Causal Inference

Ashley I Naimi

Spring 2021

Contents

- Risk Under Treatment
- Risk Under Competing Events
-

Outline

Causal Inference

- Correlation and Causation
- Introduction to Causal Inference
- Complex Longitudinal Data
- Potential Outcomes Notation
- Estimand, Estimator, Estimate
- Identifiability of the Average Treatment Effect

  a. Counterfactual Consistency

  b. No Interference

  c. Excheangability

  d. Correct Model Specification

  e. Positivity

- Non-identifiability: Bounding Effects

1 Intro

In the last section, we discussed the concepts of a cohort, censoring and truncation, the relation between cumulative distribution functions (CDFs, a.k.a cumulative incidence functions, cumulative risk, other), and how to use the Kaplan-Meier estimator to obtain an estimate of the CDF. In this section, we'll get into how to extend these concepts into several other areas. In particular, we'll discuss how to quantify the effect of an exposure of interest in terms of risk, and what happens when competing events are present when we want to quantify risk and exposure effects. We'll start with a basic introduction to causal inference.

2 Correlation and Causation

In the *The Grammar of Science,* Karl !Pearson1911 wrote: "[b]eyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect." He suggested that rather than pursue an understanding of cause-effect relations, scientists would be best served by measuring correlations through tables that classify individuals into specific categories. "Such a table is termed a contingency table, and the ultimate scientific statement of description of the relation between two things can always be thrown back upon such a contingency table."

Over a century later, a majority of statistics courses treat causal inference by simply stating that "correlation is not causation." This treatment is hardly sufficient, for at least two reasons: 1) As scientists, our primary interest is (should be) in cause-effect relations; 2) People continue to conflate correlation with causation[1]. For both of these reasons, we very much need to **clarify the conditions that would allow us to understand causality better.** This is what "causal inference" is all about.

I adopt the view that **the causal and statistical aspects of a scientific study should be kept as separate as possible.** The objective is to first articulate the conditions under which causal inference is possible, and then to understand what statistical tools will enable us to answer the causal question.[2] Causal inference tells us what we should estimate, and whether we can. Statistics

[1] Daniel Westreich and I reviewed a book whose authors were so caught up in the allure of "Big Data" was so strong, they thoroughly forgot that correlation $\neq$ causation. See !Naimi2014d

[2] Loosely speaking: Causal inference is the "what?" Statistics is the "how?" Epidemiology is the "why?"

tells us how to estimate it. By implication, we should avoid treating statistical models as if they were causal. For example, the practice of reading the risk ratio, odds ratio, or risk difference for an exposure of interest from a generalized linear (statistical) model[3] will sometimes work under very specific conditions, but is not the best approach for quantifying exposure effects [^!^Naimi2020]. In this section, we will cover exactly how the cumulative risk function can be integrated into this framework.

An additional element that we won't have an opportunity to get into in this course is how we can avoid making unnecessary assumptions when estimating causal effects. Specifically, the objective is to avoid imposing unnecessary parametric forms[4] on the causal models that we believe are generating the data. Machine learning in particular is central to this idea of not "imposing unnecessary parametric forms," which is one reason why it's becoming so popular.

3 Introduction to Causal Inference

"Causal inference" deals primarily with the formal mechanisms by which we can combine data, assumptions, and models to interpret a correlation (or association) as a causal relation.[5] The framework by which we define what we mean by "causal relation" or "causal effect" is the **potential outcomes framework**.

A central notion in the potential outcomes framework is the counterfactual. This notion stems from the intuitive and informal practice of interpreting cause-effect relations as **circumstances (e.g., health outcomes) that would have arisen had things (e.g., exposures) been different**.

While this intuition serves an important purpose, it is not sufficient for doing rigorous science. Suppose we ask: "what is the effect of smoking on the 5-year cumulative CVD risk, irrespective of smoking's effect on body weight?" This question may seem clear and intuitive. To answer this question, we would do a study in which we collect data, enter these into a computer, perform some calculations, and obtain a number (we'd usually like to interpret as the "effect").

But there is a problem.[6] The calculations performed by the computer are **rigorously defined (i.e., unambiguous) mathematical objects**. On the other hand, **English language sentences about cause effect relations are ambigu-**

[3] or the hazard ratio from a Cox model, or the mean ratio from a Poisson model, or host of other types of regression models

[4] e.g., additivity, linearity, distributional, and other.

[5] There are a number of introductory books and articles on causal inference in the empirical sciences. Here are some excellent options: [!]Hernan2015, [!]Pearl2016, [!]Imbens2015

[6] This problem was articulated by Robins 1987, and I am using a version of the example from his paper.

ous. For example, the “effect of smoking” can mean many different things:

- All people smoke any tobacco ever versus no people smoke tobacco ever.
- All people smoke 3 cigarettes per day versus all people smoke 2 cigarettes per day.
- All people who have smoked any tobacco in the last 15 years cease to smoke any tobacco whatsoever.

