Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes

Sonja A. Swanson, Miguel A. Hernán, Matthew Miller, James M. Robins & Thomas S. Richardson

To cite this article: Sonja A. Swanson, Miguel A. Hernán, Matthew Miller, James M. Robins & Thomas S. Richardson (2018) Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes, Journal of the American Statistical Association, 113:522, 933-947, DOI: 10.1080/01621459.2018.1434530

To link to this article: https://doi.org/10.1080/01621459.2018.1434530

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 522, 933–947, Reviews
https://doi.org/10.1080/01621459.2018.1434530

Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes

Sonja A. Swanson[a,b], Miguel A. Hernán[b,c,d], Matthew Miller[b,e], James M. Robins[b,c,*], and Thomas S. Richardson[f,*]

[a]Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; [b]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA; [c]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA; [d]Harvard-MIT Division of Health Sciences and Technology, Boston, MA; [e]Department of Health Sciences, Northeastern University, Boston, MA; [f]Department of Statistics, University of Washington, Seattle, WA

ABSTRACT

Several methods have been proposed for partially or point identifying the average treatment effect (ATE) using instrumental variable (IV) type assumptions. The descriptions of these methods are widespread across the statistical, economic, epidemiologic, and computer science literature, and the connections between the methods have not been readily apparent. In the setting of a binary instrument, treatment, and outcome, we review proposed methods for partial and point identification of the ATE under IV assumptions, express the identification results in a common notation and terminology, and propose a taxonomy that is based on sets of identifying assumptions. We further demonstrate and provide software for the application of these methods to estimate bounds. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2015
Revised October 2017

KEYWORDS
Average treatment effect; Causal graphical model; Instrument; Instrumental variable; Partial identification; Single world intervention graph

1. Introduction

This article provides a comprehensive review of the methods for partial identification of the average treatment effect (ATE) of a time-fixed binary treatment on a binary outcome using a binary instrumental variable (IV). These methods and their underlying assumptions have not been previously presented in a common set of notation and terminology because the methodological literature is widespread across journals of statistics, economics, epidemiology, and computer science. By unifying the notation and terminology, we provide a taxonomy of the assumptions that (combined with data) lead to partial or point identification. Our work makes apparent the heretofore obscured relationships between the different combinations of assumptions and the ATE bounds they identify. We also provide an empirical example of estimating the ATE under all proposed sets of assumptions. Finally, although software is available to implement some of these methods (Beresteanu and Manski 2000; Palmer et al. 2011; McCarthy, Millimet, and Roy 2015; Chernozhukov et al. 2015), we include comprehensive statistical software for partial identification of the ATE under all proposed sets of assumptions (supplementary materials). Space limitations preclude a detailed discussion of methods for incorporating random variability into the partial identification framework. However, in Appendix S1 we give a brief guide to the relevant literature.

This article is organized as follows. In Section 2, we describe the taxonomy of IV assumptions that lead to partial identification of the ATE of a binary treatment on a binary outcome. We relate these results to the IV inequalities in Section 3 and to graphical representations in Section 4. In Section 5, we extend this taxonomy to additional assumptions considered in combination with the IV assumptions; we briefly review some extensions to continuous outcomes and other settings in Section 6. In Section 7, we demonstrate the estimation of bounds in studying the effect of Medicaid coverage on emergency department visits from the Oregon Health Insurance Experiment (Finkelstein 2013; Taubman et al. 2014). We conclude with a brief discussion (Section 8).

2. Bounds on the Population Average Treatment Effect (ATE) Under Instrumental Variable Assumptions

Suppose that our data consist of n independent, identically distributed draws from a joint distribution P. Let X be a binary treatment (1: treated, 0: not treated) and Y a binary outcome (1: yes, 0: no). Without loss of generality, we assume a lower probability of Y is preferable. Our primary interest is in the average treatment effect (ATE) on the additive scale:

$$\text{ATE} = \text{E}[Y^{x=1}] - \text{E}[Y^{x=0}], \quad (1)$$

where the random variable $Y^{x=1}$ indicates the counterfactual outcome for a subject had she been treated ($X = 1$) and likewise $Y^{x=0}$ indicates the counterfactual outcome for a subject had she been untreated ($X = 0$). We will suppose that the observed data

CONTACT Sonja A. Swanson s.swanson@erasmusmc.nl Department of Epidemiology, Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
*Co-senior authors
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

(X, Y) is related to the counterfactual via the usual consistency assumption:

$$Y = (1 - X)Y^{x=0} + XY^{x=1} \equiv Y^X. \tag{2}$$

Before even looking at the data or making any assumptions, we know nothing about the ATE: in our all-binary setting, it could range from -1 (i.e., treatment universally prevents the outcome) to 1 (i.e., treatment universally causes the outcome). However, the data provide information that (still, without any assumptions) cuts the width of this range in half (Robins 1989; Manski 1990). This is essentially a missing data problem: we only observe one of the two counterfactuals $Y^{x=0}$ and $Y^{x=1}$ for each subject i (e.g., for a treated subject i we observe $Y^{x=1}$ but not $Y^{x=0}$). By imputing the unobserved counterfactuals to their most extreme values possible, we can identify the lower and upper bounds on the range of possible estimates for the ATE that are consistent with the observed data. The bounds will always have width 1, hence will include zero (i.e., the null), and thus cannot identify the direction of the treatment effect.

In the remainder of this section, we discuss how narrower bounds on the ATE can be obtained if one is willing to make assumptions about a binary pretreatment variable, Z, that is associated with X. This variable Z is referred to as an instrumental variable (IV), or an instrument, when two unverifiable assumptions hold: (i) the exclusion restriction, and (ii) exchangeability. The exclusion restriction (i) says that the instrument Z cannot affect the outcome except through its potential effect on treatment X, as formalized below. Exchangeability (ii) says that, at baseline, subjects with $Z = 0$ are comparable to subjects with $Z = 1$. Although these assumptions are not verifiable, they have testable implications; we will return to this point in Section 3.

There are several different versions of both the exchangeability and exclusion assumptions, and thus also of what constitutes an instrument. Before formalizing these versions, consider two settings.

First, consider the paradigmatic IV example of a double-blind placebo-controlled randomized trial with noncompliance. Let Z denote the assigned treatment arm, and X the treatment received. If the double-blinding is successfully maintained and there is no placebo effect, there can be no effect of treatment assignment on the outcome other than via the treatment, thus the exclusion assumption will hold. Furthermore, those assigned to different arms are exchangeable owing to randomization. Thus, in this circumstance Z will satisfy all versions of (i) and (ii) required of an instrument.

Second, consider the common applications of IV methods in observational studies in which investigators propose a pretreatment instrument, Z. Examples of proposed instruments Z include calendar time, geographic variation, provider preference, and genetic variants (Davies et al. 2013). Importantly, in observational studies, no version of (i) or (ii) can be guaranteed. Moreover, note that exclusion (i) and some versions of exchangeability (ii) are agnostic about whether the proposed instrument Z has a causal effect on the treatment X (like the randomization assignment in randomized trials), or is just a surrogate for an (unmeasured) causal instrument. Many commonly proposed instruments in observational studies may be conceptualized as the latter (Robins 1989; Dawid 2003; Hernán and Robins 2006).

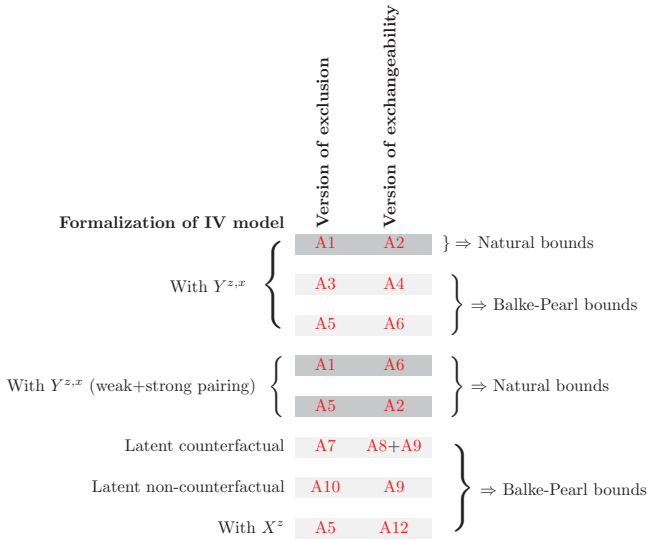

Figure 1. Combinations of assumptions for obtaining the natural or Balke-Pearl bounds on the average treatment effect for dichotomous instrument, treatment, and outcome, as discussed in Section 2. Note the latent noncounterfactual IV model further requires (A11). The row-wise pairs of assumptions that lead to the natural bounds are shaded dark gray, while the row-wise sets of assumptions that lead to the Balke-Pearl bounds are shaded light gray.

With the paradigmatic example of a double-blind trial and common applications in mind, we now turn to formal definitions of instruments. A summary of these formalizations is presented in Figure 1. Their relationships are summarized in Table 1.

2.1. Formalization of the IV Model with $Y^{z,x}$ Counterfactuals

To formally define the properties required of an instrument Z, we define "joint" counterfactual outcomes $Y^{z,x}$ corresponding to the outcome that a subject would have if (possibly contrary to fact) she had been assigned to treatment arm z and then received (again possibly contrary to fact) the treatment x.

We also define the counterfactual Y^x corresponding to the outcome the subject would have if she had her observed Z, but we intervened on $X = x$. These counterfactuals are related by a form of consistency assumption:

$$Y^x = (1 - Z)Y^{z=0,x} + ZY^{z=1,x} \equiv Y^{Z,x}. \tag{3}$$

As noted above, there are alternative definitions of exclusion and exchangeability in the literature. We begin by describing the weakest version of both.

Marginal stochastic exclusion

$$\mathrm{E}[Y^{z,x}] = \mathrm{E}[Y^{z',x}], \quad \text{for all } z, z', x. \tag{A1}$$

The weakest exclusion restriction, thus, means that at the population level the average (controlled) directed effects of Z on Y holding X fixed are zero.

Marginal exchangeability of $Y^{z,x}$ counterfactuals

$$Z \perp\!\!\!\perp Y^{z,x}; \quad \text{for all } z, x. \tag{A2}$$

Table 1. Gains in identification comparing sets of assumptions leading to partial identification of the average treatment effect for a dichotomous Z, X, and Y.

