


Defining and identifying local average treatment effects

Ashley I. Naimi ^{*,1} and Brian Whitcomb²

¹Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, 30322, United States

²Department of Epidemiology, School of Public Health and Health Sciences, University of Massachusetts at Amherst, Amherst, MA 01003, United States

*Corresponding author: Ashley I. Naimi, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road, CNR 3023, Atlanta, GA 30322 (ashley.naimi@emory.edu)

Key words: causal inference; identifiability; instrumental variables; local average treatment effect.

Introduction

Average treatment effects (ATEs) are common in epidemiology but depend heavily on key assumptions (eg, counterfactual consistency, positivity, conditional exchangeability) for validity.¹ This has prompted research on effect estimators that require fewer assumptions. For example, when identifying or accurately measuring confounders is not possible, exchangeability may be violated and alternatives to ATEs may be desirable.

Instrumental variables (IVs) are one such alternative. They exploit “instruments”—variables related to outcome only through their relationship with exposure (illustrated in [Figure 1](#)). Here, we show derivation of the IV estimator to illustrate the conditions under which IVs are valid, and what effects they target.

In epidemiology, IVs are sometimes used to estimate compliance-adjusted effects in randomized controlled trials, and define mendelian randomization studies of a confounded relationship between a gene product and outcome. As an example of the former, consider a scenario where we want to estimate the effect of aspirin on headache in a well-defined cohort. We collect information on the randomization stratum (X), the outcome of interest (Y), whether each individual actually took the pill they were randomized to (A)—as well as variables that might predict whether they took the pill or not—and whether they will experience a headache (C). To simplify the explanation below, we assume A can take 2 values: $A = 1$ (took aspirin) and $A = 0$ (took placebo). This scenario can be depicted with the directed acyclic graph in [Figure 1](#).

[Figure 1](#) illustrates important elements that result from randomization. First, X (the instrument) is exogenous, meaning the variable does not depend on other variables in the system. Under exogeneity, we can perform simple unadjusted analyses to quantify the causal effect of X on Y .

Unfortunately, the benefits of this exogeneity extend only to the treatment assignment variable X , which only provides an estimate of the effect of telling people to take the treatment (the effect of X). If interest lies in the effect of the actual treatment A , unadjusted analyses will be confounded. However, IV estimators can leverage X 's exogeneity to answer questions about the effect

of A on Y without the need to adjust for confounding of the $A \rightarrow Y$ relationship.

The instrumental variable estimator

The IV estimator relies on the fact that the effect of A on Y , defined as $E(Y^{a=1} - Y^{a=0})$, is contained in the following 2 effects:

$$E(Y^{x=1} - Y^{x=0}),$$

$$E(A^{x=1} - A^{x=0}),$$

where Y^x and A^x are the outcome and compliance variables that would be observed if X were set to x . Instrumental variables use the information in the $X \rightarrow Y$ and $X \rightarrow A$ paths ([Figure 1](#)), both of which are unconfounded, to estimate the effect of $A \rightarrow Y$.

The IV estimator is defined as:

$$\hat{\psi}_{IV} = \frac{E(Y | X = 1) - E(Y | X = 0)}{E(A | X = 1) - E(A | X = 0)}.$$

Conceptually, this equation estimates the effect of X on A (denominator), and then removes this component from the estimated effect of X on Y (numerator). What remains is the effect of A on Y .

Local average treatment effects

Instrumental variable estimators were introduced early in the 20th century. In 1996, Angrist et al² used potential outcomes to clearly define the target IV estimand. They showed that the IV estimator quantifies:

$$E(Y^{a=1} - Y^{a=0} | A^{x=1} > A^{x=0}).$$

This is referred to as the local average treatment effect (LATE). It is “local” because it estimates the effect of A on Y in a subset of the population defined by $A^{x=1} > A^{x=0}$. This conditioning

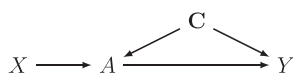


Figure 1. Directed acyclic graph showing the relationship between a randomly assigned treatment indicator X , an indicator of whether individuals took the assigned treatment A , an outcome of interest Y , and common causes of whether individuals took treatment and of the outcome (C).

statement can be understood by referring to Table 1,² which shows that those with $A^{x=1} > A^{x=0}$ would have taken aspirin if they were assigned to it, and would have not taken aspirin if not assigned to it. These individuals are considered “compliers” because this inequality holds only if $A^{x=1} = 1$ and $A^{x=0} = 0$, which (for a binary treatment) is the only way the “monotonicity” condition $A^{x=1} > A^{x=0}$ can be satisfied. As shown below, assuming monotonicity entails no defiers in the population (Table 1). Thus, the LATE quantifies the effect among a subset of the sample who would have taken aspirin had they been assigned to it, AND who would have taken placebo had they been assigned to it.