Similarly, “irrespective of” can mean a number of things:

- The effect of smoking on CVD risk that would be observed in a hypothetical world where smoking did not affect body mass?
- The effect of smoking on CVD risk if everyone were set to “normal” body mass?
- The effect of smoking on CVD risk if everyone were held at the body mass they had in the month prior to study entry?

But the numerical strings of data and the computer algorithms applied to these data are well defined mathematical objects, which do not admit such ambiguity. Depending on several choices, including the data, how variables are coded, and the modeling strategy, the computer is being told which question to answer. There is a lot of potential uncertainty in the space between the English language sentences we use to ask causal questions, and the computer algorithms we use to answer those questions. Causal inference is about clarifying this uncertainty.

4 Potential Outcomes Notation

The building blocks for causal inference are **potential outcomes** [^!^Rubin2005].

Importantly, these are conceptually distinct from **observed outcomes**. That is, the outcome that one might observe in a dataset is not the same as the potential outcome.

Potential outcomes are functions of exposures. For a given exposure x, we will write the potential outcome as Y^x.[7] **This is interpreted as “the outcome (Y) that would be observed if X were set to some value x”**. For example, if X is binary [denoted $X \in (0,1)$], then Y^x is the outcome that would be observed

[7] Alternate notation includes: Y_x, $Y(x)$, $Y \mid Set(X = x)$, and $Y|do(X = x)$.

if $X = 0$ or $X = 1$. If we wanted to be specific about the value of x, we could write $Y^{x=0}$ or $Y^{x=1}$ (or, more succinctly, Y^0 or Y^1).

> **Study Question**:
> Suppose you collect data from a single person and find that they are exposed. Can you interpret the outcome you observe to be the potential outcome that would have been observed had they been exposed? Why or why not?

When the exposure and/or outcome are measured repeatedly over follow-up, notation must account for that. We thus use subscripts to denote when the variable was measured. For example, if the exposure is measured twice, we can denote the first measurement X_0 and the second X_1. Additionally, we use overbars to denote the history of a variable over follow-up time. For example, $\overline{X}_1$ denotes the set $\{X_0, X_1\}$. More generally, for some arbitrary point over follow-up m, $\overline{X}_m$ denotes $\{X_0, X_1, X_2, \dots X_m\}$. We can then define potential outcomes as a function of these exposure histories: For two exposure measurements, $\overline{X}_j = \{1, 1\}$, $Y^{\overline{x}_j = \overline{1}}$ is the outcome that would be observed if X_0 were set to 1 and X_1 were set to 1.

Estimand, Estimator, Estimate

Causal inference starts with a clear idea of the effect of interest (the target causal parameter). To do this, it helps to distinguish between estimands, estimators, and estimates.

STUDY QUESTION: You are familiar with the well known odds ratio equation for a 2×2 table: (ab/cd). Is this an estimand, estimator, or estimate?

The **estimand** is the (mathematical) object we want to quantify. It is, for example, the causal risk difference, risk ratio, or odds ratio for our exposure and outcome of interest. In our smoking CVD example, we might be interested in:

$$E(Y^1 - Y^0), \quad \frac{E(Y^1)}{E(Y^0)}, \quad \frac{Odds(Y^1 = 1)}{Odds(Y^0 = 1)},$$

where $Odds(Y^x = 1) = E(Y^x)/[1 - E(Y^x)]$, and where $E(.)$ is the expectation operator taken with respect to the total population.[8] There are

[8] Throughout this course, if the outcome Y is binary, then $E(Y) \equiv P(Y = 1)$. Or, the expectation of Y is equivalent to the probability that $Y = 1$. For the more technially oriented,

$$E(Y) = \int y f(y) dy$$

where $f(y)$ is the probability density function of Y.

many other estimands besides these.

All of the above estimands represent **average treatment effects** (on the risk difference, risk ratio, and odds ratio scale, respectively). This effect is also referred to as a marginal treatment effect, because it averages (or marginalizes) the effect over the entire sample. For instance, if we consider the risk ratio, it is easy to show that[9]

$$E(Y^1 - Y^0) = \sum_{i=1}^{N} Y_i^1 - \sum_{i=1}^{N} Y_i^0$$

However, we may want to estimate this effect in a subset of the population. For instance, $E(Y^1 - Y^0 \mid C = c)$ is the effect of $x = 1$ versus $x = 0$ among those with $C = c$. There are many different conditional treatment effects, this latter one being the simplest. Another common conditional treatment effect is the effect of treatment on the treated (ETT):

$$E(Y^1 - Y^0 \mid X = 1)$$

This effect compares the outcomes that would be observed if the exposure were set to 1 (Y^1) versus if the exposure were set to 0 (Y^0) among those who were observed to be exposed in the sample ($X = 1$).