Initial assumption set	Strengthened assumption set	Gains in identification (If any)
No data and no assumptions	Data only	Width of bounds reduced by 1/2 (width of bounds = 1)
Data only	A1 + A2	Width of bounds = $\Pr[X=0\|Z=1]+\Pr[X=1\|Z=0]$[†]
A1 + A2	A1 + A6	No gains
A1 + A2	A5 + A2	No gains
A1 + A2	A3 + A4	Narrower bounds if and only if inequalities (6) are violated
A3 + A4	A5 + A6	No gains
A5 + A6	A5 + A12	No gains
A5 + A12	A5 + A12 + A13	Potentially narrower bounds depends on specified proportion in A13
A5 + A12 + A13	A5 + A12 + A13 + A14	Improvement depends on assumed limits in A14
A1 + A2	A1 + A2 + A15	Identifies direction of effect with the same upperbound
A1 + A2	A1 + A2 + A16	May improve lowerbound on each mean counterfactual
A1 + A2	A1 + A2 + A17	Point identification
A1 + A2	A1 + A2 + A18	Point identification
A7 + A8 + A9	A7 + A8 + A9+(A19 or A20)	Point identification

NOTES: Note the following assumptions imply one another and therefore are not included in nested assumption sets: A5 ⇒ A3 ⇒ A1; A6 ⇒ A4 ⇒ A2.
[†]Here we implicitly suppose that $\Pr[X=0|Z=1]+\Pr[X=1|Z=0] < \min\{\Pr[X=0|Z=0]+\Pr[X=1|Z=1],\ 1\}$.

This assumption follows from randomization of Z, but is a weaker condition.

Theorem 1. Under (A1) and (A2), we have:

$$Z \perp\!\!\!\perp Y^x, \qquad \text{for all } x, \tag{4}$$

and further $\mathrm{E}[Y^x] = \mathrm{E}[Y^{z,x}]$ for all z.

Robins (1989) and Manski (1990) obtained sharp lower and upper bounds on the ATE under (4). These bounds are given in Tables 2 and 3, respectively. Sharp bounds for the mean counterfactuals $\mathrm{E}[Y^{x=0}]$ and $\mathrm{E}[Y^{x=1}]$ are given in Appendix S2. These bounds on the ATE and the counterfactual means are also sharp under the larger model given by (A1) and (A2). These ATE bounds are often referred to as the "natural" or the "Robins-Manski" bounds in the literature. The width of the natural bounds is no greater than the sum of the noncompliance proportions in each arm: $\Pr[X=1|Z=0]+\Pr[X=0|Z=1]$ (Balke and Pearl 1997). As such, the width of the bounds may be substantially narrower than those identified from the data on X and Y alone (which were of width 1).

Table 2. Lower bounds for identification of the average treatment effect under sets of assumptions described in Figure 1.

Assumption set	Lower bound*
Data only	$-p_{y_0,x_1} \;-\; p_{y_1,x_0} = (p_{y_1\|x_1}-1)p_{x_1} - p_{y_1\|x_0}p_{x_0}$
A1 + A2**	$\max\left\{\begin{array}{l} -p_{y_0,x_1\|z_0} - p_{y_1,x_0\|z_0} \\ -p_{y_0,x_1\|z_1} - p_{y_1,x_0\|z_1} \\ p_{y_1\|z_0} - p_{y_1\|z_1} - p_{y_1,x_0\|z_0} - p_{y_0,x_1\|z_1} \\ p_{y_1\|z_1} - p_{y_1\|z_0} - p_{y_1,x_0\|z_1} - p_{y_0,x_1\|z_0} \end{array}\right\}$
A3 + A4***	$\max\left\{\begin{array}{l} -p_{y_0,x_1\|z_0} - p_{y_1,x_0\|z_0} \\ -p_{y_0,x_1\|z_1} - p_{y_1,x_0\|z_1} \\ p_{y_1\|z_0} - p_{y_1\|z_1} - p_{y_1,x_0\|z_0} - p_{y_0,x_1\|z_1} \;=\; p_{y_1,x_1\|z_0} + p_{y_0,x_0\|z_1} - 1 \\ p_{y_1\|z_1} - p_{y_1\|z_0} - p_{y_1,x_0\|z_1} - p_{y_0,x_1\|z_0} \;=\; p_{y_1,x_1\|z_1} + p_{y_0,x_0\|z_0} - 1 \\ p_{y_1,x_1\|z_0} - p_{y_1,x_1\|z_1} - p_{y_1,x_0\|z_1} - p_{y_0,x_1\|z_0} - p_{y_1,x_0\|z_0} \\ p_{y_1,x_1\|z_1} - p_{y_1,x_1\|z_0} - p_{y_1,x_0\|z_0} - p_{y_0,x_1\|z_1} - p_{y_1,x_0\|z_1} \\ p_{y_0,x_0\|z_1} - p_{y_0,x_1\|z_1} - p_{y_1,x_0\|z_1} - p_{y_0,x_1\|z_0} - p_{y_0,x_0\|z_0} \\ p_{y_0,x_0\|z_0} - p_{y_0,x_1\|z_0} - p_{y_1,x_0\|z_0} - p_{y_0,x_1\|z_1} - p_{y_0,x_0\|z_1} \end{array}\right\}$
A5 + A12 + A13	see Appendix
A5 + A12 + A13 + A14	see Appendix
A1 + A2 + A15	same as A1 + A2
A1 + A2 + A16	$\max\left\{\begin{array}{c} p_{y_1\|x_1,z_1}p_{x_1\|z_1} + p_{y_1\|x_0,z_1}p_{x_0\|z_1} \\ p_{y_1\|x_1,z_0}p_{x_1\|z_0} \end{array}\right\} - \min\left\{\begin{array}{c} p_{y_1\|x_0,z_0}p_{x_0\|z_0} + p_{x_1\|z_0} \\ p_{y_1\|x_0,z_1}p_{x_0\|z_1} + p_{x_1\|z_1} \end{array}\right\}$
A1 + A2 + A17****	$\frac{p_{y_1\|z_1}-p_{y_1\|z_0}}{p_{x_1\|z_1}-p_{x_1\|z_0}}$
A1 + A2 + A18	$p_{y_1\|x_0}p_{x_0}(\exp(\psi_0)-1) + p_{y_1\|x_1}p_{x_1}(1-\exp(-\psi_0))$ where $\exp(-\psi_0) = 1 - \frac{p_{y_1\|z_1}-p_{y_1\|z_0}}{p_{y_1\|x_1,z_1}p_{x_1\|z_1}-p_{y_1\|x_1,z_0}p_{x_1\|z_0}}$

NOTES: *$p_{y_k,x_j|z_i} \equiv \Pr[Y=k, X=j|Z=i]$; $p_{y_k|x_j,z_i} \equiv \Pr[Y=k|X=j,Z=i]$; $p_{y_k|x_j} \equiv \Pr[Y=k|X=j]$; $p_{y_k|z_i} \equiv \Pr[Y=k|Z=i]$; $p_{x_j|z_i} \equiv \Pr[X=j|Z=i]$; $p_{x_j} \equiv \Pr[X=j]$; $p_{z_i} \equiv \Pr[Z=i]$.
**Some authors use the term "natural bounds" to refer solely to the fourth term here.
***See Section 2 for additional assumption sets that likewise lead to the Balke-Pearl bounds.
****Assumption set A7 + A8 + A9+(A19 or A20) also leads to this same expression.

Table 3. Upper bounds for identification of the average treatment effect under sets of assumptions described in Figure 1.

Assumption set	Upper bound*
Data only	$p_{y_1,x_1} + p_{y_0,x_0} = (1 - p_{y_1\|x_0})p_{x_0} + p_{y_1\|x_1}p_{x_1}$
A1 + A2**	$\min\left\{\begin{array}{l} p_{y_1,x_1\|z_0} + p_{y_0,x_0\|z_0} \\ p_{y_1,x_1\|z_1} + p_{y_0,x_0\|z_1} \\ p_{y_1\|z_0} - p_{y_1\|z_1} + p_{y_0,x_0\|z_0} + p_{y_1,x_1\|z_1} \\ p_{y_1\|z_1} - p_{y_1\|z_0} + p_{y_0,x_0\|z_1} + p_{y_1,x_1\|z_0} \end{array}\right\}$
A3 + A4***	$\min\left\{\begin{array}{l} p_{y_1,x_1\|z_0} + p_{y_0,x_0\|z_0} \\ p_{y_1,x_1\|z_1} + p_{y_0,x_0\|z_1} \\ p_{y_1\|z_0} - p_{y_1\|z_1} + p_{y_0,x_0\|z_0} + p_{y_1,x_1\|z_1} = 1 - p_{y_0,x_1\|z_0} + p_{y_1,x_0\|z_1} \\ p_{y_1\|z_1} - p_{y_1\|z_0} + p_{y_0,x_0\|z_1} + p_{y_1,x_1\|z_0} = 1 - p_{y_0,x_1\|z_1} + p_{y_1,x_0\|z_0} \\ -p_{y_0,x_1\|z_0} + p_{y_0,x_1\|z_1} + p_{y_0,x_0\|z_1} + p_{y_1,x_1\|z_0} + p_{y_0,x_0\|z_0} \\ -p_{y_0,x_1\|z_1} + p_{y_0,x_1\|z_0} + p_{y_0,x_0\|z_0} + p_{y_1,x_1\|z_1} + p_{y_0,x_0\|z_1} \\ -p_{y_1,x_0\|z_1} + p_{y_1,x_1\|z_1} + p_{y_0,x_0\|z_1} + p_{y_1,x_1\|z_0} + p_{y_1,x_0\|z_0} \\ -p_{y_1,x_0\|z_0} + p_{y_1,x_1\|z_0} + p_{y_0,x_0\|z_0} + p_{y_1,x_1\|z_1} + p_{y_1,x_0\|z_1} \end{array}\right\}$
A5 + A12 + A13	see Appendix
A5 + A12 + A13 + A14	see Appendix
A1 + A2 + A15	$-\lvert p_{y_1\|z_1} - p_{y_1\|z_0}\rvert$
A1 + A2 + A16	$\min\left\{\begin{array}{c} p_{y_1\|x_1,z_1}p_{x_1\|z_1} + p_{x_0\|z_1} \\ p_{y_1\|x_1,z_0}p_{x_1\|z_0} + p_{x_0\|z_0} \end{array}\right\} - \max\left\{\begin{array}{c} p_{y_1\|x_0,z_0}p_{x_0\|z_0} + p_{y_1\|x_1,z_0}p_{x_1\|z_0} \\ p_{y_1\|x_0,z_1}p_{x_0\|z_1} \end{array}\right\}$
A1 + A2 + A17****	$\frac{p_{y_1\|z_1} - p_{y_1\|z_0}}{p_{x_1\|z_1} - p_{x_1\|z_0}}$
A1 + A2 + A18	$p_{y_1\|x_0}p_{x_0}(\exp(\psi_0) - 1) + p_{y_1\|x_1}p_{x_1}(1 - \exp(-\psi_0))$ where $\exp(-\psi_0) = 1 - \frac{p_{y_1\|z_1} - p_{y_1\|z_0}}{p_{y_1\|x_1,z_1}p_{x_1\|z_1} - p_{y_1\|x_1,z_0}p_{x_1\|z_0}}$