The IV estimator as a LATE is a recognized problem.³ The conditioning statement for LATEs is defined as a function of potential outcomes; we do not know whom the LATE applies to beyond the vague abstraction referred to as “compliers” (bottom left cell of Table 1), which are not observable with data. Thus, when assuming monotonicity, the LATE is identical to the complier average causal effect (CACE).

In the sections that follow, we show how the IV estimator can be derived, why this derivation requires assumptions not often articulated in applied research, and how these assumptions are important for interpreting IV effects. By definition, a variable X is an instrument if:

Assumption 1. X has a causal effect on A . Formally, this can be written as:

$$A^{x=1} - A^{x=0} \neq 0.$$

Note Hernán and Robins³ generalized this condition to be “ X is associated with A .”

Assumption 2. X affects Y only through A . Formally, this can be written as

$$Y^{a,x=1} = Y^{a,x=0} = Y^a$$

and is often referred to as the exclusion restriction assumption.

Assumption 3. X does not share common causes with the outcome (ie, no confounding of the effect of X on Y), which implies mean exchangeability:

$$E(Y^{A^{x=1},x=1} | X) = E(Y^{A^{x=1},x=1}) \text{ and } E(A^{x=1} | X) = E(A^{x=1}).$$

With these assumptions, we can rewrite the equation for the IV estimator above to equal the LATE, which allows us to establish identifiability. In effect, we will prove that under certain assumptions:

$$\frac{\overbrace{E(Y | X = 1) - E(Y | X = 0)}^{\text{IV Estimator}}}{\overbrace{E(A | X = 1) - E(A | X = 0)}^{\text{Local Average Treatment Effect}}} = E(Y^{a=1} - Y^{a=0} | A^{x=1} > A^{x=0}).$$

Stable unit treatment value assumption (consistency and no interference)

If we assume counterfactual consistency and no interference,¹ we can rewrite the terms in the numerator of the IV estimator. Throughout, we demonstrate derivations under the scenario where $X = 1$:

$$E(Y | X = 1) = E(Y^{a=A,x=1} | X = 1).$$

Here, $Y^{a=A,x=1}$ is the outcome that would be observed if X were set to 1 ($x = 1$ in the arguments of the potential outcome), under the A value observed in the sample ($a = A$).

Exclusion restriction

As noted above, the exclusion restriction states that $Y^{a,x=1} = Y^{a,x=0} = Y^a$, allowing simplification of the right-hand side of the above equation:

$$E(Y^{a=A,x=1} | X = 1) = E(Y^{a=A} | X = 1).$$

Exchangeability

If Assumption 3 holds, we can assume mean exchangeability, which allows us to write:

$$E(Y^{a=A} | X = 1) = E(Y^{a=A}).$$

Table 1. Table of response types showing never-takers, always-takers, compliers, and defiers, as well as the effects (defined via potential outcomes) that result for each scenario^a.

		$A^{x=0}$	
		0	1
$A^{x=1}$	0	Never-taker: $Y[x=1,A^{x=1}=0] - Y[x=0,A^{x=0}=0] = 0$	Defier: $Y[x=1,A^{x=1}=0] - Y[x=0,A^{x=0}=1] = -[Y^{x=1} - Y^{x=0}]$
	1	Complier: $Y[x=1,A^{x=1}=1] - Y[x=0,A^{x=0}=0] = Y^{x=1} - Y^{x=0}$	Always-taker: $Y[x=1,A^{x=1}=1] - Y[x=0,A^{x=0}=1] = 0$

^aTable variable definitions: A^x , compliance that would be observed if randomized to x ; Y^x , outcome that would be observed if randomized to x ; $Y^{[x,A^x]}$, outcome that would be observed if randomized to x under compliance that would be observed if randomized to x . Table modified from Angrist et al.²

It is important to reiterate here that $Y^{a=A}$ refers to the potential outcome that would be observed if the compliance variable was set to the value that was observed in the sample. In the actual dataset, this consists of 2 possible potential outcomes: $Y^{a=1}$ for those with $A = 1$, and $Y^{a=0}$ for those with $A = 0$. We can then reformulate the right-hand side of the above equation as:

$$E(Y^{a=A}) = E[Y^{a=0} + (Y^{a=1} - Y^{a=0})A^{x=1}].$$

In this equation, note that if $A^{x=1} = 1$, the expectation on the right-hand side resolves to $Y^{a=1}$, whereas if $A^{x=1} = 0$ it resolves to $Y^{a=0}$. Thus, the above equality holds.