To illustrate the relevance of this effect, consider the following (entirely fictional) scenario: During gestation of a high-risk pregnancy, two clinical options are available to manage the risk of death: induction of premature delivery induction and expectant management. Suppose a researcher is interested in quantifying the effect of inducing delivery prematurely on fetal and infant death. This researcher collects data on a cohort of high-risk pregnant women, including whether delivery was induced prematurely, fetal/infant death, and a host of confounding variables. All parties involved agree the study is designed perfectly (no confounding, measurement error, loss to follow-up). They calculate the average treatment effect of premature delivery induction on fetal and infant death on the risk difference scale:

$$E(Y^1 - Y^0) = 0.15$$

This researcher concludes that, if all high-risk pregnancies were induced prematurely ($X = 1$), 15 more out of every 100 would end in death, relative to what

[9] Recall that Y^x is not the observed (or sample) value of the outcome, so how do we actually get this average? When we discuss identifiability, we will see how we use observed data to quantify these contrasts.

would happen if all high-risk pregnancies were left to expectant management ($X = 0$). In light of this incredibly high excess risk of death, this researcher would naturally advise abandoning the practice of premature delivery induction entirely.

Another researcher questions the relevance of the average treatment effect. They argue that physicians would never induce delivery prematurely in all versus no high-risk pregnancies. Rather, the more interesting question is: **for those women whose pregnancies were actually induced**, what would the risk of death have been had they not been induced? This researcher thus calculates the effect of treatment on the treated:

$$E(Y^1 - Y^0 \mid X = 1) = -0.05$$

This other researcher concludes that, among those whose pregnancies were actually delivered prematurely, the risk of death would have been higher had they not been delivered prematurely.

This example demonstrates the fundamental difference between the ATE and the ETT: for those high-risk pregnancies that were not induced prematurely, the act of inducing premature delivery would not be beneficial. But for those high-risk pregnancies that were induced prematurely, the act of inducing premature delivery was beneficial. The ATE averages the beneficial and non-beneficial effects in the entire population, to give an overall non-beneficial effect. The ETT isolates the beneficial effect among those who actually received the intervention. Thus, in this example, premature delivery actually did benefit those who received it, even though it would not benefit everybody.

There are many other estimands that can be defined, including the local average treatment effect, the survivor average causal effect, the complier average causal effect, and a host of principal strata effects. We will not discuss these in the context of this course, but it’s good to be aware of their existence.

STUDY QUESTION 2B: List some estimators that can be used to quantify the odds ratio.

The estimand is the object we want to estimate. The **estimator** is an equation that allows us to use our data to quantify the estimand. Suppose, for example, we were explicitly interested in quantifying the causal risk difference for the relation between smoking and CVD risk. To do this, we have to start by quantifying the associational risk difference, but there are many ways to do this (e.g., ordinary least squares, maximum likelihood, or the method of moments).

To be specific, let's simulate some hypothetical data on the relation between smoking and CVD. Let's look at ordinary least squares, maximum likelihood, and the generalized method of moments as estimators:

```
# define the expit function
expit<-function(z){1/(1+exp(-(z)))}
set.seed(123)
n<-1e6
confounder<-rbinom(n,1,.5)
smoking<-rbinom(n,1,expit(-2+log(2)*confounder))
CVD<-rbinom(n,1,.1+.05*smoking+.05*confounder)

# the data
head(data.frame(CVD,smoking,confounder))
```

```
##   CVD smoking confounder
## 1   0       0          0
## 2   0       0          1
## 3   1       0          0
## 4   1       0          1
## 5   0       0          1
## 6   0       0          0
```

```
round(mean(confounder),3)
```

```
## [1] 0.499
```

```
round(mean(smoking),3)
```

```
## [1] 0.166
```

```
round(mean(CVD),3)
```

```
## [1] 0.133
```

```
#OLS
round(coef(lm(CVD~smoking+confounder)),4)
```

```
## (Intercept)     smoking  confounder
##      0.1000      0.0485      0.0501
```

```
#ML1
round(coef(glm(CVD~smoking+confounder,family=poisson("identity"))),4)
```

```
## (Intercept)     smoking  confounder
##      0.0999      0.0487      0.0502
```

```
#ML2
round(coef(glm(CVD~smoking+confounder,family=binomial("identity"))),4)
```

```
## (Intercept)     smoking  confounder
##      0.1000      0.0487      0.0501
```

```
#GMM
round(gmm(CVD~smoking+confounder,x=cbind(smoking, confounder))$coefficients,4)
```

```
## (Intercept)     smoking  confounder
##      0.1000      0.0485      0.0501
```

In our simple setting with 1 million observations, ordinary least squares, maximum likelihood, and the generalized method of moments yield the same associational risk difference (as expected) even though they are different **estimators**. Finally, the values obtained from each regression approach are our **estimates**.