NOTES: $^*p_{y_k,x_j|z_i} \equiv \Pr[Y = k, X = j|Z = i]$; $p_{y_k|x_j,z_i} \equiv \Pr[Y = k|X = j, Z = i]$; $p_{y_k|x_j} \equiv \Pr[Y = k|X = j]$; $p_{y_k|z_i} \equiv \Pr[Y = k|Z = i]$; $p_{x_j|z_i} \equiv \Pr[X = j|Z = i]$; $p_{x_j} \equiv \Pr[X = j]$; $p_{z_i} \equiv \Pr[Z = i]$.

**Some authors use the term "natural bounds" to refer solely to the fourth term here.

***See Section 2 for additional assumption sets that likewise lead to the Balke-Pearl bounds.

****Assumption set A7 + A8 + A9+(A19 or A20) also leads to this same expression.

Next, consider the strengthened exchangeability and exclusion assumptions.

Joint stochastic exclusion

$$\begin{aligned}&\Pr[Y^{z=0,x=0} = y, Y^{z=0,x=1} = y'] \\ &\quad = \Pr[Y^{z=1,x=0} = y, Y^{z=1,x=1} = y'] \text{ for all } y, y'; \quad \text{(A3)}\end{aligned}$$

Partial joint exchangeability of $Y^{z,x}$ counterfactuals

$$Z \perp\!\!\!\perp \{Y^{z=0,x=0}, Y^{z=0,x=1}\}, \quad Z \perp\!\!\!\perp \{Y^{z=1,x=0}, Y^{z=1,x=1}\}. \tag{A4}$$

Theorem 2. Under (A3) and (A4), the following joint exchangeability holds:

$$Z \perp\!\!\!\perp \{Y^{x=0}, Y^{x=1}\}. \tag{5}$$

Richardson and Robins (2014) obtained sharp bounds on the ATE under (5), which are also sharp under (A3) and (A4). These bounds are identical to those obtained by Balke and Pearl (1997) under stronger assumptions that we discuss below. Again, the bounds are given in Tables 2 and 3. In the literature, these expressions are often referred to as the "Balke-Pearl" or the "sharp IV" bounds. The latter terminology can cause some confusion, as the natural bounds can likewise be considered sharp, albeit under different assumptions—for example, (4) but not (5).

The natural bounds obtained under (4) will be wider than the bounds obtained under (5) if and only if at least one of the following inequalities is violated:

$$\begin{aligned}\Pr[Y = 1, X = 0 \mid Z = 1) &\leq \Pr[Y = 1, X = 0 \mid Z = 0]; \\ \Pr[Y = 0, X = 0 \mid Z = 1] &\leq \Pr[Y = 0, X = 0 \mid Z = 0]; \\ \Pr[Y = 1, X = 1 \mid Z = 1] &\geq \Pr[Y = 1, X = 1 \mid Z = 0]; \\ \Pr[Y = 0, X = 1 \mid Z = 1] &\geq \Pr[Y = 0, X = 1 \mid Z = 0]. \quad (6)\end{aligned}$$

At the end of this section, we provide an interpretation of these equations in terms of counterfactual variables X^z that will make these conditions more intuitive.

Next, we consider even stronger versions of exclusion and exchangeability:

Individual-level exclusion

$$Y^{z,x} = Y^{z',x} = Y^x \text{ for all } x, z, z'. \tag{A5}$$

In other words, there is no individual direct effect of Z on Y relative to X.

Full joint exchangeability of $Y^{z,x}$ counterfactuals

$$Z \perp\!\!\!\perp \{Y^{z=0,x=0}, Y^{z=0,x=1}, Y^{z=1,x=0}, Y^{z=1,x=1}\}. \tag{A6}$$

The assumptions (A5) and (A6) do not lead to narrower bounds than the Balke-Pearl bounds.

We can consider bounds under other combinations of these exclusion and exchangeability assumptions. Interestingly, the model defined by the weakest exclusion restriction (A1) and the strongest exchangeability assumption (A6) leads to the natural bounds even when (6) fails to hold. Analogously, the model

defined by the strongest exclusion restriction (A5) and the weakest exchangeability assumption (A2) also gives the natural bounds. Both of these claims were verified with direct calculation using the computational geometry package (rcdd) in R.

2.2. Latent Formulation of the IV Model

Many articles formulate the IV model in terms of an unobserved confounder, U between X and Y; see, for example, Dawid (2003) and Didelez, Meng, and Sheehan (2010). Under this framework, we may alternatively define an IV model via the following three assumptions:

$$\mathrm{E}[Y^{z,x} \mid U] = \mathrm{E}[Y^{z',x} \mid U], \quad \text{for all } z, z'; \tag{A7}$$

$$Y^{z,x} \perp\!\!\!\perp (Z, X^z) \mid U; \tag{A8}$$

$$Z \perp\!\!\!\perp U. \tag{A9}$$

We will refer to the combination of (A7), (A8), and (A9) as the latent counterfactual IV model. In words, (A7) states that within strata defined by U, Z has no population-level direct effect on Y, relative to X (i.e., a version of exclusion). Assumptions (A8) and (A9) are forms of exchangeability. The assumption (A8) states that given U, the joint effect of Z and X on Y is unconfounded, where X^z is defined as the counterfactual treatment that a participant would receive under instrument level $Z = z$. The assumption (A9) would be true, for example, if Z were randomly assigned, and U were baseline covariates.

The assumption (A8) plus consistency implies

$$\mathrm{E}[Y^{z,x}|U] = \mathrm{E}[Y \mid Z{=}z, X{=}x, U] \tag{7}$$

and

$$\mathrm{E}[Y^{z,x}] = \int \mathrm{E}[Y \mid X{=}x, U{=}u, Z{=}z]p(u)du. \tag{8}$$

It then follows that (A7) and (A8) together imply that

$$\mathrm{E}[Y^{z,x}] = \mathrm{E}[Y^{Z,x}] \equiv \mathrm{E}[Y^x]; \tag{9}$$

$$\mathrm{E}[Y^x] = \int \mathrm{E}[Y \mid X{=}x, U{=}u]p(u)du. \tag{10}$$

Similarly, (A7) and (7) imply:

$$Z \perp\!\!\!\perp Y \mid X, U. \tag{A10}$$

Some authors have defined the IV model without referring to counterfactuals at all (e.g., Dawid 2003). These authors define the latent noncounterfactual IV model by assumptions (A9) and (A10), together with the assumption that

$$\mathrm{E}^{\mathrm{int},x}[Y] = \int \mathrm{E}[Y \mid X{=}x, U{=}u]p(u)du, \tag{A11}$$

where $\mathrm{E}^{\mathrm{int},x}[Y]$ is the expectation of Y under an intervention to set X to x. The ATE is then defined by

$$\mathrm{ATE} = \int (\mathrm{E}[Y \mid X{=}1, U{=}u] - \mathrm{E}[Y \mid X{=}0, U{=}u])\, p(u)du. \tag{11}$$

Note that the above latent counterfactual IV model defined by (A7), (A8), and (A9) implies the noncounterfactual IV model given by (A9), (A10), and (A11).

Interestingly, the latent counterfactual IV model defined by (A7), (A8), and (A9) is exactly the IV model defined by (A5) and (A6), discussed earlier, when $U = (Y^{z=0,x=0}, Y^{z=1,x=0}, Y^{z=0,x=1}, Y^{z=1,x=1})$. To see this, note that with this choice of U, (A7) becomes the individual-level exclusion restriction (A5), (A8) becomes a tautology, and (A9) is (A6).

Consequently, the bounds on the ATE obtained under the latent counterfactual model defined via (A7), (A8), and (A9) are logically at least as large as the Balke-Pearl bounds of model (A5) and (A6). Furthermore, it was shown by Dawid (2003) that the sharp bounds in the noncounterfactual IV model (given by (A9), (A10), and (A11)) were also the Balke-Pearl bounds. It follows that the bounds are also sharp for the latent counterfactual model, since it is a submodel of the noncounterfactual model.

In the course of proving the aforementioned result, Dawid (2003) showed that the sharp bounds for $Y^{x=1}$ were variation independent of those for $Y^{x=0}$. Variation independence is needed to conclude that the upper bound for the ATE is the upper bound for $\mathrm{E}[Y^{x=1}]$ minus the lower bound for $\mathrm{E}[Y^{x=0}]$ and that the lower bound for the ATE is the lower bound for $\mathrm{E}[Y^{x=1}]$ minus the upper bound for $\mathrm{E}[Y^{x=0}]$; see also Manski (2003) and Kitagawa (2009).

2.3. Formalization of the IV Model Including Counterfactual Treatments X^z

We now consider the strongest version of exchangeability:

Randomization assumption

$$Z \perp\!\!\!\perp \{Y^{z=0,x=0}, Y^{z=0,x=1}, Y^{z=1,x=0}, Y^{z=1,x=1}, X^{z=0}, X^{z=1}\}. \tag{A12}$$

Balke and Pearl (1997) formulated the IV model as individual-level exclusion (A5) and

$$Z \perp\!\!\!\perp \{Y^{x=0}, Y^{x=1}, X^{z=0}, X^{z=1}\}. \tag{12}$$

Note that (A5) and (12) are equivalent to (A5) and (A12).

