Causal Inference

Ashley I Naimi

January 2024

Contents

1 Introduction

Bias is a foundation concept in data science, but the term is unfortunately used to represent a number of very distinct concepts. For example, as mentioned, the term bias is often used in epidemiology to convey the concept of an inconsistent estimator in statistics. In machine learning, the term "bias" is often used to refer to an intercept in a linear regression model, the parameter that shifts predicted values in a prediction algorithm, or the parameter in a deep neural network that shifts the magnitude of the activation function [Bishop (2016); p 138].

In this section, we'll clarify two notions of bias: identification bias, and estimation bias. We'll start with identification bias, which is related to the consistency of an estimator, and depends on key unverifiable assumptions. We'll show what conditions are needed to establish identifiability for the average treatment effect.

Next, we'll discuss estimation bias, which depends on the properties of the estimator being used to analyze the data.

2 Identification Bias

In our simulation example, we estimated the associational (as opposed to causal) contrast using four different estimators (maximum likelihood, penalized maximum likelihood, generalized method of moments, and AIPW). 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. This is true with ANY estimator, including IP-weighting, g computation, g estimation, or double robust approaches, such as AIPW (as demonstrated) or targeted maximum likelihood estimation.

But our primary interest is often in causal quantities. In our simulated case, we want to estimate the causal mean difference for the effect of smoking on CVD. We can only do so if this causal risk difference is **identified**. Formally, *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,[1] we want to estimate the causal risk

[1] 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 discuss in a subsequent section.

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** $X = x$.

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

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

This equation 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 $X = x$. 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 ($Y^x = 1$). Because potential outcomes are unobservable abstractions, this equivalence will only hold if we can make some assumptions.

2.1 Counterfactual Consistency

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 (Hernán, 2005, Hernan and Taubman (2008),Hernán and VanderWeele (2011),VanderWeele and Hernán (2013)).[2] 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.

One way to grasp what counterfactual consistency is about is to use the

[2] While somewhat convoluted, this assumption is primarily about legitimizing the connection between our observational study, and future interventions in actual populations based on this study. 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. The differences between seeing and setting can be profound.

example of the "effect" of obesity on mortality (Hernan and Taubman, 2008). We know that obesity is associated with an increased risk of mortality, but interpreting this excess risk into a causal statement is tricky. In an observational study, the association between obesity and mortality is obtained by contrasting the risk of mortality among, say, obese versus non-obese individuals. However, causally acting on this information would require us to find a way to make obese individuals non-obese. This might consist of getting obese individuals to diet, exercise, start smoking, or to undergo a single leg amputation (!). Each of these interventions could reduce BMI, and thus getting obese individuals to become non-obese. However, each intervention will likely have (dramatically) different effects on mortality.

They key here is that obesity is not a manipulable construct (on the other hand, dieting, exercise, smoking, and leg amuptation, more or less, are). As a result, precisely translating what we mean by "the effect of obesity" is difficult. The same problem arises with other variables, such as the "effect of education," the "effect of race/ethnicity," and the "effect of socieoeconomic status," to name a few (Naimi and Kaufman, 2015).

2.2 Interference

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 (Hudgens and Halloran, 2008, Naimi and Kaufman (2015)). 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: $Y_i^{x_i,x_j}$.[3] Notation and methods that account for interference can become very complex very quickly (Tchetgen Tchetgen and VanderWeele, 2012, Halloran and Hudgens (2016), Hudgens and Halloran (2008)). As a result, we will not consider the impact of interference here, except only to say that different estimands and estimators should be used to properly account for them.

[3] Together, counterfactual consistency and no interference make up the stable-unit treatment value assumption (SUTVA), first articulated by Rubin (1980).

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:

$$E(Y^x) = E(Y \mid X = x) \tag{1}$$
$$= E(Y^x \mid X = x) \tag{2}$$

2.3 Exchangeability

A third assumption is **exchangeability**, which implies that the potential outcomes under a specific exposure (Y^x) are independent of the observed exposures X (Greenland and Robins, 1986, Greenland et al. (1999),Greenland and Robins (2009)). To explain the intuition behind exchangeability (Hernán and Robins, 2020), consider a setting in which we are estimating the effect of aspirin on headache incidence in a cohort of individuals aged 18-40 years.[4] 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 (red card) versus placebo (blue card).