It is important to note that these are not causal risk differences, but are associational. To interpret them as causal effects, we have to evaluate whether we can identify the effect. We discuss this next.

Identifiability: Average Treatment Effect

In our simulation example, we estimated the associational (as opposed to causal) risk difference using four different estimators (ordinary least squares, two different maximum likelihood estimators, and the generalized method of moments). Estimating associations is all we can do with empirical data. Any time you use software to obtain a point estimate, you get an associational measure, irrespective of the method used.[10]

But our primary interest is (most often, see note 2 below) in causal quantities. In our simulated case, we want to estimate the causal risk difference for the effect of smoking on CVD. We can only do so if this causal risk difference is **identified**. *A parameter (e.g., causal risk difference) is identified if we can write it as a function of the observed data.*

The causal risk difference is defined as a contrast of potential outcomes. Referring back to our simulated example,[11] we want to estimate the causal risk difference which is an example of an average treatment effect:

$$E(Y^1 - Y^0),$$

where Y^1, Y^0 are the potential CVD outcomes that would be observed if smoking were set to 1 and 0, respectively. On the other hand, the associational risk difference is defined as a contrast of observed outcomes:

$$E(Y \mid X = 1) - E(Y \mid X = 0),$$

where each term in this equation is interpreted as the risk of CVD **among those who had** $X = x$.

Study Question: The causal risk difference is one example of an average treatment effect. Can you identify other examples?

The causal risk difference is identified if the following equation holds:[12]

$$E(Y^x) = E(Y \mid X = x)$$

which says that the risk of CVD that would be observed if everyone were set to $X = x$ is equal to the risk of CVD that we observe among those with $X =$

[10] Note that this is not just true whether we use MLE, OLS, or the method of moments, it's also true when we use IP-weighting, g computation, g estimation, some doubly robust estimator (e.g., TMLE, AIPW), or really any other estimation approach.

[11] To simplify the explanation here, I am ignoring the fact that we conditioned on (or adjusted for) confounders C. Of course, without adjusting for C, we get a confounded estimate. However, if we adjust for C, we no longer obtain the average treatment effect. Instead, we obtain the conditional treatment effect. Their are important distinctions between average and conditional treatment effects that we will unfortunately not have time to discuss.

[12] Throughout this course, we will assume that the target parameter of interest is a causal contrast of potential outcomes. Sometimes, the target parameter of interest is an associational contrast, and the assumptions needed are less demanding. See, e.g., !Naimi2016c.

x. In this equation, the right hand side equation is written entirely in terms of observed data ($Y = 1$). The left hand side is a function of unobserved potential outcomes ($Y^x = 1$). This equivalence will only hold if we can make some assumptions.

The first is **counterfactual consistency**, which states that the potential outcome that would be observed if we set the exposure to the observed value is the observed outcome [^!^Hernan2005b,^!^Hernan2008a,^!^Hernan2011a,^!^VanderWeele2013b].[13] Formally, counterfactual consistency states that:

$$\text{if } X = x \text{ then } Y^x = Y$$

The status of this assumption remains unaffected by the choice of analytic method (e.g., standard regression versus g methods). Rather, this assumption's validity depends on the nature of the exposure assignment mechanism.

We must also assume **no interference**, which states that the potential outcome for any given individual does not depend on the exposure status of another individual [^!^Hudgens2008,^!^Naimi2015]. If this assumption were not true, we would have to write the potential outcomes as a function of the exposure status of multiple individuals. For example, for two different people indexed by i and j, we might write: $Y_i^{x_i, x_j}$.[14] Notation and methods that account for interference can be somewhat complex [^!^Tchetgen2012,^!^Halloran2016], and we will not consider the impact of interference here.