As noted earlier, the bounds derived by Balke and Pearl (1997) using these strengthened exclusion and exchangeability assumptions are the same as those obtained under (5). That these assumptions were stronger than necessary was previously conjectured and demonstrated in the most extreme special case (Robins and Greenland 1996).

This maximal exchangeability (A12) in addition to (A5) also implies:

$$\begin{aligned} &Z \perp\!\!\!\perp \{Y^{x=0}, X^{z=0}\};\ Z \perp\!\!\!\perp \{Y^{x=1}, X^{z=0}\};\\ &Z \perp\!\!\!\perp \{Y^{x=0}, X^{z=1}\};\ Z \perp\!\!\!\perp \{Y^{x=1}, X^{z=1}\}. \end{aligned} \tag{13}$$

Conditions (A5) and (13), which are particularly suited to the single-world intervention graph (SWIG) framework discussed in Section 4, are also sufficient for the Balke-Pearl bounds (Richardson and Robins 2014). This seems surprising since although (with (A5)) (A12) implies both (5) and (13), neither of these imply one another.

When counterfactual treatments X^z are defined, we may characterize subjects by one of four mutually exclusive compliance types: always-takers ($X^{z=0} = X^{z=1} = 1$); never-takers ($X^{z=0} = X^{z=1} = 0$); compliers ($X^{z=0} = 0, X^{z=1} = 1$); and defiers ($X^{z=0} = 1, X^{z=1} = 0$). With these compliance types

defined, we can now provide an interpretation of the inequalities (6) presented earlier, at least one of which will be violated if and only if the natural bounds are wider than the Balke-Pearl bounds. Specifically, when one of the inequalities (6) is violated, the proportion of defiers is greater than zero. Consequently, whenever the natural bounds differ from the Balke-Pearl bounds, then under (A5) and (A12), there is evidence in the data for the existence of defiers (Pearl 2000). Huber, Laffers, and Mellace (2015) showed this is also true under (A5) and an exchangeability assumption equivalent to

$$Z \perp\!\!\!\perp \{Y^{x=0}, X^{z=0}, X^{z=1}\}; Z \perp\!\!\!\perp \{Y^{x=1}, X^{z=0}, X^{z=1}\}, \quad (14)$$

which is weaker than (A12) but stronger than (13).

3. IV Inequalities

The exclusion and exchangeability assumptions that define the IV model are not empirically verifiable. However, it is sometimes possible to falsify these assumptions, that is, to find empirical evidence against them. Balke and Pearl (1997) showed that the most restrictive model defined by (A5) and (A12) implies all of the following inequalities:

$$\begin{aligned}
&\Pr[Y = 0, X = 0|Z = 0] + \Pr[Y = 1, X = 0|Z = 1] \leq 1;\\
&\Pr[Y = 0, X = 1|Z = 0] + \Pr[Y = 1, X = 1|Z = 1] \leq 1;\\
&\Pr[Y = 1, X = 0|Z = 0] + \Pr[Y = 0, X = 0|Z = 1] \leq 1;\\
&\Pr[Y = 1, X = 1|Z = 0] + \Pr[Y = 0, X = 1|Z = 1] \leq 1.
\end{aligned} \quad (15)$$

Conversely, Bonet (2001) showed that any observable distribution that satisfies (15) is compatible with the assumptions (A5) and (A12).

It follows from Richardson and Robins (2010) that these inequalities are also implied by condition (4) alone, that is, the least restrictive model we have considered! To see this, consider the following argument. For $i, j, k \in \{0, 1\}$,

$$\begin{aligned}
\Pr[Y^{x=i} = j] &= \Pr[Y^{x=i} = j \mid Z = k]\\
&= \Pr[Y^{x=i} = j, X = i \mid Z = k]\\
&\quad + \Pr[Y^{x=i} = j, X = 1 - i \mid Z = k]\\
&= \Pr[Y = j, X = i \mid Z = k]\\
&\quad + \Pr[Y^{x=i} = j, X = 1 - i \mid Z = k]\\
&\leq \Pr[Y = j, X = i \mid Z = k]\\
&\quad + \Pr[X = 1 - i \mid Z = k]\\
&= 1 - \Pr[Y = 1 - j, X = i \mid Z = k], ;
\end{aligned} \quad (16)$$

where the first equality follows from (4) and the third from consistency. It follows that

$$\begin{aligned}
&\max_k \Pr[Y = 1, X = i \mid Z = k]\\
&\quad \leq \Pr[Y^{x=i} = 1] \leq \min_{k^*} 1 - \Pr[Y = 0, X = i \mid Z = k^*],
\end{aligned} \quad (17)$$

where the lower bound is obtained from (16) taking $j = 0$. The requirement that the lower bound be less than the upper bound (where $k \neq k^*$) then directly implies (15).

Since (A5) and (A12) imply (4), it follows that any observable distribution is compatible with the most restrictive model (A5) and (A12) if and only if it is also compatible with the least restrictive (4). This is surprising since the Balke-Pearl bounds for the ATE implied by (A5) and (A12) are narrower than the ATE bounds implied by (4) whenever at least one of the inequalities (6) fails to hold. (This statement is not vacuous because the set of distributions obeying (6) is a strict subset of those satisfying (15).)

Many authors have considered the power of these tests and the interpretation of specific violations. Richardson and Robins (2010) noted that any distribution can violate at most one of these four inequalities (15). In addition, they are invariant under relabeling of any variable. Cai et al. (2008) gave a simple interpretation of the inequalities in terms of bounds on average controlled direct effects in the counterfactual model assuming (A12). Specifically, they showed that if either of the IV inequalities associated with a given level of $X = x$ is violated, then under (A12) one can conclude that there is a nonzero population controlled direct effect of Z on Y fixing X to x, and further the sign may be determined. In fact, this conclusion follows directly from our weakest exchangeability assumption (A2) by an argument similar to that following (16).

Furthermore, returning to the model based on our strongest exclusion (A5) and exchangeability (A12) assumptions, the violation of either of the inequalities involving $X=1$ (or analogously, $X=0$), may be viewed as the presence of a direct effect for always-takers (or analogously, never-takers); see Zhang and Rubin (2003), Hudgens, Hoering, and Self (2003), and Imai (2008).

Finally, some authors have considered the testable implications of combining an IV model with additional assumptions, including some assumptions we review in Section 5. For example, see Huber and Mellace (2015), Mourifie and Wan (2014), and Kitagawa (2015).

4. Graphical Representations

In many disciplines such as epidemiology and computer science, the IV model is nearly exclusively represented using graphs. In the prior section, we showed several different IV models defined by variants of the exclusion and exchangeability assumptions. Here, we will show how many of these causal models may be associated with graphs by different semantics. We begin with a brief introduction to graphical representations; interested readers unfamiliar with graphical causal models may consider consulting additional resources (Spirtes, Glymour, and Scheines 1993; Greenland, Pearl, and Robins 1999; Pearl 2000; Richardson and Robins 2013).

A *causal* directed acyclic graph (DAG) is a DAG in which (i) the absence of an arrow from node A to B can be interpreted as the absence of a direct causal effect of A on B (relative to the other variables on the graph), (ii) all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph, and (iii) the Causal Markov Assumption (CMA) holds. The CMA links the causal structure represented by the DAG to the statistical data obtained in a study. It states that the distribution of the (factual) variables on the graph factor according to the DAG if the joint density is the product of the conditional

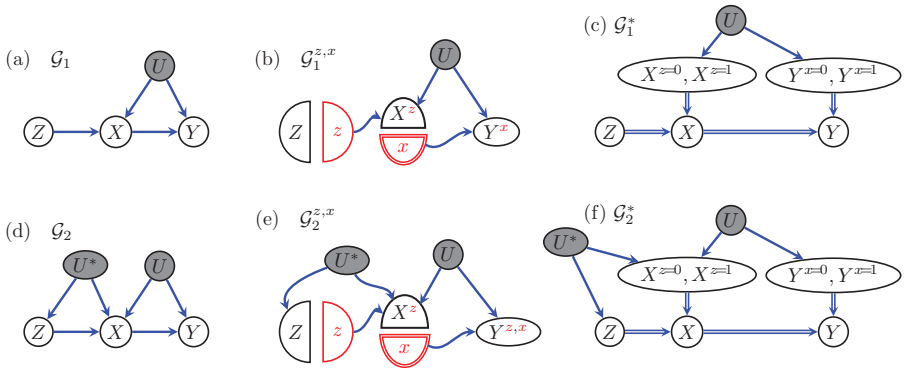

Figure 2. Graphical representations of IV models discussed in Section 4. The setting with no confounding between Z and X is considered in (a), (b), and (c); (d), (e), and (f) concern the setting with confounding between Z and X. Double edges ($\Rightarrow$) indicate deterministic relationships in (c) and (f).

densities for each variable given its "parents" in the DAG. This factorization is equivalent to the statement that each variable is independent of its nondescendants given its parents, where variable B is a descendant of variable A if there is a sequence of directed paths from A to B. For a causal DAG, this last statement can be reformulated as, given its parents, any variable is independent of all variables for which it is not a (direct or indirect) cause. These defining independencies logically imply additional independencies that can be read off the graph. The graphical method for determining these additional independencies is known as d-separation (Pearl 2000).

We now turn to the graphs in Figure 2. We note that the factorization associated with the DAG $\mathcal{G}_1$ is

$$\begin{aligned}&\Pr[Z=z, U=u, X=x, Y=y]\\&\quad=\Pr[Z=z]\Pr[U=u]\Pr[X=x\mid u, z]\Pr[Y=y\mid x, u].\end{aligned}\tag{18}$$

The factorization (18) directly implies the independencies (A9) and (A10).