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

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:

$$\text{Aspirin (Red Card): } E(Y \mid X = 1) = 0.6$$
$$\text{Placebo (Blue Card): } E(Y \mid X = 0) = 0.1$$

However, after reviewing the study protocol, he realizes that he accidentally assigned placebo to those with the red card, and aspirin to those with the blue 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 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 ($X = 0$) or aspirin ($X = 1$):

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

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 progress the identifiability statement above:

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

2.4 Conditional Exchangeability

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:

$$E(Y^x) = \sum_c E(Y \mid X = x, C) \tag{7}$$

$$= \sum_c E(Y^x \mid X = x, C) \text{ by consistency and no interference} \tag{8}$$

$$= \sum_c E(Y^x \mid C) \text{ by conditional exchangeability} \tag{9}$$

$$= E(Y^x) \text{ by marginalization} \tag{10}$$

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.

2.5 Positivity

Although it seems that we have successfully written the potential risk as a function of the observed data, we are in need of one more assumption, known as **positivity**.[5] Positivity requires exposed and unexposed individuals within all confounding levels (Mortimer et al., 2005, Westreich and Cole (2010)). There are two kinds of positivity violations (non-positivity): structural (or deterministic) and stochastic[6] (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.

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

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

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 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, a 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 (Petersen et al., 2012).[7] Alternative estimands include the effect of treatment on the treated or untreated, various types of stochastic effects [including incremental propensity score effects

[7] Keep in mind: one cannot simply "avoid" positivity. In extreme setting, nonpositivity means that those who were unexposed 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.

(Kennedy, 2019), which do not require that positivity hold (Naimi et al., 2021)], or "blip" effects that are encoded in structural nested models, and can be estimated with g estimation. One can also use collaborative targeted minimum loss-based estimation,[8] and the parametric g formula, which tend to be less sensitive to positivity violations (Cole et al., 2013; Porter et al., 2011; Ju et al., 2017).

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

There are a number of different procedures one can use to evaluate whether positivity is a problem. Among these include propensity score overlap plots. Consider again our data from the last section. To get the propensity score for a binary exposure, we can fit a logistic model to the exposure data, conditional on confounders. Here, we use the Lalonde dataset, which is well known in econometric circles. This dataset was originally obtained from a study used to evaluate the effect of a training program (treat) on income:

```
library(MatchIt)
data("lalonde")

head(lalonde)
```

```
##      treat age educ   race married nodegree re74 re75       re78
## NSW1     1  37   11  black       1        1    0    0  9930.0460
## NSW2     1  22    9 hispan       0        1    0    0  3595.8940
## NSW3     1  30   12  black       0        0    0    0 24909.4500
## NSW4     1  27   11  black       0        1    0    0  7506.1460
## NSW5     1  33    8  black       0        1    0    0   289.7899
## NSW6     1  22    9  black       0        1    0    0  4056.4940
```

```
propensity_score <- glm(treat ~ age + educ +
    re74 + re75, data = lalonde, family = binomial(link = "logit"))$fitted.values

head(propensity_score)
```

```
##      NSW1      NSW2      NSW3      NSW4      NSW5      NSW6
## 0.3916080 0.3824366 0.4084567 0.3999990 0.3627466 0.3824366
```

```
## 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:

```
set.seed(123)

exposure <- lalonde$treat

plot_data <- data.frame(propensity_score,
    Exposure = as.factor(lalonde$treat))

p1 <- ggplot(data = plot_data) + scale_y_continuous(expand = c(0,
    0)) + scale_x_continuous(expand = c(0,
    0)) + ylab("Density") + xlab("Propensity Score") +
    scale_color_manual(values = c("#000000",
        "deepskyblue")) + scale_fill_manual(values = c("#000000",
    "deepskyblue")) + geom_density(aes(x = propensity_score,
    group = Exposure, color = Exposure),
    bw = 0.03) + geom_histogram(aes(y = ..density..,
    x = propensity_score, group = Exposure,
    fill = Exposure), alpha = 0.25, ) + xlim(0,
    1)