Together, counterfactual consistency and no interference allow us to make some progress in writing the potential risk $E(Y^x)$ as a function of the observed risk $E(Y \mid X = x)$. Specifically, by counterfactual consistency and no interference, we can do the following:

$$\begin{aligned} E(Y^x) &= E(Y \mid X = x) && (1) \\ &= E(Y^x \mid X = x) && (2) \end{aligned}$$

A third assumption is **exchangeability**, which implies that the potential outcomes under a specific exposure (Y^x) are independent of the observed exposures X [^!^Greenland1986,^!^Greenland1999,^!^Greenland2009]. To explain the intuition behind exchangeability [^!^Hernan2015], consider a setting in which we are estimating the effect of aspirin on headache incidence in a cohort

[13] While somewhat convoluted, this assumption is about legitimizing the connection between our observational study, and future interventions in actual populations. In our observational study, we **see** people with with a certain value of the exposure. In a future intervention, we **set** people to a certain value of the exposure.

[14] Together, counterfactual consistency and no interference make up the stable-unit treatment value assumption (SUTVA), first articulated by !Rubin1980.

of individuals aged 18-40 years.[15] To do this experiment, a researcher randomly assigns 50% of the cohort to aspirin, and the remaining 50% to placebo. However, to overcome some logistical complications, before actually giving them aspirin/placebo, this researcher hands out cards that indicate whether the participant was assigned to aspirin (white card) versus placebo (black card).

After the cards/aspirin/placebo are distributed and the follow-up period transpires, the researcher tallies up the number of headaches in each exposure group. He finds the following results:

$$\begin{aligned} \text{Aspirin (White Card): } & E(Y \mid X = 1) = 0.6 \\ \text{Placebo (Black Card): } & E(Y \mid X = 0) = 0.1 \end{aligned}$$

However, after reviewing the study protocol, he realizes that he accidentally assigned placebo to those with the white card, and aspirin to those with the black card, instead of the other way around. Fortunately, this has no actual impact on the study, with the exception of needing to switch the aspirin label with the placebo label. Why? Randomization (in a sufficiently large enough sample) creates an independencies between outcome that would be observed under some exposure value (the potential outcome) and the observed exposure. In our case, $E(Y^{x=1}) = 0.1$, and this is the case whether the exposure received was placebo (0) or aspirin (1):

$$E(Y^{x=1}) = 0.1 \implies \begin{cases} E(Y^{x=1} \mid X = 1) = 0.1 \\ E(Y^{x=1} \mid X = 0) = 0.1 \end{cases}$$

Thus, because of randomization the following mathematical relation is implied:

$$E(Y^x \mid X) = E(Y^x) \tag{3}$$

which is exactly what we need to complete the identifiability statement above:

$$\begin{aligned} E(Y^x) &= E(Y \mid X = x) && (4) \\ &= E(Y^x \mid X = x) \text{ by consistency and no interference} && (5) \\ &= E(Y^x) \text{ by exchangeability} && (6) \end{aligned}$$

[15] Assume that our sample size is sufficiently large so as to avoid any sampling variability problems.

With exchangeability, we are able to drop the observed exposure on the right side of the conditioning statement. However, we motivated this exchangeability assumption via simple randomization. What about when we have an observational study where the exposure is not randomized? It turns out that the validity of results from an observational study still rests upon the idea of randomization. For example, if we conduct an analysis in observational data where we adjust for 3 confounding variables, and we believe these three variables are sufficient to control for all confounding (and there are no other threats to validity, such as selection or information bias), then we can show that the same set of steps required to equate the average potential outcomes $E(Y^x)$ with the average observed outcome among those with $X = x$: $E(Y \mid X = x)$.

Consider our aspirin and headache example above, instead rather than randomly assign 50% of the individuals to aspirin and 50% to placebo, imagine that for people who in an average week sleep < 7 hours per night, we use a coin that chooses heads 75% if the time to assign aspirin, and 25% of the time to assign placebo. And for people who sleep $\geq$ 7 hours per night, we use a 50:50 coin to assign aspirin and placebo.

Using an aspirin:placebo assignment proportion of 75:25 for “non-sleepers”, and 50:50 for “sleepers” creates an association between sleeping quantity and aspirin assignment. If sleeping quantity also has an association with headache, what we’ve done is created a confounding relation between aspirin versus placebo and headache via sleeping quantity. Because of this confounding relation, we can no longer re-write the conditional expectation $E(Y^x \mid X = x)$ as $E(Y^x)$.

However, if we adjust for sleeping quantity in our analysis, we can partly recover the procedure we need to equate these quantities:

$$\begin{aligned}
E(Y^x) &= \sum_c E(Y \mid X = x, C) && (7)\\
&= \sum_c E(Y^x \mid X = x, C) \text{ by consistency and no interference} && (8)\\
&= \sum_c E(Y^x \mid C) \text{ by conditional exchangeability} && (9)\\
&= E(Y^x) \text{ by marginalization} && (10)
\end{aligned}$$

The only difference is that now we have to incorporate an additional step in which we "average" or marginalize over the distribution of C to obtain a weighted average of the $E(Y^x)$ in the sample or population.