Spirtes, Glymour, and Scheines (1993) showed that the CMA implies the distribution resulting from a causal intervention that fixes or sets a given variable to a specific value is obtained by simply removing the term from the factorization corresponding to the variable that has been intervened on. Thus, in the case of DAG $\mathcal{G}_1$ in Figure 2, for the intervention fixing X to x the distribution after intervention $\Pr^{int,x}[z, u, y]$ is given by

$$\Pr^{int,x}[z, u, y] = \Pr[Z=z]\Pr[U=u]\Pr[Y=y\mid x, u].\tag{19}$$

This Equation (19) directly implies (A11) by integrating out Z and U. The right-hand side of (19) is a particular instance of the g-formula (Robins 1986).

Therefore, we have seen that the latent noncounterfactual IV model of Section 2 that is defined by (A9), (A10), and (A11) is directly encoded in $\mathcal{G}_1$ via this interpretation. Thus, the causal DAG $\mathcal{G}_1$ would be appropriate if we were to suppose that (i) U represents all unmeasured common causes of X and Y, and further (ii) Z has been randomized, and hence is not confounded with X, Y, or U.

Causal graphs that directly incorporate counterfactual variables can also be used to represent counterfactual causal models. The two most widely considered such models are the Finest Fully Randomized Causally Interpreted Structural Tree Graph (FFRCISTG) and the nonparametric structural equation model with independent errors (NPSEM-IE) (Robins 1986; Pearl 2000). The NPSEM-IE is a strict submodel of the FFRCISTG model. The counterfactual independencies implied by an FFRCISTG model are sufficient to identify the effects of any intervention when all the variables on the graph are observed. The NPSEM-IE model can further identify counterfactual estimands that do not correspond to any intervention, such as the pure (or natural) direct effect.

The counterfactual independencies defining the FFRCISTG model can be encoded graphically by single-world intervention graphs (SWIGs) introduced in Richardson and Robins (2013). The nodes on a SWIG represent the counterfactual random variables corresponding to a single specific hypothetical intervention on a subset of the variables in the graph (Robins and Richardson 2011).

The graph $\mathcal{G}_1^{z,x}$ depicted in Figure 2(b) is a SWIG that represents the causal structure in graph $\mathcal{G}_1$ in a counterfactual world where Z has been set to z and X has been set to x. It is constructed from the original DAG $\mathcal{G}_1$ by the following three steps: (i) split all nodes that are being intervened upon into a random piece and a fixed piece, (ii) the random piece inherits all incoming edges on the original graph and the fixed piece inherits all out-going edges, and (iii) replace nodes that are descendants of the fixed portion with counterfactual nodes associated with this intervention. The random half of a split node represents the random variable that would be observed if that node had not been intervened on. Richardson and Robins (2013) showed that under the (naturally associated) FFRCISTG model, the distribution of the counterfactual random variables on the SWIG factors according to the graph. Consequently, since Z is not the parent of any variable on the SWIG $\mathcal{G}_1^{z,x}$, we immediately obtain the counterfactual independence (13). In addition, because the last node is Y^x and not $Y^{z,x}$, the SWIG encodes the individual-level exclusion assumption (A5).

The counterfactual DAG $\mathcal{G}_1^*$ in Figure 2(c) encodes additional independencies implied by the NPSEM-IE model that are not implied by the FFRCISTG model. In particular, the graph implies the independencies (12) used by Balke and Pearl (1997) because Z is d-separated from all the counterfactuals on the graph. Graph $\mathcal{G}_1^*$ is not a SWIG because (12) is not implied by the FFRCISTG model.

The graph $\mathcal{G}_1$ depicted in Figure 2(a) is often provided as "the" canonical IV graph, sometimes with the implication that this is the only situation where IV techniques may be applied.

However, this is inaccurate. To see this, consider the graph $\mathcal{G}_2$ that, unlike the simple graph in $\mathcal{G}_1$, includes confounding between the instrument Z and treatment X (Figure 2(d)). Like $\mathcal{G}_1$, the factorization of conditional densities represented by $\mathcal{G}_2$ implies the latent noncounterfactual model of Section 2 given by (A9), (A10), and (A11). (Dawid (2003) did not consider $\mathcal{G}_2$, but it implies all the assumptions (A9), (A10), and (A11) needed for his analysis.)

Consider next the graph for the SWIG $\mathcal{G}_2^{z,x}$ depicted in Figure 2(e). This graph is a population-level SWIG because the variable Y is indexed by both z and x. The absence of the edge from Z to Y encodes the population-level exclusion (A7) without imposing the individual-level exclusion (A5). Furthermore, by d-separation, the graph implies the constraints (A8) and (A9). Thus, it implies the latent counterfactual IV model described in Section 2.

Finally, the graph $\mathcal{G}_2^*$ in Figure 2(f) is the natural extension of $\mathcal{G}_1^*$ in Figure 2(c) and thus is not implied by the FFRCISTG model and therefore is not a SWIG. On this graph, Z is no longer independent of the counterfactuals X^z, and therefore does not imply the exchangeability assumptions discussed in Section 2 involving X^z (such as (12)). However, the graph does imply (5) because Z is d-separated from the node containing the counterfactuals $Y^{x=1}$, $Y^{x=0}$.

In summary, we have seen that all of the different counterfactual formulations of the IV model that lead to the Balke-Pearl bounds can be expressed graphically, thus unifying the graphical and counterfactual approaches.

5. Bounds on the Population Average Treatment Effect (ATE) Combining an Instrument with Further Assumptions

Now that we have thoroughly discussed the various versions of the IV assumptions, we turn to partial and point identification results when combining an instrument with additional assumptions. As with the bounds discussed in Section 2, the gains in identification for the bounds for the ATE discussed in this section are presented in Table 1, while expressions are presented in Tables 2 and 3.

5.1. Further Assumptions Requiring Counterfactual Treatments X^z

As discussed in Section 2, some IV models require the existence of the counterfactual treatment X^z. As noted earlier, with X^z defined, we may characterize subjects by one of four mutually exclusive compliance types: always-takers ($X^{z=0} = X^{z=1} = 1$); never-takers ($X^{z=0} = X^{z=1} = 0$); compliers ($X^{z=0} = 0, X^{z=1} = 1$); and defiers ($X^{z=0} = 1, X^{z=1} = 0$).

We can then consider further assumptions about the distribution of compliance types and effects within compliance types. For example, Richardson and Robins (2010) considered the geometry of the IV model under individual-level exclusion (A5), full exchangeability (A12), and the assumption that the proportion of defiers is known, that is,

$$\Pr[X^{z=0} = 1, X^{z=1} = 0] \text{ is known.} \tag{A13}$$

Note that the set of possible proportions of defiers is restricted by the assumptions (A5) and (A12) in conjunction with the observed joint distribution of (Y, X, Z). Interestingly, the full joint data imply restrictions beyond those implied by the marginal data on (X, Z). However, once given the proportion of defiers, the proportion of the other three compliance types is determined solely by the marginal distribution (X, Z). See Richardson and Robins (2010) and our Appendix S3 for details.

Assumption (A13) is of interest because, in the special case when it is assumed that there are no defiers, the effect in the compliers (a.k.a., the local average treatment effect [LATE]) is identified by the usual IV estimand

$$\frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[X|Z=1] - \mathrm{E}[X|Z=0]}. \tag{20}$$

This result was described in seminal work by Imbens and Angrist (1994), Baker and Lindeman (1994), and Angrist, Imbens, and Rubin (1996). It does not require (A12). In fact, for the usual IV estimand to equal the LATE it suffices that Z is independent of $(X^{z=0}, X^{z=1})$ and assumption (4) both hold; recall that (4) is our weakest outcome independence assumption. These assumptions are weaker than those of Huber, Laffers, and Mellace (2015) who also showed that (A12) could be relaxed. Richardson and Robins (2010) generalized the results from Angrist, Imbens, and Rubin (1996) by giving bounds for all four compliance types and for the entire population as a function of (A13). Such results can be used as a sensitivity analysis, as subject matter experts are often willing to give bounds on the proportion of defiers in their study population. If the proportion of defiers specified is nonzero, the effect in the compliers becomes only partially identified. Angrist, Imbens, and Rubin (1996) made this latter point but did not consider the restrictions placed on the proportion of defiers implied by the joint distribution of (Y, X, Z) (Richardson and Robins 2010). Huber, Laffers, and Mellace (2015) also studied bounds for compliance types under weaker assumptions.

5.2. Further Assumptions Restricting the Heterogeneity of the Effect of X on Y

The observed data $P(Y, X, Z)$ and assumptions (A5), (A12), and (A13) provide no information, by definition, about the mean counterfactual outcome in the never-takers had they been forced to take treatment nor in the always-takers had they been forced to forgo treatment. A corollary of this observation is that we may tighten the bounds on the ATE by assuming bounds on:

$$\mathrm{E}[Y^{x=0} \,|\, X^{z=0} = X^{z=1} = 1] \;\text{ or }\; \mathrm{E}[Y^{x=1} \,|\, X^{z=0} = X^{z=1} = 0]. \tag{A14}$$

For instance, when an outcome is rare, it is unlikely that the stratum of always-takers would universally experience the outcome had they not been treated; rather we might consider assuming that at most a certain proportion would experience the outcome.

Some authors have made specific proposals for how to use the observed data to inform specific versions of the assumption (A14). For example, an approach described by Baiocchi, Cheng, and Small (2014) corresponds to bounding the differences in treatment effects between compliance types under the special case of no defiers. A condition proposed by Siddique (2013) corresponds to, under the assumption of no defiers, bounding the mean counterfactual outcome under treatment among the never-takers by $\Pr[Y = 1|Z = 1, X = 1]$ and the mean counterfactual outcome under no treatment among the always-takers by

$\Pr[Y = 1|Z = 0, X = 0]$. By definition, a strategy for imposing limits for (A14) based on the observed data and/or prior knowledge would be subject-matter-dependent.

Another assumption that has been used to limit possible heterogeneity of the effects X can have on Y is to assume a monotonic relationship that specifies nobody is hurt by treatment:

$$Y^{x=1} \leq Y^{x=0} \tag{A15}$$

for all individuals (Manski 1997; Manski and Pepper 2000). It then follows that the upper bound is necessarily nonpositive and the lower bound is the same as that identified under the IV assumptions alone. Note the direction of inequality described in assumption (A15) could be flipped depending on the study setting. Related bounds can be found by assuming a monotonic relationship between X and Y without specifying the direction of the effect but further assuming a monotonic relationship between Z and X (Bhattacharya, Shaikh, and Vytlacil 2008).