More mathematical manipulation

What we have shown is that, under the stable unit treatment value assumption (SUTVA), exclusion restriction (Assumption 2), and exchangeability (Assumption 3), we can rewrite the first term in the numerator of the IV estimator as:

$$E(Y | X = 1) = E[Y^{a=0} + (Y^{a=1} - Y^{a=0})A^{x=1}],$$

which means the full numerator of the IV estimator can be written as:

$$E[Y^{a=0} + (Y^{a=1} - Y^{a=0})A^{x=1}] - E[Y^{a=0} + (Y^{a=1} - Y^{a=0})A^{x=0}].$$

At this point, the above equation can be rewritten as the product of 2 differences. Rearranging, the numerator becomes:

$$E[(Y^{a=1} - Y^{a=0})(A^{x=1} - A^{x=0})],$$

which holds only under the assumptions listed (SUTVA + Assumptions 2 and 3). This latter equation states that the causal effect of X on Y (the numerator of the IV estimator) is the product of the effect of A on Y and X on A . Therefore, dividing by the effect of X on A gives us the effect of A on Y .

Performing further manipulation on the product of these 2 effects provides additional insights on what the IV estimator quantifies. Consider that for a binary A the difference in potential outcomes ($A^{x=1} - A^{x=0}$) can only result in 3 possible values: $\{1, 0, -1\}$. Again, Assumption 1 states that $A^{x=1} - A^{x=0} \neq 0$, which allows for the following decomposition:

$$\begin{aligned} E[(Y^{a=1} - Y^{a=0})(A^{x=1} - A^{x=0})] \\ = E(Y^{a=1} - Y^{a=0} | A^{x=1} - A^{x=0} = 1)P(A^{x=1} - A^{x=0} = 1) \\ + E(Y^{a=1} - Y^{a=0} | A^{x=1} - A^{x=0} = 0)P(A^{x=1} - A^{x=0} = 0) \\ + E(Y^{a=1} - Y^{a=0} | A^{x=1} - A^{x=0} = -1)P(A^{x=1} - A^{x=0} = -1). \end{aligned}$$

In effect, this states that product of the effect of A on Y and X on A (the numerator of the IV estimator) is a weighted average of the ATE among the compliers (those with $A^{x=1} - A^{x=0} = 1$) and the defiers (those with $A^{x=1} - A^{x=0} = -1$).

Monotonicity

Monotonicity (formally, $A^{x=1} > A^{x=0}$) states that there are no defiers in the population. If we add this assumption, we can

remove the defiers from the equation above:

$$\begin{aligned} E[(Y^{a=1} - Y^{a=0})(A^{x=1} - A^{x=0})] &= E(Y^{a=1} - Y^{a=0} | A^{x=1} - A^{x=0} = 1) \\ &\quad P(A^{x=1} - A^{x=0} = 1). \end{aligned}$$

Noting that $P(A^{x=1} - A^{x=0} = 1) = E(A^{x=1} - A^{x=0})$, we get:

$$\begin{aligned} \frac{E(Y | X = 1) - E(Y | X = 0)}{E(A | X = 1) - E(A | X = 0)} \\ = \frac{E[(Y^{a=1} - Y^{a=0})(A^{x=1} - A^{x=0})]}{E(A^{x=1} - A^{x=0})} \\ = E(Y^{a=1} - Y^{a=0} | A^{x=1} > A^{x=0}). \end{aligned}$$

Conclusion

Instrumental variable methods are popular, and often used because they do not require “no unmeasured confounding” assumptions. However, other assumptions are required to interpret IV results as causal effects, including counterfactual consistency, no interference, and (to interpret the IV estimand as a complier average causal effect) monotonicity. This has important implications for common uses, such as mendelian randomization.

IV methods can also be used to quantify effects other than the CACE.⁴ We have shown how observed data (randomized or observational) can be connected to causal effects defined using potential outcomes when IV estimators are used. Precisely understanding what these steps imply will enable researchers to better interpret and understand empirical results when causal effects are estimated using instrumental variables.

Funding

A.I.N. was supported by NIH grant R01HD102313.

Conflict of interest

The authors declare no conflicts of interest.

Data availability

No data was used for this work.

References

1. Naimi AI, Whitcomb BW. Defining and identifying average treatment effects. *Am J Epidemiol*. 2023;192(5):685-687. <https://doi.org/10.1093/aje/kwad012>
2. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91(434):444-455. <https://doi.org/10.1080/01621459.1996.10476902>
3. Swanson SA, Hernán MA. Think globally, act globally: an epidemiologist's perspective on instrumental variable estimation. *Stat Sci*. 2014;29(3):371-374.
4. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17(4):360-372. <https://doi.org/10.1097/01.ede.0000222409.00878.37>