STUDY QUESTION: Why is the word ”exchangeable” used to describe this concept? What, precisely, is being ”exchanged”?

Although it seems that we have successfully written the potential risk as a function of the observed data, we are in need of two more assumptions. The first is **correct model specification**. This assumption is required when we rely on models to estimate effects, which is particularly relevant in an observational study when we have several confounders we have to adjust for.

Consider the example above, where we had to adjust for C to equate the potential and observed outcomes. In our simple example, we only considered one confounding variable (sleep quantity), but in a typical observational study, we'd typically adjust for quite a few variables. Consider further that we'd typically employ a statistical regression model (e.g., linear, logistic, Poisson, Cox, or other) to actually implement our adjustment, which might look something like[16]:

$$E(Y \mid X, C_1, C_2, C_3, C_4) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_3 + \beta_5 C_4$$

The problem with using the above model is that it makes fairly strong assumptions about exactly *how* Y is related to X and the confounders. Specifically, this equation states (or assumes) that the conditional mean of Y is related to all the variables additively such that a single unit increase in each variable results in a linear and independent increase in the mean of Y.

However, consider that for five variables there can be a total of

$$\binom{5}{2} = 10$$

two-way interactions that we could potentially add to the model. Alternatively, we could also include up to:

[16] Such a model would be what we'd use in SAS, Stata, R, or any other software when we use the regression function and include only main effects terms in the model

$$2^5 - 5 - 1 = 26$$

k-way interactions (including 2, 3, 4, and 5 way). If we exclude any of the relevant interactions from among this set, we could end up with a biased estimator, which could create an dependence between the observed exposure and the potential outcomes, and would thus not allow us to equate the potential and observed outcomes the way we did when we could assume exchangeability.

There are other ways in which this correct model specification assumption can be violated, including making incorrect linearity (or nonlinearity) assumptions, choosing the wrong link function in a generalized linear model [see, e.g., $^{!}$Weisberg1994], or making the wrong distributional assumption about the conditional mean of the outcome. It is for these reasons (among others) that machine learning methods are becoming so popular. They do not make such (parametric) assumptions about how the data were generated. However, they do come with some important trade-offs that we will get into in subsequent sections.

Study Question: Can you think of a relation between correct model misspecification and exchangeability?

Their is another assumption known as **positivity**,[17] and requires exposed and unexposed individuals within all confounding levels [^!^Mortimer2005,^!^Westreich2010a]. There are two kinds of positivity violations (non-positivity): structural (or deterministic) and stochastic[18] (or random).

Structural non-positivity occurs when individuals with certain covariate values cannot be exposed. For example, in occupational epidemiology work-status (employed/unemployed in workplace under study) is a confounder, but individuals who leave the workplace can no longer be exposed to a work-based exposure. Alternatively, stochastic non-positivity arises when the sample size is not large enough to populate all confounder strata with observations.

Problems because of positivity arise for two reasons. The first is definitional. Consider the step in our equation above where we marginalize over C to equate the potential and observed outcomes. In the case where C is binary

[17] Also known as the experimental treatment assignment assumption.

[18] The word **stochastic** is derived from the greek word "to aim," as in "to aim for a target."

and we want to estimate the potential outcome if everyone were exposed to $X = 1$, this step could be re-written as:

$$E(Y^{x=1}) = E(Y \mid X = 1, C = 1)P(C = 1)+E(Y \mid X = 1, C = 0)P(C = 0)$$

Now imagine that for those with $C = 1$, it is either impossible to have $X = 1$ (structural nonpositivity) or we just don't have anyone in our sample with $X = 1$ (stochastic nonpositivity). Mathematically, it does not make sense to write $E(Y \mid X = 1, C = 1)$ because there are no individuals with $X = 1$ and $C = 1$. We thus cannot define this conditional average.

The second problem with positivity violations has to do with estimators. Consider, for example, the simple inverse probability weight that corresponds to the above scenario (i.e., if $C = 1$, there are no individuals with $X = 1$):

$$\frac{1}{P(X = 1 \mid C = 1)}$$

In this case, the probability in the denominator is zero. And because $1/0$ is undefined, we can't use IP-weighting to estimate the effect we're after with this estimator. The same type of problem arises even if there are only a very small number people in the sample with $X = 1$ if $C = 1$. In this latter case, imagine that the probability of being exposed is very small, say 0.0001. Then, the above weight would be equivalent to $1/0.0001 = 10,000$. The above weight means that one or more of these individuals will contribute $10,000$ observations to the weighted analysis (usually well more than the original sample). These types of problems result in instability of the estimator (because the results end up being heavily dependent on only a few individuals in the sample with large weights).