Another assumption that limits treatment heterogeneity, described by Siddique (2013), specifically imposes restrictions on the counterfactual outcomes among those for whom $Z \neq X$. Specifically, she assumes that among those who decided not to take their assigned treatment, this decision was, on average, correct. That is, the outcome would be minimized under the observed treatment relative to the "compliant" unobserved treatment level:

$$\begin{aligned} &\mathrm{E}[Y^{x=1}|Z = 1, X = 0] - \mathrm{E}[Y^{x=0}|Z = 1, X = 0] \geq 0, \\ &\mathrm{E}[Y^{x=1}|Z = 0, X = 1] - \mathrm{E}[Y^{x=0}|Z = 0, X = 1] \leq 0. \end{aligned} \tag{A16}$$

When combined with the IV assumptions, (A16) can lead to improved lower bounds for each of the mean counterfactual outcomes, $\mathrm{E}[Y^{x=1}]$ and $\mathrm{E}[Y^{x=0}]$; however, the specific gains in identifying the ATE have not been described previously. As with (A15), note the direction of the relationship described in (A16) could be flipped depending on the study setting. Assumption (A16) is related to the "mean dominance" assumptions sometimes proposed in the econometrics literature (Huber, Laffers, and Mellace 2015; Huber and Mellace 2015).

Even stronger assumptions limiting the effect heterogeneity lead to point identification. Assuming additive effect homogeneity across levels of Z in the treated and the untreated,

$$\begin{aligned} &\mathrm{E}[Y^{x=1}|X = 1, Z = 1] - \mathrm{E}[Y^{x=0}|X = 1, Z = 1] \\ &\quad = \mathrm{E}[Y^{x=1}|X = 1, Z = 0] - \mathrm{E}[Y^{x=0}|X = 1, Z = 0], \\ &\mathrm{E}[Y^{x=1}|X = 0, Z = 1] - \mathrm{E}[Y^{x=0}|X = 0, Z = 1] \\ &\quad = \mathrm{E}[Y^{x=1}|X = 0, Z = 0] - \mathrm{E}[Y^{x=0}|X = 0, Z = 0] \end{aligned} \tag{A17}$$

identifies the ATE. The first equality identifies the effect of treatment on the treated, and the second equality identifies the effect on the untreated, both by the standard IV estimand (20). The first equality was given as an identifying assumption under an additive structural mean model (Robins 1989, 1994). Assuming effect homogeneity on the multiplicative rather than additive scale

$$\begin{aligned} \frac{\mathrm{E}[Y^{x=1}|X = 1, Z = 1]}{\mathrm{E}[Y^{x=0}|X = 1, Z = 1]} &= \frac{\mathrm{E}[Y^{x=1}|X = 1, Z = 0]}{\mathrm{E}[Y^{x=0}|X = 1, Z = 0]}, \\ \frac{\mathrm{E}[Y^{x=1}|X = 0, Z = 1]}{\mathrm{E}[Y^{x=0}|X = 0, Z = 1]} &= \frac{\mathrm{E}[Y^{x=1}|X = 0, Z = 0]}{\mathrm{E}[Y^{x=0}|X = 0, Z = 0]} \end{aligned} \tag{A18}$$

also results in point identification for the ATE. The identifying formula under additive versus multiplicative effect homogeneity assumptions, however, differs whenever the effect is nonnull (Robins 1989, 1994; Hernán and Robins 2006). In other words, except when the effect is null, it is impossible for both additive (A17) and multiplicative (A18) effect homogeneity to hold.

5.3. Further Assumptions Regarding Unmeasured Covariates

Recently, Wang and Tchetgen (2016) have provided new identifying assumptions for the ATE under the latent counterfactual IV model. Specifically, they showed that if, in addition to (A7), (A8), and (A9), either

$$\begin{aligned} &\mathrm{E}[X|Z = 1, X, U] - \mathrm{E}[X|Z = 0, X, U] \\ &\quad = \mathrm{E}[X|Z = 1, X] - \mathrm{E}[X|Z = 0, X] \end{aligned} \tag{A19}$$

or

$$\mathrm{E}[Y^{x=1} - Y^{x=0}|X, U] = \mathrm{E}[Y^{x=1} - Y^{x=0}|X] \tag{A20}$$

holds, then the ATE is identified by the usual IV estimand (20).

Other researchers have considered bounds, without point identification, in specific settings for which there is some subject matter knowledge about a specific unmeasured covariate, U, in combination with the IV assumptions. Such settings typically require a number of further parametric assumptions concerning the state space of U and the relationship between U and either the treatment X, the outcome Y, or both. For example, Chesher (2010) derived bounds that rely on assumptions about a single scalar U. Manski and Pepper (2000) described an assumption about how a specific dichotomous U informs treatment X; see also Siddique (2013), who further considered the same assumption in conjunction with (A16).

The bounds for the ATE under any of these additional assumptions concerning U often lead to markedly narrower bounds than the IV assumptions alone. However, the strong and/or specific assumptions about the unmeasured U are only substantively justified in limited domains (e.g., econometric models that imply the existence of a scalar U). For this reason, we do not compute bounds under these assumptions in our empirical example in Section 7.

5.4. Proposed Relaxations of IV Assumptions

In Section 2, we reviewed IV models defined by combinations of exclusion and exchangeability assumptions, but a natural question may be what happens if we relax (rather than add to) one or both of these types of assumptions. For example, Kaufman, Kaufman, and MacLehose (2009) noted that assuming only (A5) or (A12), but not both together, provides no improvement over the bounds obtained under the data alone. Other authors have considered "imperfect instruments" (Manski and Pepper 2000; Ramsahai 2012; Flores and Flores-Lagunes 2013; Huber 2014). As one example, Ramsahai (2012) considered an IV model that replaced (A10) with

$$\begin{aligned} 0 \leq |\Pr[Y = 1|Z = 1, X = x, U] \\ - \Pr[Y = 1|Z = 0, X = x, U]| \leq \epsilon, \end{aligned} \tag{21}$$

where $\epsilon = 0$ would reduce to (A10), $\epsilon = 1$ would place no restriction, and $0 < \epsilon < 1$ would represent a weakened exclusion restriction. Such relaxations of course lead to wider bounds than those derived under the IV assumptions, but can serve as a sensitivity analysis between having an instrument and having no instrument at all.

6. Extensions to Other Study Settings

6.1. Continuous Outcomes

When Y is a continuous variable, partial identification of the ATE (i.e., the difference in means) under the IV assumptions requires specification both of an upper bound that exceeds the maximum of the supports for $Y^{x=1}$ and $Y^{x=0}$ and a lower bound that is less than the minimum of the supports. As one example, given such upper and lower bounds for the support of Y, the bounds that follow from marginal exchangeability (4) are identified for continuous outcomes by assuming mean exchangeability (Manski 1990; Robins 1994; Manski and Pepper 2000; Hernán and Robins 2006):

$$\mathrm{E}[Y^x \mid Z = 1] = \mathrm{E}[Y^x \mid Z = 0] \text{ for all } x. \tag{22}$$

Though implicit for dichotomous outcomes (which are bounded by 0 and 1), bounding the support of Y^x is an additional assumption needed when outcomes are continuous. For many continuous outcomes this may be plausible: for example, cholesterol levels cannot be negative nor can they approach infinity. However, the limits on Y^x may not be known. For example, it is physically impossible for cholesterol levels to be below 0 mg/dl or above 105,200 mg/dl given the density of cholesterol, implying Y^x must be bounded between 0 and 105,200 mg/dl. One may further argue that the cholesterol levels are less than a certain threshold (e.g., 1100 mg/dl) based on extreme hypercholesterolemia case studies (Sprecher et al. 1984). For a specific study population, experts may argue the range of plausible cholesterol levels is narrower still. In practice, the choice of the bounds on Y^x can greatly affect the width of the bounds on the ATE.

For continuous outcomes, other assumptions that have been used to construct bounds for the ATE found in the literature include: relaxing mean exchangeability to instead assume $\mathrm{E}[Y^x|Z = 1] \geq \mathrm{E}[Y^x|Z = 0]$ (Manski and Pepper 2000) and nonparametric selection models (Heckman and Vytlacil 2001). In the face of continuous Y, a number of authors have made their object of inference contrasts between the quantiles of $Y^{x=1}$ and $Y^{x=0}$ rather than between the means; see Chernozhukov and Hansen (2005) and Blanco, Flores, and Flores-Lagunes (2013).

In Section 2, we observed that a range of exchangeability assumptions leads to the same bounds. In particular, the SWIG exchangeability (13) and full randomization (A12) assumptions both lead to the Balke-Pearl bounds. This phenomenon may be specific to the case where Y is binary; Huber, Laffers, and Mellace (2015) showed that, in the case where Y is continuous, they obtain wider bounds under (A5) and (14) than those obtained by Kitagawa (2009) who assumed (A5) and (A12).

6.2. Nonbinary Instruments

The IV model can be extended to settings with continuous or categorical instruments. Beresteanu, Molchanov, and Molinari (2012) described identification regions in the general case under (4) and the natural extension of (5). Ramsahai (2012) and Richardson and Robins (2014) described bounds on the ATE under the IV assumptions for a categorical instrument with an arbitrary (finite) number of categories with binary treatment and outcome. Palmer et al. (2011) provided software for implementing bounds using three-level instruments.

