American Journal of Epidemiology

Vol. 192, No. 5
https://doi.org/10.1093/aje/kwad012
Advance Access publication:
January 18, 2023

The *AJE* Classroom

Defining and Identifying Average Treatment Effects

Ashley I. Naimi* **and Brian W. Whitcomb**

* Correspondence to Dr. Ashley I. Naimi, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road CNR 3023, Atlanta, GA 30322 (e-mail: ashley.naimi@emory.edu).

Initially submitted October 31, 2022; accepted for publication January 10, 2023.

Methods for causal inference have experienced tremendous recent growth; this paper introduces key concepts underlying causal effect methods in epidemiologic research. "Causal effects" can be described using counterfactuals, which consider alternate versions of reality, and have a long history of use for interpreting cause-effect relationships in terms of what would have happened had things (e.g., exposures) been one way vs. another. Despite their intuitive appeal for thinking about causality, use of counterfactuals in formal causal inference requires attention to potential limitations. For example, consider a study of the effect of smoking on the 5-year cumulative risk of cardiovascular disease (CVD). Because of ambiguity in the "effect of smoking," the corresponding counterfactual contrast is unclear and could correspond to a comparison between risk that would have happened if all study participants had ever smoked vs. never smoked, smoked 3 vs. 2 cigarettes per day, continued to smoke vs. quitting, and so on (1). There is an important degree of ambiguity in the language we use informally to ask causal questions. Causal inference is largely devoted to clarifying this ambiguity through use of potential outcomes as a formal framework, and a set of assumptions to allow interpretation of associations causally.

POTENTIAL OUTCOMES

Common epidemiologic associations (risk difference, risk ratio, and odds ratio) are based on risks estimated in observed data. For a binary exposure (X) and outcome (Y), we might use a risk difference to compare risk in the exposed, $E(Y \mid X = 1)$, with that in the unexposed, $E(Y \mid X = 0)$. In contrast, the building blocks for causal inference are potential outcomes (2). Potential outcomes are the hypothetical outcomes that would occur under different exposure realities, and are thus functions of possible exposures, and not the observed exposure. We write the potential outcomes as Y^x for a given exposure x. This is interpreted as the outcome, Y, that would be observed if X were equal to x (e.g., smoke exactly 1 pack of cigarettes per day). For a binary $X \in (0, 1)$, Y^x is the outcome that would be observed if $X = 0$ or $X = 1$.

Potential outcomes are used to define causal parameters, or estimands. The average treatment effect (ATE) is a commonly used causal estimand. On the difference scale, this effect can be defined as $E(Y^{x=1} - Y^{x=0})$. Here $Y^{x=1}$ and $Y^{x=0}$ are the potential outcomes that would be observed if all individuals' exposure was set to 1 and 0, respectively. This ATE can thus be interpreted as the difference in risks that would be observed if everyone in the population were exposed, versus if everyone in the population were unexposed, which is equivalent to the average of all individual risk differences under $x = 1$ and $x = 0$.

Unfortunately, one can never observe an individual simultaneously in both exposed and unexposed states, with all other things being equal. This "fundamental problem of causal inference" implies that one can never use data to directly compute these ATEs. Instead, we must seek to interpret associational differences such as:

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

where each term in this equation can be interpreted as the risk of cardiovascular disease (CVD) among those who had $X = 1$ *or* $X = 0$. Formally, the causal risk difference is identified if the following equation holds:

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

This equation says that the risk of CVD that would be observed if everyone were set to $X = x$ (i.e., $E(Y^x)$, a function of potential outcomes) is equal to the risk of CVD that we observe among those with $X = x$ (i.e., $E(Y \mid X = x)$, a function of observed data). Because potential outcomes are unobservable, this equivalence will only hold if we can make some assumptions.

Varying sets of assumptions (3) can be used to identify target estimands. Most commonly, the assumptions for identifying ATEs are counterfactual consistency, no interference,

exchangeability, and positivity. If these assumptions hold, then we can use our data to quantify causal effects. This, in effect, is what we mean by "identifiability."

COUNTERFACTUAL CONSISTENCY

For the causal risk difference $E(Y^{x=1} - Y^{x=0})$ to represent a meaningful quantity, we need unambiguous definitions of the exposure in terms of risk. Smoking status can vary in terms of duration, amount, and timing. Thus, in quantifying the 5-year risk of CVD comparing "ever" versus "never" smokers, the potential outcome is not well-defined because smoking 1 cigarette 20 years ago is very different from smoking 2 packs a day for 20 years. Counterfactual consistency assumes that the potential outcome that would be observed if we set the exposure to the observed value is the observed outcome (4). Formally, counterfactual consistency is:

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

This convoluted statement is necessary to justify interventions that may be enacted based on observed data. For example, the outcome in an observational data set from someone classified as an "ever smoker" may not be close enough to the outcome we would observe if we set someone to become an "ever smoker."

INTERFERENCE

The "no interference" assumption states that the potential outcome for any given individual does not depend on the exposure status of another individual (5). If interference is present in a given setting, the potential outcomes have to be written as a function of the exposure status of multiple individuals. For example, for 2 different people indexed by i and j, we might write: $Y_i^{x_i,x_j}$. Interference would be violated if a participant in our study was a nonsmoker but lived in the same residence as a heavy smoker, and was thus exposed to heavy second-hand smoke. The no-interference assumption can be written as:

$$Y_i^{x_i,x_j} = Y_i^{x_i}.$$

Together, counterfactual consistency and no interference make up the stable-unit treatment value assumption (SUTVA), first articulated by Rubin (6).