When faced with positivity violations, one should either re-define the estimand so that there is no positivity violation, choose an estimator that is less affected by positivity problems, or both [^!^Petersen2012].[19] Alternative estimands include the effect of treatment on the treated or untreated, various types of stochastic effects [including incremental propensity score effects [^!^Kennedy2019], which do not require that positivity hold [^!^Naimi2021]], or "blip" effects that are encoded in structural nested models, and can be estimated with g estimation. One can also use, collaborative targeted minimum

[19] Keep in mind: one cannot simply "avoid" positivity. In an extreme setting, nonpositivity means that those who were exposed in the sample are very unlikely to be exposed (and vice versa). In such a situation, it may not make sense to estimate the average treatment effect, because there is a subset of the population who may never realistically be exposed (or unexposed). In this case, g estimation, cTMLE, and the parametric g formula can actually estimate parameters that differ slightly or profoundly from the ATE.

loss-based estimation,[20] and the parametric g formula, which tend to be less sensitive to positivity violations [^!^Cole2013;^!^Porter2011;^!^Ju2017].

There are a number of different procedures one can use to evaluate whether posititivity is a problem. Among these include propensity score overlap plots. Consider again our simple simulation example:

```
# define the expit function
expit<-function(z){1/(1+exp(-(z)))}
set.seed(123)
n<-1e3
confounder<-rbinom(n,1,.5)
smoking<-rbinom(n,1,expit(-2+log(2)*confounder))
```

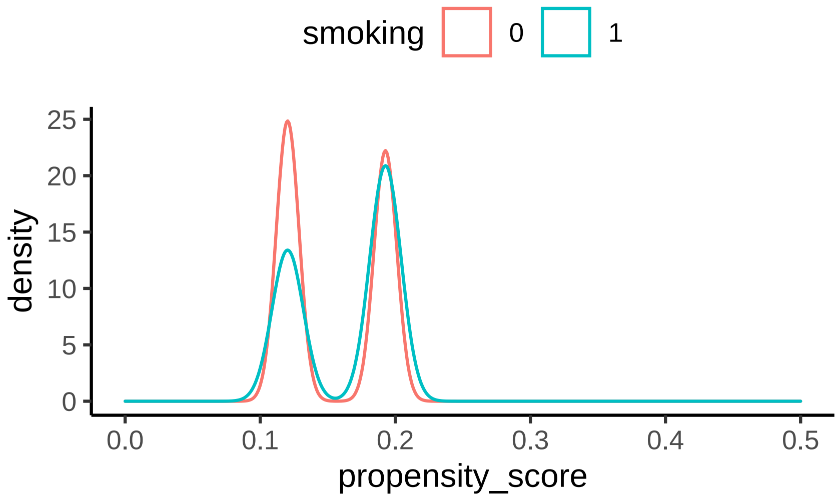
To get the propensity score for a binary exposure, we can fit a logistic model to the exposure data, conditional on all confounders:

```
propensity_score <- glm(smoking ~ confounder,family=binomial(link="logit"))$fitted.values

## by appending a "$fitted.values" to the end
## of this glm function, we are
## keeping the predicted values from the model
## under the observed data settings.
```

We can now plot the density of this propensity score for each exposure group to see how they overlap:

```
plot_data <- data.frame(propensity_score,smoking=as.factor(smoking))

library(ggplot2)
ggplot(data=plot_data) + geom_density(aes(x=propensity_score,group=smoking,color=smoking)) + xlim(0,.5)
```

[20] there is mounting evidence that standard (not collaborative) TMLE is highly sensitive to positivity violations.



Since the mass of the density for the exposed occurs in the same place as the density mass for the unexposed, positivity does not seem to be much of an issue here. Another way to check positivity is to create stabilized inverse probability weights[21] and look at their descriptive statistics.

```
sw <- (mean(smoking)/propensity_score)*smoking +
  ((1 - mean(smoking))/(1 - propensity_score))*(1 - smoking)

summary(sw)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.8096  0.9594  0.9594  1.0000  1.0455  1.2966
```

The mean of the stabilized weights is 1, and the max weight is not large at all, suggesting very well-behaved weights. Thus, in this particular case, we are not concerned with violations of the positivity assumption.

Non-Identifiability: Bounding Effects

What happens when the effect we want to estimate is not identifiable? Suppose, for example, exchangeability is violated because we could not randomize our exposure? Or perhaps there was some loss to follow-up that could not be accounted for with absolute certainty? More likely there is both unmeasured confounding and loss to follow-up. When this happens, we get a point estimate for the causal effect of interest, but it could either be smaller or larger in magnitude due to the influence of the unmeasured confounder and loss to follow-up.