The instrumental inequalities discussed in Section 3 can likewise be generalized for nonbinary instruments and outcomes. When the instrument Z and outcome Y are categorical, Pearl (1995) showed the IV model satisfied:

$$\max_x \sum_y \max_z \Pr[y, x|z] \leq 1. \tag{23}$$

(When Z or Y is continuous, this can be reexpressed with respect to the conditional density function of Y given X, Z.) When Z has more than two levels, Bonet (2001) showed that the IV model defined by the individual-level exclusion restriction (A5) and the full exchangeability assumption (A12) implies additional constraints on the joint distribution of the observed data beyond (23) or the more general form. Results in Richardson and Robins (2014) imply that the constraints on the observed data distribution under the model given by joint exchangeability (5) are the same as those implied by the more restrictive model (A5) plus (A12). In Appendix S4, the additional constraints of Bonet are given for the special case $Z = 3$; furthermore, an observed data distribution satisfying Pearl's constraints (23) but not Bonet's additional constraint is displayed. An argument similar to that of Equation (16) shows that Pearl's constraints (23) hold under the weakest IV model (4); however, this is not the case for Bonet's additional constraint, which need not hold even under the model defined by individual-level exclusion (A5) and marginal exchangeability (4)—rather joint exchangeability (5) is required. It follows that the surprising finding that the least and most restrictive IV models are associated with the same instrumental inequalities (discussed in Section 3) is specific to the binary instrument setting.

6.3. Other Study Designs

Thus far we have considered observed data distributions that may have been generated in one of two study designs: randomized trials and observational follow-up studies. We may also consider the so-called "two-sample" study design, where we obtain information on the distribution of (Z, X) from one sample and (Z, Y) from a second sample. Intuitively, such a design has less information than one that observes the full joint distribution of (Z, X, Y). Bounds and point identification results have been derived under IV assumptions in these settings: see Palmer et al. (2011) for data analysis and implementation, and see Ramsahai (2012) for further discussion. Note an implicit assumption in "two-sample" study designs is that the two samples are both random samples from the same source population.

IV analyses in case–control studies have also been considered, and indeed bounds under IV assumptions have been derived if the marginal distribution $\Pr[Y = 1]$ of the binary outcome in the source population is known. If this probability is not known, one may consider assuming upper and lower bounds on $\Pr[Y = 1]$ to obtain bounds for the ATE (Didelez and Sheehan 2007; Palmer et al. 2011).

6.4. Incorporating Measured Covariates

Particularly in observational studies, the above-described assumptions may often be unlikely to hold unconditionally, but perhaps would seem more palatable within levels of measured covariates occurring prior to Z. We can relax any of these sets of assumptions by first bounding effects within strata, and then using standardization techniques to partially identify the ATE. For example, suppose we assumed that some set of assumptions held within levels of a measured categorical covariate, L. This implies that we can estimate lower and upper bounds, LB_l and UB_l, for the treatment effect within any level of $L = l$:

$$LB_l \leq \mathrm{E}[Y^{x=1}|L = l] - \mathrm{E}[Y^{x=0}|L = l] \leq UB_l. \tag{24}$$

It follows that bounds for the ATE can be derived by standardizing these bounds:

$$\sum_{L=l} LB_l \Pr[L = l] \leq \mathrm{E}[Y^{x=1}] - \mathrm{E}[Y^{x=0}] \leq \sum_{L=l} UB_l \Pr[L = l]. \tag{25}$$

See Swanson et al. (2015) for an applied example. For an approach to modeling stratum-specific ATE bounds as a function of preinstrument covariates (including potentially continuous covariates); see Richardson, Evans, and Robins (2010).

Rather than incorporating preinstrument covariates, another strategy in some studies may be to incorporate information on auxiliary outcomes. For example, Mealli and Pacini (2013) considered a setting for which the IV conditions were satisfied for a secondary outcome, and developed bounds for the intention-to-treat effect within compliance types for a primary outcome.

7. Empirical Example

7.1. Study Setting

To demonstrate the reviewed bounding approaches in one empirical example, we used publicly available data from the Oregon Health Insurance Experiment (Finkelstein 2013). Details of the study have been provided elsewhere (Taubman et al. 2014). In brief, Oregon initiated an expansion of the Medicaid program in 2008, extending benefits to include uninsured, low-income, able-bodied adults who would not have previously qualified for Medicaid coverage. This expansion was done by drawing names from a waiting list lottery, thus offering an opportunity to study the effects of healthcare coverage in a randomized design. Taubman et al. (2014) analyzed the effects of Medicaid coverage on emergency department visits during the 18-month follow-up period using IV methods, and generally concluded that Medicaid coverage increased emergency department use over this study period. Their study primarily focused on point estimates for the LATE, that is, the average treatment effect in persons who would have received Medicaid coverage had they won the lottery draw but not otherwise.

Here, we estimate bounds for the ATE, that is, the effect in the entire study population. The ATE is arguably more relevant for policy questions (Robins and Greenland 1996)—in particular for questions about the possible effects of universal healthcare coverage (Kreider and Hill 2009)—but typically requires stronger assumptions for point identification. We estimate bounds for the ATE of Medicaid coverage on (i) any emergency department visit and (ii) any emergency department visit for chest pain or a heart condition. These outcomes were chosen to demonstrate results for a common and a rare dichotomous outcome, respectively. We used the data made publicly available by Taubman et al. (2014) with complete information on these outcomes, and further restricted analyses to single-person households, leaving an analytic sample of $N = 18{,}854$. For our primary analyses, we considered as our treatment variable Medicaid coverage defined as any enrollment in Medicaid during the study period. As a secondary treatment definition, we considered Oregon Health Program (OHP) Standard coverage, defined as enrollment in the lotteried healthcare program (OHP Standard) during the study period. Using these two treatment definitions allow for comparison of the flexibility of bounds derived for feasible distributions of compliance types.

Table 4. Distribution of randomization, Medicaid/OHP coverage, and outcomes.

N	Randomization Z	Coverage X	Any visit $\mathrm{E}[Y\|X, Z]$	Heart visit $\mathrm{E}[Y\|X, Z]$
Medicaid				
10,594	0	0	0.342	0.025
1819	0	1	0.330	0.031
3810	1	0	0.339	0.028
2631	1	1	0.372	0.023
OHP				
12,094	0	0	0.341	0.026
319	0	1	0.320	0.025
4464	1	0	0.345	0.029
1977	1	1	0.369	0.019

The distribution of these variables can be found in Table 4. Note that, for both treatment definitions, there is noncompliance with lottery assignment in both levels of lottery assignment. Persons who were selected in the lottery may not have obtained coverage for a number of reasons, including simply failing to pursue the application process. Persons who were not selected in the lottery could have obtained coverage if they qualified for coverage through other means, for example, because of changes in their income or disability status over the study period.

7.2. Bounds

Estimates of the bounds for the ATE of Medicaid and OHP coverage on the risk of the two dichotomous outcomes are presented in Table 5. As expected, bounds of unit length are obtained using the data only. Narrower bounds are achieved under an IV model, with the natural and Balke-Pearl bounds being identical in these data: for example, $[-0.287, 0.452]$ for the effect of Medicaid coverage on any emergency department visit. Because of randomization, we might expect all versions of exchangeability to hold. A justification for the individual-level exclusion restriction (A5) was provided by the authors (Taubman et al. 2014). The assumption (A5) would not hold if a participant's lottery draw encouraged some patients to adopt other healthy behaviors or seek other public services. Moreover, even if the exclusion restriction was satisfied for the continuous, time-varying treatment of Medicaid coverage (e.g., number of months with coverage), our dichotomization of the treatment as "any" versus "none" can lead to violations of the condition (VanderWeele et al. 2014).

Table 5. Identification of the average treatment effect of Medicaid coverage on 18-month risk of emergency department visits under the sets of assumptions described in Figure 1.

Assumption set	Lower bound	Upper bound
Effect of Medicaid coverage on any visit		
Data only	−0.413	0.587
A1 + A2	−0.287	0.452
A3 + A4	−0.287	0.452
A1 + A2 + A15	−0.287	−0.012
A1 + A2 + A16	−0.086	0.403
A1 + A2 + A17	0.046	
A1 + A2 + A18	0.044	
Effect of Medicaid coverage on heart visit		
Data only	−0.250	0.750
A1 + A2	−0.159	0.579
A3 + A4	−0.159	0.579
A1 + A2 + A15	−0.159	−0.001
A1 + A2 + A16	−0.143	0.575
A1 + A2 + A17	−0.002	
A1 + A2 + A18	−0.003	
Effect of OHP coverage on any visit		
Data only	−0.378	0.622
A1 + A2	−0.245	0.474
A3 + A4	−0.245	0.474
A1 + A2 + A15	−0.245	−0.012
A1 + A2 + A16	−0.005	0.466
A1 + A2 + A17	0.043	
A1 + A2 + A18	0.044	
Effect of OHP coverage on heart visit		
Data only	−0.143	0.855
A1 + A2	−0.046	0.673
A3 + A4	−0.046	0.673
A1 + A2 + A15	−0.046	−0.001
A1 + A2 + A16	−0.026	0.673
A1 + A2 + A17	−0.002	
A1 + A2 + A18	−0.003	

Bounds under an additional assumption restricting the direction of individual-level effects (A15) led to identifying the direction of the ATE, as expected. Because it is plausible that coverage could cause visits for some people and protect against visits in other people, neither (A15) or its inverse would be reasonable in this study. Assuming (A16) improves the bounds relative to the bounds under the IV assumptions alone. However, (A16) is also unlikely to hold in the current study, particularly because it is implicit in this assumption that minimizing risk of an emergency department visit completely informs the enrollment decision.

We obtained similar point estimates with wide confidence intervals (not shown) under the additional assumptions of effect homogeneity on the additive (A17) and multiplicative (A18) scales. To justify the reasonableness of these additional assumptions, we might apply subject-matter knowledge to assess the sufficient condition of effect homogeneity by the unmeasured confounders described by Hernán and Robins (2006). In this study, age, preexisting conditions, and health literacy are all likely strong effect modifiers, thus effect homogeneity assumptions may not be appropriate.

For Medicaid coverage, a proportion of defiers between 0 and 14.7% was consistent with the observed data, individual-level exclusion restriction (A5), and randomization (A12); for OHP coverage, the feasible proportion of defiers was much more restrictive (0 to 2.6%). In this particular study, both of the upper bounds on the proportion of defiers are $\Pr[X = 1|Z = 0]$, and thus the small number of subjects who obtained OHP coverage despite not being selected in the lottery drives this narrow range.