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:

$$\begin{aligned} E\left(Y^x\right) &= E\left(Y \mid X = x\right) \\ &= E\left(Y^x \mid X = x\right). \end{aligned}$$

While this revised equation $[E(Y^x \mid X = x)]$ is almost what we are trying to identify $[E(Y^x)]$, we must next find a way to remove the exposure from the conditioning statement. We do this via exchangeability, or conditional exchangeability.

EXCHANGEABILITY AND CONDITIONAL EXCHANGEABILITY

Consider an ideal randomized trial with 10 individuals, where individuals with identifers (ID) = 1, 2, 3, 4, and 5 receive treatment and those with ID = 6, 7, 8, 9, and 10 receive placebo. Suppose further that in this trial, the risk of the outcome in the treated individuals (with ID = 1 to 5) is 60%, and the risk of the outcome in placebo individuals (with ID = 6 to 10) is 40%, leading to a risk difference of 20%.

Importantly, randomization implies that if the individuals with ID = 1, 2, 3, 4, and 5 had been assigned to placebo, the risk of the outcome in this group would have been identical to placebo group risk (ID = 6 to 10 with risk = 40%). In other words, under randomization, the risk of the outcome in the placebo group ($X = 0$) is exchangeable with the risk of the outcome in the treatment group ($X = 1$) had the treatment group been assigned to placebo:

$$\begin{aligned} E\left(Y^{x=0}\right) = 0.4 \Rightarrow & E\left(Y^{x=0} \mid X = 1\right) = 0.4 \\ & E\left(Y^{x=0} \mid X = 0\right) = 0.4. \end{aligned}$$

In effect, exchangeability describes the circumstance where the potential outcomes under a specific exposure (Y^x) are independent of the observed exposures X (7).

It is widely understood that adequately large sample size is necessary for successful randomization. However, we use this simple example to illustrate that exchangeability enables us to assume that the observed risk in one group (e.g., placebo) can be used as a proxy for the unobserved risk in the other group (e.g., counterfactual outcome risk among the treatment group had they been assigned to placebo) (8).

In nonexperimental settings, exchangeability may be achievable conditional on adjusting for a set of (e.g.) confounding variables C. For example, if we assume that exchangeability holds conditional on a binary confounder C, then:

$$\begin{aligned} E\left(Y^x\right) &= E\left(Y \mid X = x\right) \\ &= E\left(Y^x \mid X = x\right) \quad \text{by consistency and no interference} \\ &= \sum\nolimits_c E\left(Y^x \mid C\right) \quad \text{by conditional exchangeability} \end{aligned}$$

showing how these assumptions can be used to equate potential and observed outcomes in observational settings.

POSITIVITY

Because we've adjusted for a confounder in the above equation, we require one more assumption. Positivity requires exposed and unexposed individuals within all confounder levels (9). Problems because of positivity arise for 2 reasons. The first is definitional. Consider the final step in our equation above where we marginalize over C. This step could be

rewritten as:

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

If positivity is violated, then for those with (for example) $C = 1$, there are no individuals with $X = 1$. As a result, it does not make sense to write $E\left(Y \mid X = 1, C = 1\right)$ because there are no individuals with $X = 1$ and $C = 1$. The conditional average of Y cannot thus be defined in this group. A second problem with positivity violations has to do with estimators. For example, a simple inverse-probability weighting estimator that requires taking the reciprocal of $P\left(X = 1 \mid C = 1\right)$ is not computable if this reciprocal is equal to $1/0$, which is undefined.

CONCLUSION

As described here, the potential outcomes framework and key assumptions that can be used to make causal inferences about the ATE can be drawn from associational measures. In cases when positivity and/or exchangeability assumptions may be violated, strategies include bias or sensitivity analysis, parametric or nonparametric bounding of the treatment effect, employing estimators that rely on alternative assumptions (e.g., instrumental variables), and/or modifying the target parameter.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, United States (Ashley I. Naimi); and Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, Massachusetts, United States (Brian W. Whitcomb).

This work was supported by the National Institutes of Health (grant R01HD098130, A.I.N.; and R21ES029686 and R01ES028298, B.W.W.).

Conflict of interest: none declared.

REFERENCES

1. Robins JM. Association, causation, and marginal structural models. *Synthese.* 1999;121(1–2):151–179.
2. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc.* 2005;100(469): 322–331.
3. Greenland S. For and against methodologies: some perspectives on recent causal and statistical inference debates. *Eur J Epidemiol.* 2017;32(1):3–20.
4. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology.* 2009; 20(1):3–5.
5. Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc.* 2008;103(482):832–842.
6. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test—comment. *J Am Stat Assoc.* 1980; 75(371):591–593.
7. Hernan MA, Robins JM. *Causal Inference: What If?* Boca Raton, FL: Chapman & Hall / CRC; 2020.
8. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol.* 1986;15(3): 413–419.
9. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol.* 2010;171(6):674–677.