ggsave(here("_images", "2022_01_10-ps_overlap.pdf"),
    plot = p1)
```

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[9] and look at their descriptive statistics.

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

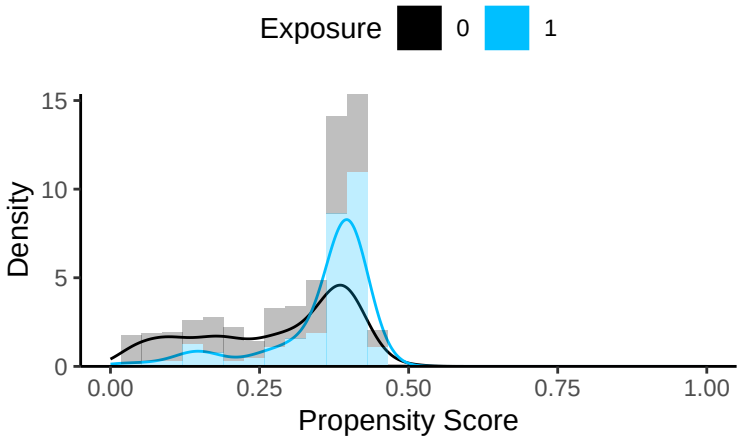

Figure 1: Propensity score overlap plot for the training intervention in 614 individuals in the Lalonde dataset.

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

summary(sw)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6548  0.7788  0.9340  1.0451  1.1245 27.6692
```

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.

Technical Note:

In a large body of methods literature, particularly econometrics, you are likely to encounter these causal assumptions articulated in different ways. Most commonly, researchers will often invoke **ignorability** as a core assumption in causal inference. There are at least two versions of ignorability: strong and weak. **Strong ignorability** is defined as the combination of the conditional independence assumption, and the positivity assumption. Technically, strong ignorability holds if, for individual i with a binary exposure $X \in \{0, 1\}$:

$$(Y_i^{x=0}, Y_i^{x=1}) \perp\!\!\!\perp X_i \mid \mathbf{C}_i, \text{ and}$$
$$0 < P(X_i = 1 \mid \mathbf{C}_i) < 1,$$

where the $\perp\!\!\!\perp$ symbol denotes independence (in this case, conditional independence since we include $\mid \mathbf{C}_i$). In this case, the ignorability is "strong" because the independence is assumed to exist *jointly* between both potential outcomes (Y_i^{x=0}, Y_i^{x=1}) for individual i, and the exposure X_i, conditional on $\mathbf{C}_i$. Sometimes, a weaker version of the assumption is made:

$$(Y_i^{x}) \perp\!\!\!\perp X_i \mid \mathbf{C}_i, \text{ and}$$
$$0 < P(X_i = 1 \mid \mathbf{C}_i) < 1,$$

This version of the assumption is weaker in that we need not worry about whether the potential outcomes are jointly independent of the observed exposure. Rather, we only need each potential outcome (Y_i^x) to be independent of the observed exposure.

Even still, these assumptions are stronger than what we need to identify the causal risk difference, risk ratio, odds ratio, or other typical summary contrasts we often quantify in epidemiology. In the above proof, we demonstrated that identifiability is obtained under **mean exchangeability** $E(Y^x \mid X, \mathbf{C}) = E(Y^x \mid \mathbf{C})$. This assumption is even weaker than those articulated in the formalization of strong and weak ignorability.

The distinctions between strong ignorabilty, weak ignorability, and mean exchangeability are of little practical consequence. While it is important to recognize that ignorability typically consists of the combination of exchangeability and positivity. Additional details on and intuition behind the different versions of exchangeability can be found in Technical Point 2.1 of Hernán and Robins (2020).

3 Estimation Bias

The second form of bias that we will discuss is referred to as "estimation bias", which is distinct from identification bias (Díaz, 2020). While identification bias involves establishing that the data can be used to quantify the target estimand, estimation bias involves choosing the estimator that best uses the data to quantify the target estimand.

To explain the concept of estimation bias, we start with an explanation of why the change in estimate approach can yield erroneous conclusions about a cause-effect relation of interest.

