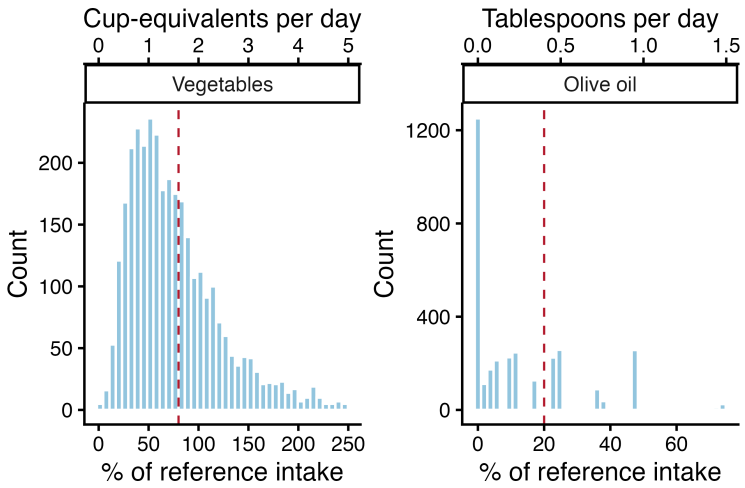

Figure 4: Vegetable and olive-oil intake distributions among n = 3,281 in the NuMoM2b-Heart Health Study 2013-2019. Histograms of visit-1 intake, with the bottom axis on the percent-of-reference scale used to define the exposure and the top axis in native units (100% of reference = 2 cup-equivalents per day for vegetables, 2 tablespoons per day for olive oil). The dashed line is an exposure threshold of interest, read on either axis: 80% (1.6 cup-equivalents per day) for vegetables and 20% (0.4 tablespoons per day) for olive oil. The highest 1% of values for olive oil is omitted from the display

We are here interested in periceptional vegetable and olive oil consumption, and whether they might be associated with metabolic syndrome several years after giving birth. Figure 5 shows the distribution of vegetable and olive oil consumption in these women, each with two x-axes. For both vegetable and olive oil consumption, the lower axes represent percent adherence to the guidelines for a Mediterranean diet. For example, 100% here represents consuming two cups of vegetables per day, which is what you would need to consume if one adopted a Mediterranean diet. In contrast, to meet the olive oil consumption guidelines, women would need to consume at least 4 tablespoons of olive oil per day. However, this is a fairly uncommon level of olive oil consumption in American populations. Since no individuals in the data consumed this much olive oil, we modified the guideline to at least 2 tablespoons, and the lower axis represents adherence to this modified guideline. The upper axes of each

plot represent actual consumption in the based on cup-equivalents per day for vegetables[5] and based on tablespoons consumed per day.

Note here how the two axes represent different operationalizations of the same distribution. This is important to recognize. When creating an analytic dataset with an exposure variable, it's essential to note exactly what contrast you are interested in as a researcher and why. In the NuMoM2b-HHS example, we can be interested in whether an individual consumes vegetables / olive oil in absolute terms per day (cup-equivalents and tablespoons) OR relative to some standard, such as the Mediterranean diet.

Alternatively, consider a different type of exposure: quitting smoking, as captured in the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS), the dataset used throughout the book by Hernán and Robins (2020).[6] The NHEFS was jointly initiated by the National Center for Health Statistics and the National Institute on Aging to investigate how the clinical, nutritional, and behavioral factors measured in NHANES I relate to subsequent morbidity and mortality.

The subset used in the book consists of 1,629 cigarette smokers aged 25-74 who completed a baseline visit in 1971-75 and a follow-up visit in 1982. The exposure `qsmk` is defined by comparing smoking status across the two visits: the 428 participants who reported having quit by the 1982 visit are "treated," and the 1,201 who continued smoking are "untreated." Because the exposure is ascertained at the 1982 visit, everyone in the dataset was necessarily alive until then; all 318 recorded deaths occur between 1983 and 1992, as the figure below shows.