[21] We won't get too deep into the theory for / definition of weights here. But here is the code for creating stabilized weights and evaluating positivity.

In order to get a precise measure of **all the values the point estimate can possibly take** as a result of unmeasured confounding and loss to follow-up, we can estimate bounds for the point estimate of interest. Confidence intervals are bounds on the point estimate of interest that capture the uncertainty that results from random variation [^!^Wasserman2004]. In contrast, identification bounds capture the uncertainty that results from potential violations of certain assumptions required for identification [^!^Manski2003].

Consider a study by [!]Cole2019 in which they sought to quantify the effect of injection drug use on time to AIDS or death in a cohort of 1164 adult HIV-positive, AIDS-free women. These women were followed for AIDS or death up to 10 years from 12/6/95 in the Women's Interagency HIV Study [^!^Barkan1998]. Overall, 127 of 1164 women (11%) were lost to follow up. Adjusted risk differences were obtained via inverse probability weighting. Adjustment was made for age, race and nadir CD4 cell count.

Figure 4 shows the results from the analysis (obtained via personal communication with Stephen R. Cole). The top left panel shows the unadjusted risk difference over follow-up. The top right panel shows the corresponding risk difference after adjusting for loss to follow-up and measured confounders. The bottom left panel shows the identification bounds that result from loss to follow-up. And the bottom right panel shows the identification bounds that result from both loss to follow-up and unmeasured confounding. Specifically, the black area shows all possible risk differences that could arise given the data.

The bottom right panel in Figure 4 tells us something critically important that we often fail to consider when conducting an empirical study. Without assumptions, data alone rarely provide much information about a causal effect of interest. Rather, when we interpret that a point estimate from a statistical model as a causal effect estimate, we are invoking a whole set of assumptions (knowingly or unknowingly) that allow us to get a single number out of our data, rather than a range of possible values. One of these sets of assumptions we discussed here (counterfactual consistency, no interference, positivity, exchangeability, correct model specification). Nonparametric bounds such as those depicted in the study by [!]Cole2019 help us understand exactly how much support our data provide for an effect of interest, and how much of our results rely on unverifiable assumptions.

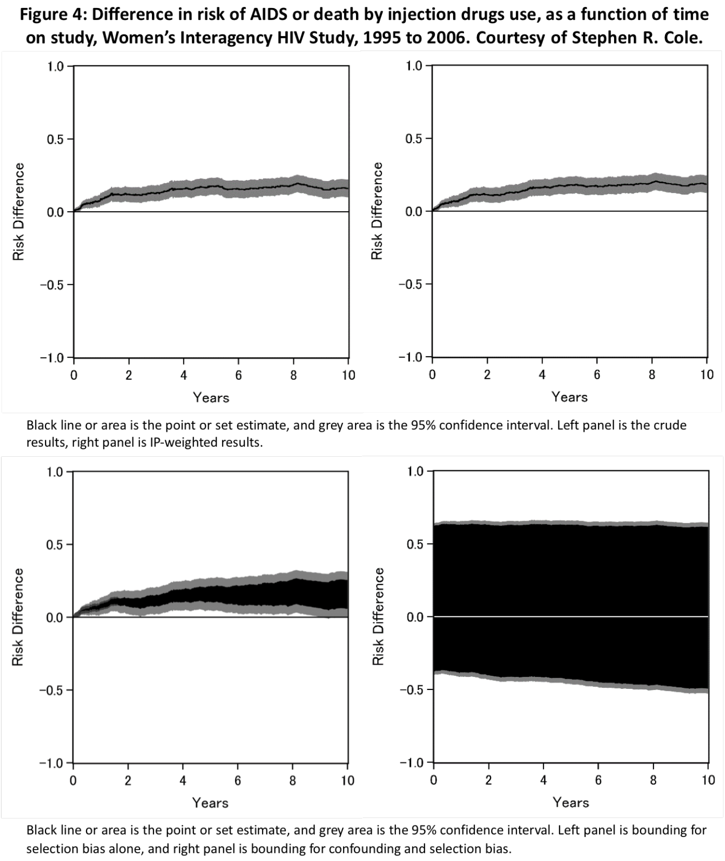


Figure 4: Difference in risk of AIDS or death by injection drugs use, as a function of time on study, Women's Interagency HIV Study, 1995 to 2006. Courtesy of Stephen R. Cole.

Black line or area is the point or set estimate, and grey area is the 95% confidence interval. Left panel is the crude results, right panel is IP-weighted results.

Black line or area is the point or set estimate, and grey area is the 95% confidence interval. Left panel is bounding for selection bias alone, and right panel is bounding for confounding and selection bias.

Figure 1: Bounds figure