3.1 Change in Estimate Approach

The "change in estimate" approach to confounder selection use to be a popular appraoch to determining whether a variable needs to be adjusted for in an analysis as a confounder. The basic procedure can be implemented as follows. For example, suppose you have an exposure X, an outcome Y, and five variables that you suspect are confounders: $C_1, C_2, \ldots C_5$. You then fit a regression model regressing the outcome against the exposure:

$$E(Y \mid X) = \alpha_0 + \alpha_1 X$$

You then fit a separate regression model in which you also adjust for C_1:

$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 C_1$$

The change in estimate approach then evaluates the following (Talbot et al., 2021)[10]:

$$\xi = \frac{(\hat{\alpha}_1 - \hat{\beta}_1)}{\hat{\alpha}_1}$$

The variable C_1 would then be deemed a "confounder" if the computed value ξ exceeds some threshold, usually 10% (VanderWeele, 2019).

The problems with this approach are well documented (Talbot et al., 2021, VanderWeele (2019)). In particular, the change-in-estimate approach cannot distinguish between a confounder and a collider, and will fail in the presence of more complex scenarios, such as M-bias (Shrier and Platt, 2008). Ultimately,

[10] Note there are many different ways to implement the change in estimate approach, including via forward, backward, or stepwise selection. All of them are wrong.

choosing confounders cannot be done without a causal model encoding knowledge that is not present in the data alone (Hernán, 2019, Díaz (2020)).

3.2 Bias Variance Tradeoff

There is another way to approach variable selection, using bias variance-tradeoff arguments. In this setting, consider that we've used the best subject matter possible to select a set of confounders that are deemed sufficient for confounding adjustment. Again, let's say these variables are denoted $C_1, C_2, \ldots C_5$. This would entail fitting a model such as:

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

Suppose we fit this model using our data, and obtain an estimate for $\beta_1 = 1.5$, with $SE(\hat{\beta}_1) = 1.25$. However, if we remove C_4 from the model, we get an estimate of $\beta_1 = 1.46$, with $SE(\hat{\beta}_1) = 0.66$. These results, including 95% confidence intervals, are included in the following Table:

Model	$\hat{\beta}$	$SE(\hat{\beta})$	LCL	UCL
Including C_4	1.50	1.25	-0.95	3.95
Excluding C_4	1.46	0.66	0.17	2.75

In this particular case, it looks like removing a confounder from the model leads to a change in the point estimate, from 1.5 to 1.46. This is a change of 2.7%. In contrast, the change in the standard error we obtain is much larger, from 1.25 to 0.66 (a near 50% change). Thus, while we are inducing a suspected "bias" in our point estimate of 2.7%, we are reducing the variance by nearly 50%.

The natural question is then, is it worth including the C_4 confounder?

From a bias-variance tradeoff perspective, the answer is no, even though subject matter knowledge deems it a confounder that should be included in the adjustment set. Including C_4 could end up inducing an estimation bias that we should try to avoid.

3.3 Functional Form Specification

In a typical observational study, researchers would have to adjust for quite a few potential confounding variables. This makes the practicality of conducting the bias-variance tradeoff evaluation above a little tedious. But the task becomes more challenging when we consider other issues. Consider again with our same linear regression model. This linear model is quite common in empirical research settings:

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

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[11]

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10$$

two-way interactions that we could potentially add to the model. Additionally, we could include higher-order interactions, for example, a three-way interaction between X, C_1, and C_3. In fact, if we considered higher order interactions, for this simple model would could have up to:

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

k-way interactions (including 2, 3, 4, and 5 way). If we exclude any of the relevant interactions from among this set, our model would be misspecified. This misspecification could result in bias.

Other functional form issues can be considered. For example, suppose that a subject matter expert believes that the confounder C_3 and C_4 are related to the outcome in a very specific way:

$$f(C_3, C_4) = \frac{C_3}{C_3 C_4 + \exp C_4}$$

This would be very useful information, and we would want to include it in

[11] This equation is referred to as the binomial coefficient.

our regression model, such as:

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

This may be the case, but then we can also ask about other transformations for other variables. The total number of possible variations we could introduce here is exceptionally large.