Now look at what the figure does *not* show: for the first twelve years of the displayed window, there is not a single death. This is not because smoking or quitting protected anyone in the 1970s — it is a feature of how the exposure was built. Because `qsmk` is ascertained at the 1982 visit, a person appears in this dataset — in *either* exposure group — only if they survived to 1982 to be classified. Every year of person-time drawn from the 1971 origin up to 1983 is therefore *immortal*: no one who died during those years can be in the data at all. Starting the follow-up clock at the 1971 baseline would mean classifying individuals by information from their future (smoking status roughly a decade later), while crediting both groups with more than a decade of guaranteed survival. This is exactly why the analyses in Hernán and Robins (2020) anchor

[5] A "cup-equivalent" is a measure that takes into account volume. For example, a standard cup of spinach is not really the same as a standard cup of chopped cabbage. Cup-equivalents are meant to standardize across these types of discrepancies.

[6] The book and the NHEFS data are freely available at `https://miguelhernan.org`.

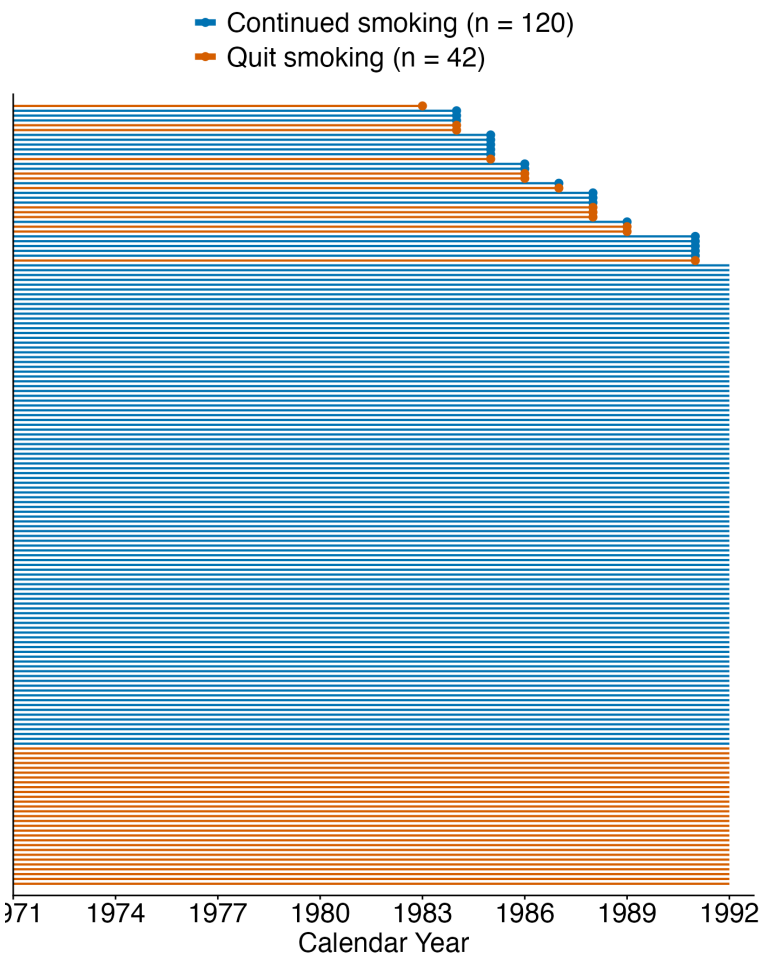

Figure 5: Follow-up for a 10 percent random sample of NHEFS participants, drawn separately within each level of quitting status, from a common origin in 1971 to the administrative end of follow-up in 1992. The sample comprises 162 of the 1,629 participants: 120 who continued smoking and 42 who quit. Each horizontal line is one participant, running from 1971 to the year of death for the 33 who died, where a circle marks the death, and to 1992 for the 129 who did not. Lines are colored by quitting status (qsmk) and ranked by the end of follow-up, shortest at the top and longest at the bottom, with ties broken by quitting status. No deaths are recorded before 1983, so the first twelve years of the displayed window contain no events.

time zero at the 1982 visit instead: at that moment the exposure is fully determined, everyone in the dataset is alive, and mortality follow-up runs cleanly from 1983 through 1992. Even then, note that the quitters at this time zero are *prevalent* quitters — they stopped at some unknown point during the preceding decade — a residual mismatch that the new-user versus prevalent-user question below is designed to catch.

What's essential to takeaway from these descriptions of "the exposure" is the implied intervention that arises from the variable definitions and coding choices. In coding an exposure for a research project, there are several questions that must be considered carefully, that involve more than just "how is the exposure defined?" The key considerations include:

- What is the contrast implied by the intervention (in the exposed, in the referent) that I am interested in studying?
- How (or can) my data be used to construct a variable that represents this contrast?
- How was the variable that I am using for construct the contrast measured? Is the measurement tool consistent with the contrast/intervention of interest?
- If I could assign this exposure, exactly what would the process be? If I actually implemented this process, could I expect to see the data that I have?
- How many people in my data can actually be classified as exposed/referent in the contrast?
- Is the implied intervention a one-time act at time zero, or a strategy sustained over follow-up?
- Does my variable identify people starting the exposure, or people already using it? Can the start time be demarcated? (New-user / prevalent-user)
- What process generated each person's exposure value. Is that process related to prognosis?
- Does classifying a person require information from after time zero, or surviving to a later measurement?

5.1 Exposure Timing

When the exposure is measured, relative to time zero, matters as much as *what* it is. In our cohort ART is recorded at baseline — at the origin, before any follow-up accrues — and that is deliberate. An exposure defined using information from *after* time zero can manufacture immortal time bias: a person must survive long enough to become "exposed," which guarantees them event-free time that is then misattributed to the exposure. The NHEFS example above has exactly this structure: with a 1971 time zero, `qsmk` is information from the future — no one, quitter or continuer, could even be classified without surviving to the 1982 visit — and the twelve event-free years in the figure are that guaranteed survival made visible. Week 1 made the same point through time zero and the target trial — eligibility and assignment must be settled at the origin. The data-collection consequence is concrete: to define the exposure cleanly you must capture it as of time zero, and record its timing, not merely its value.

5.2 Continuous Exposures

Many exposures are continuous — here, the baseline `cd4` count or `age`. A continuous exposure does not name two groups, so the implied intervention has to be stated explicitly, and different codings imply different interventions. It's usually not, and instead we often express interest in a dose-response curve.Entering `cd4` linearly implies a *per-unit* contrast, the effect of shifting CD4 by one cell, assumed the same everywhere along the scale. A threshold coding (CD4 below vs. above 200, say) implies a *categorical* intervention and discards variation within each side. A spline (Week 9) implies a smooth dose-response and lets the per-unit effect vary across the range. None is uniquely correct, because each answers a different question and imposes its own data requirement. A threshold analysis needs only which side of 200 a person falls on, but a dose-response needs the full distribution.

Note: Target and Nuisance Functions in Health Research

Suppose in our running example we modeled the five-year risk of death as a function of ART (A) and the two baseline confounders, age and CD4 (gathered into C): $\text{logit}\,[P(T \leq 5 \mid A, C)] = \alpha_0 + \alpha_1 A + \alpha_2 C$. The right hand side of this model can be viewed as a combination of two functions with very different jobs, $\mu(A) + \eta(C)$. The first piece, $\mu(A) = \alpha_1 A$, relates the exposure to the outcome and carries the contrast we actually care about: this is the **target function**. The second piece, $\eta(C) = \alpha_0 + \alpha_2 C$, captures the relationship between the confounders and the outcome, and exists only to adjust for confounding: this is a **nuisance function**, in the language of the semiparametric estimation literature (Tsiatis, 2006) — a function we must estimate on the way to the target, without being substantively interested in it.

This distinction resolves what might otherwise look like inconsistent advice about coding. Variables in the target function should be coded for *interpretation*: as we have seen, coding the exposure is deciding the implied intervention, so the scientific question settles it. Variables in the nuisance function should be coded for *performance*: their coefficients will never be interpreted, so considerations of information loss, efficiency, and bias take over, and we are free to spend flexibility there. Splines for a continuous confounder, or standardized (z-scored) covariates, are perfectly acceptable in the nuisance function even though the resulting coefficients mean nothing on their own. The same moves can be hazardous on the other side of the divide: a z-scored *exposure*, for instance, entangles the contrast of interest with the exposure's own standard deviation, changing the implied intervention.

Collapsing the distinction produces the **Table 2 fallacy** (Westreich and Greenland, 2013): presenting exposure and confounder coefficients from a single model side by side, and reading each as an "effect." The study was designed to identify the effect of the exposure — the confounders were chosen, measured, and collected for exactly that purpose. The "effects" of the confounders themselves are generally not identified by that same design; each would require its own adjustment set, and in effect its own study. The nuisance function is there to help us estimate the exposure effect; it is not itself a finding.