There are also many other kinds of choices that can lead to bias with the estimator, including making linearity (or nonlinearity) assumptions, choosing the link functions in a generalized linear model (see, e.g., Weisberg and Welsh, 1994), or making the distributional assumption about the conditional mean of the outcome. It is for these reasons (among others) that machine learning methods are becoming so popular. Generally, machine learning methods do not tend to rely as heavily on such (parametric) assumptions about how the data were generated. However, they do come with some important trade-offs that should be considered before use. We will consider these tradeoffs shortly.

References

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2016.

Stephen R. Cole, David B. Richardson, Haitao Chu, and Ashley I. Naimi. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. *Am J Epidemiol*, 177(9):989–996, 2013.

Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2020.

Sander Greenland and James Robins. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov*, 6(1):4, 2009.

Sander Greenland and JM Robins. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*, 15(3):413–419, 1986.

Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Stat Sci*, 14(1):29–46, 1999.

M Elizabeth Halloran and Michael G Hudgens. Dependent happenings: A recent methodological review. *Curr Epidemiol Rep*, 3(4):297–305, Dec 2016.

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

M A Hernan and S L Taubman. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *Int J Obes*, 32(S3): S8–S14, 2008.

MA Hernán. Spherical cows in a vacuum: Data analysis competitions for causal inference. *Statistical Science*, 34(1):69–71, 2019.

Miguel A. Hernán. Invited commentary: Hypothetical interventions to define causal effects—afterthought or prerequisite? *Am J Epidemiol*, 162(7):618–620, 2005.

Miguel A Hernán and Tyler J VanderWeele. Compound treatments and transportability of causal inference. *Epidemiol*, 22(3):368–377, May 2011. DOI: 10.1097/EDE.0b013e3182109296.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *J Am Stat Assoc*, 103(482):832–842, 2008.

Cheng Ju, Susan Gruber, Samuel D Lendle, Antoine Chambaz, Jessica M Franklin, Richard Wyss, Sebastian Schneeweiss, and Mark J van der Laan. Scalable collaborative targeted learning for high-dimensional data. *Statistical Methods in Medical Research*, 28(2):532–554, 2017.

Edward H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Kathleen M Mortimer, Romain Neugebauer, Mark van der Laan, and Ira B Tager. An application of model-fitting procedures for marginal structural models. *Am J Epidemiol*, 162(4):382–388, Aug 2005. DOI: 10.1093/aje/kwi208.

Ashley I. Naimi and Jay S. Kaufman. Counterfactual theory in social epidemiology: Reconciling analysis and action for the social determinants of health. *Curr Epidemiol Reports*, 2(1):52–60, 2015.

Ashley I. Naimi, E Rudolph, H Kennedy, A Cartus, SI Kirkpatrick, DM Haas, H Simhan, and LM Bodnar. Incremental propensity score effects for time-fixed exposures. *Epidemiology*, 32(2):202–208, 2021.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Stat Methods in Med Res*, 21(1):31–54, 2012.

Kristin E Porter, Susan Gruber, Mark J van der Laan, and Jasjeet S Sekhon. The relative performance of targeted maximum likelihood estimators. *Int J Biostat*, 7(1), 2011.

Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *J Am Stat Assoc*, 75(371):591–593, 1980.

Ian Shrier and Robert W. Platt. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, 8(1):70, 2008. ISSN 1471-2288. DOI: 10.1186/1471-2288-8-70. URL http://dx.doi.org/10.1186/1471-2288-8-70.

Denis Talbot, Awa Diop, Mathilde Lavigne-Robichaud, and Chantal Brisson. The change in estimate method for selecting confounders: A simulation study. *Stat Methods Med Res*, 30(9):2032–2044, 2021.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Stat Methods in Med Res*, 21(1):55–75, 2012.

Tyler J. VanderWeele. Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219, 2019.

Tyler J VanderWeele and Miguel Ángel Hernán. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013.

S. Weisberg and A. H. Welsh. Adapting for the missing link. *The Annals of Statistics*, 22(4):1674—1700, 1994.

Daniel Westreich and Stephen R. Cole. Invited commentary: Positivity in practice. *Am J Epidemiol*, 171(6):674–677, 2010.