The idea extends beyond confounder terms in an outcome model. The censoring models behind inverse probability of censoring weights (Cain and Cole, 2009), and the adjustment machinery of Week 8 more generally (e.g., propensity score models), are also nuisance functions: models fit not because their parameters answer our question, but because the target cannot be reached without them. The data-collection lesson is symmetric. Measure and code the exposure so that the target function encodes the contrast you want; collect covariates rich enough, and at the right time, for the nuisance functions to do their job.

6 Covariate Collection and Coding

Beyond the exposure and the outcome, a study must decide *which other variables to collect, when, and why*. The role a variable plays determines whether collecting it helps or hurts. A **confounder** — a common cause of exposure and outcome — must be collected and adjusted for; omitting it biases the contrast. In our cohort, baseline `cd4` and `age` are prognostic and plausibly related to who started ART, so a crude ART-versus-no-ART comparison mixes the treatment effect together with baseline differences between the groups; collecting CD4 and age is what makes adjustment possible at all (the machinery of adjustment is Week 8's subject). A **mediator** — a variable on the causal path from exposure to outcome — should generally *not* be adjusted for when the total effect is the target, because conditioning on it blocks part of the effect we are trying to measure. A **collider** — a common effect of two variables — should not be conditioned on, because doing so can *create* an association where none existed.

The practical rule for this course is that covariate roles are design decisions, ideally settled with a causal diagram *before* collection. Two reasons echo the hierarchy of information. First, a variable you failed to measure at baseline cannot be recovered later. Second, a variable measured but mis-timed — a "confounder" recorded after the exposure has already acted — may be unusable, or worse, may actually be a mediator or collider in disguise. Coding matters too: a confounder collapsed into a few coarse categories only partially adjusts, leaving residual confounding, in the same way that a coarsely coded exposure only partially defines the intervention.

7 Measurement Concerns

Finally, every variable is measured with some error, and the *kind* of error often matters more than its size. **Misclassification** of a categorical variable and **measurement error** in a continuous one degrade the outcome, the exposure, and the covariates in different ways. Error that is *non-differential* — unrelated to the other variables — usually, though not always, biases an exposure–outcome contrast toward the null, weakening what you are able to detect. *Differential* error — for instance, an outcome ascertained more aggressively among the

exposed — can bias in either direction and is far more dangerous.

The concrete consequences run right through this week's material. If `cd4` is measured with error, adjustment for it is incomplete and residual confounding remains — a covariate measured badly is a covariate only partly collected. If the cause of death is misclassified, the split between AIDS and other-cause deaths shifts, and with it the cause-specific and sub-distribution risks. If ART status is self-reported, exposure misclassification blurs the treated and untreated groups and attenuates the contrast. The reliability and validity of each measurement instrument should be established as part of the design, not discovered in the analysis: measurement is where the hierarchy of information meets the real world, and the cleanest estimand is only ever as good as the data collected to estimate it.

8 Conclusion

This week reframed data collection as a question of *targets*. Decide what you could estimate — a risk, a rate, an odds, or a hazard; a cause-specific or a sub-distribution risk; a contrast under some implied intervention — and that decision dictates exactly what to collect and how to code it. Risk sits at the top of the hierarchy because everything below is a function of it, but each function, each contrast, and each adjustment demands its own information, collected at time zero and measured well.

Lab 1 puts this into practice. Starting from raw data, you will apply eligibility criteria, set time zero, code the exposure and covariates, define follow-up, and compute the cumulative risk, rate, and odds on the cohort you build — the mechanical counterpart of the conceptual choices laid out here. Week 3 then turns to what happens when, rather than following a whole cohort, we *sample* on the outcome.

References

Odd O. Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978.

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. *American Journal of Epidemiology*, 189(11):1408–1411, May 2020.

Bradley Efron. The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853. University of California Press, Berkeley, 1967.

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

M. A. Hernán and JM Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 53(282):457–481, 1958.

Edward H. Kennedy. Semiparametric theory. In *Wiley StatsRef: Statistics Reference Online*, pages 1–7. John Wiley & Sons, Ltd, 2018.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv:2203.06469 [stat]*, 2022.

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

Bryan Lau, Stephen R. Cole, and Stephen J. Gange. Competing risk regression models for epidemiologic data. *Am J Epidemiol*, 170(2):244–256, 2009.

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, September 2020. ISSN 2196-2995. DOI: 10.1007/s40471-020-00240-7.

Anastasios A. Tsiatis. *Semiparametric theory and missing data*. Springer, New York, 2006.

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

Daniel Westreich and Sander Greenland. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4):292–298, 2013.

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, April 2020.