The ATE and the effects within compliance types were estimated under the IV assumptions and various feasible compliance type distributions (Table 6). The estimated LATE was sensitive to the proportion of defiers assumed: if we assumed 1.3% or more of the study population were defiers, the direction of the effect of Medicaid coverage on any visits was no longer identified. In Table 7, we also estimate the ATE under further assumptions restricting the proportion of always-takers who would have had a heart-related emergency department visit had they not had coverage, and the proportion of never-takers who would have had a heart-related emergency department visit had they had coverage. Because only 2.6% of the study population had an emergency department visit for chest pain or a heart condition, it is perhaps plausible to assume that at most a certain proportion of the always-takers and never-takers under their unobserved counterfactual treatment level would have had such a visit. However, subject-matter experts may have different opinions on what is a reasonable upper bound. A similar strategy of assumed restrictions could also be applied to our more common outcome of any emergency department visit,

Table 6. Identification of the average treatment effect globally and within compliance types under assumptions (A5), (A12), and specified feasible versions of (A13).

Distribution [DE,CO,AT,NT]*	Defier	Complier	Always-taker	Never-taker	Global
Effect of Medicaid coverage on any visit					
[0.00, 0.26, 0.15, 0.59]	[−1.000, 1.000]	0.046	[−0.670, 0.330]	[−0.349, 0.661]	[−0.287, 0.452]
[0.05, 0.31, 0.10, 0.54]	[−1.000, 0.968]	[−0.122, 0.194]	[−1.000, 0.502]	[−0.370, 0.722]	[−0.287, 0.450]
[0.10, 0.36, 0.05, 0.49]	[−0.981, 0.484]	[−0.238, 0.167]	[−1.000, 1.000]	[−0.408, 0.796]	[−0.285, 0.400]
Effect of Medicaid coverage on heart visit					
[0.00, 0.26, 0.15, 0.59]	[−1.000, 1.000]	−0.002	[−0.969, 0.031]	[−0.028, 0.972]	[−0.159, 0.579]
[0.05, 0.31, 0.10, 0.54]	[−0.326, 0.090]	[−0.054, 0.012]	[−1.000, 0.467]	[−0.030, 1.000]	[−0.125, 0.534]
[0.10, 0.36, 0.05, 0.49]	[−0.163, 0.045]	[−0.047, 0.011]	[−1.000, 0.097]	[−0.033, 1.000]	[−0.075, 0.484]
Effect of OHP coverage on any visit					
[0.00, 0.28, 0.03, 0.69]	[−1.000, 1.000]	0.043	[−0.680, 0.320]	[−0.345, 0.655]	[−0.245, 0.474]
[0.01, 0.29, 0.02, 0.68]	[−1.000, 0.822]	[0.007, 0.070]	[−1.000, 0.523]	[−0.350, 0.664]	[−0.245, 0.472]
[0.02, 0.30, 0.01, 0.67]	[−0.874, 0.411]	[−0.018, 0.067]	[−1.000, 1.000]	[−0.355, 0.674]	[−0.242, 0.462]
Effect of OHP coverage on heart visit					
[0.00, 0.28, 0.03, 0.69]	[−1.000, 1.000]	−0.002	[−0.975, 0.025]	[−0.029, 0.971]	[−0.046, 0.673]
[0.01, 0.29, 0.02, 0.68]	[−1.000, 0.064]	[−0.037, −0.000]	[−1.000, 0.041]	[−0.029, 0.986]	[−0.046, 0.664]
[0.02, 0.30, 0.01, 0.67]	[−0.994, 0.032]	[−0.068, −0.000]	[−1.000, 0.113]	[−0.030, 1.000]	[−0.045, 0.654]

NOTES: *[DE,CO,AT,NT] denotes the proportion or defiers, compliers, always-takers, and never-takers under each specified proportion of defiers. For Medicaid coverage, we specified the proportion of defiers as 0%, 5%, and 10%. For OHP coverage, we specified the proportion of defiers as 0%, 1%, and 2%.

Table 7. Identification of the average treatment effect of Medicaid and OHP coverage on any heart visit under assumptions restricting the unobserved counterfactuals within compliance types, assuming no defiers (assumptions (A5)+(A12)+(A13)+(A14)).

Restriction on unobserved strata (A14)	Effect of Medicaid coverage	Effect of OHP coverage
No restriction	[−0.159, 0.579]	[−0.046, 0.673]
[0, 0.9]	[−0.144, 0.520]	[−0.043, 0.604]
[0, 0.8]	[−0.130, 0.461]	[−0.040, 0.535]
[0, 0.7]	[−0.115, 0.402]	[−0.038, 0.465]
[0, 0.6]	[−0.100, 0.342]	[−0.035, 0.396]
[0, 0.5]	[−0.086, 0.283]	[−0.033, 0.327]
[0, 0.4]	[−0.071, 0.224]	[−0.030, 0.257]
[0, 0.3]	[−0.056, 0.165]	[−0.028, 0.188]
[0, 0.2]	[−0.042, 0.106]	[−0.025, 0.119]
[0, 0.1]	[−0.027, 0.047]	[−0.022, 0.049]
[0, 0.05]	[−0.020, 0.017]	[−0.021, 0.015]
[0, 0.02]	[−0.015, −0.001]	[−0.020, −0.006]

however, because it is not rare it is less clear what subject-matter knowledge can be applied to impose reasonable restrictions.

8. Discussion

We reviewed and provided a taxonomy of methods for partial identification of the ATE under IV assumptions in a common notation. We laid out the gains and trade-offs involved when making increasingly strong assumptions, and presented an empirical example of estimating bounds in a common dataset. As such, readers can now readily compare all the proposed sets of assumptions and resulting bounds.

Focusing on partial identification has two key benefits. First, presenting bounds clarifies the role of unverifiable assumptions when deriving estimates for causal effects. This in turn makes clear how much any data-driven decision (e.g., policy, clinical, personal) relies on strong assumptions. Our empirical example illustrates that the IV assumptions alone resulted in wide bounds, and it is only when we restrict the possible effects of treatment on outcome that we identify the direction of the effect. Like others before (Robins and Greenland 1996), we suggest it is often useful to present bounds and point estimates under various sets of assumptions, particularly when using IV methods. Indeed, presenting bounds and point estimates under an array of assumptions clarifies that the scientific debate should focus on what assumptions we are most confident in, which in turn relates to the range of plausible effect sizes we are most confident in. For a recent application highlighting this benefit, see Manski and Pepper (2015).

Second, presenting bounds allows us to refocus on the ATE as the causal parameter of interest. Because the effect homogeneity assumptions necessary for point identification for the ATE are often unrealistic, the LATE has often been favored in IV applications (including in the primary publications associated with our empirical example). However, the LATE applies to a subset of the population that generally cannot be identified, and is not directly informative for decision-making. Further, our example shows that the LATE estimate can be sensitive to relatively small monotonicity assumption violations.

The development of partial identification methods has been transdisciplinary, with important advances on this topic made by statisticians, economists, epidemiologists, and computer scientists. Nevertheless, these methods are seldom applied outside of some applications by economists. It is time to place these partial identification methods into the standard arsenal for causal inference. If nothing else, estimating the bounds can serve as a reminder to remain humble about how much information the data really provide.

Supplementary Materials

In Appendix S1, we provide a brief overview of estimation procedures. In Appendix S2, we provide expressions for the bounds on the mean counterfactuals under the assumption sets described in Section 2. In Appendix S3, we provide expressions for the ATE, the ATE within compliance types, and the mean counterfactuals under assumption sets including (A13) and (A14). In Appendix S4, we provide a comparison of constraints implied by the IV model for a trichotomous instrument. The supplementary materials further include SAS code; any updates to this code following publication will be posted at *https://www.hsph.harvard.edu/causal/software/*.

Funding

This work is supported in part by grants from NIH (R01 AI102634; DP1 ES025459; R01 AI112339; R01 AI27271), ONR (N00014-15-1-2672), and NWO/ZonMW (Veni personal grant 91617066).

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. [940]

Baiocchi, M., Cheng, J., and Small, D. S. (2014), "Instrumental Variable Methods for Causal Inference," *Statistics in Medicine*, 33, 2297–2340. [940]

Baker, S. G., and Lindeman, K. S. (1994), "The Paired Availability Design: A Proposal for Evaluating Epidural Analgesia During Labor," *Statistics in Medicine*, 13, 2269–2278. [940]

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. [935,936,937,938,939]

Beresteanu, A., and Manski, C. F. (2000), "Bounds for Stata," available at *http://faculty.wcas.northwestern.edu/cfm754/bounds_stata.pdf*. [933]

Beresteanu, A., Molchanov, I., and Molinari, F. (2012), "Partial Identification Using Random Set Theory," *Journal of Econometrics*, 166, 17–32. [942]

Bhattacharya, J., Shaikh, A. M., and Vytlacil, E. (2008), "Treatment Effect Bounds Under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization," *The American Economic Review*, 98, 351–356. [941]

Blanco, G., Flores, C. A., and Flores-Lagunes, A. (2013), "Bounds on Average and Quantile Treatment Effects of Job Corps Training on Wages," *Journal of Human Resources*, 48, 659–701. [942]

Bonet, B. (2001), "Instrumentality Tests Revisited," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., pp. 48–55. [938,942]

Cai, Z., Kuroki, M., Pearl, J., and Tian, J. (2008), "Bounds on Direct Effects in the Presence of Confounded Intermediate Variables," *Biometrics*, 64, 695–701. [938]

Chernozhukov, V., and Hansen, C. (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. [942]

Chernozhukov, V., Kim, W., Lee, S., and Rosen, A. M. (2015), "Implementing Intersection Bounds in Stata," *Stata Journal*, 15, 21–44. [933]

Chesher, A. (2010), "Instrumental Variable Models for Discrete Outcomes," *Econometrica*, 78, 575–601. [941]

Davies, N. M., Smith, G. D., Windmeijer, F., and Martin, R. M. (2013), "Issues in the Reporting and Conduct of Instrumental Variable Studies: A Systematic Review," *Epidemiology*, 24, 363–369. [934]

Dawid, A. (2003), "Causal Inference Using Influence Diagrams: The Problem of Partial Compliance," in *Highly Structured Stochastic Systems*, eds. P. J. Green, N. L. Hjort, and S. Richardson, Oxford: Oxford University Press, pp. 45–81. [934,937,940]

Didelez, V., Meng, S., and Sheehan, N. A. (2010), "Assumptions of IV Methods for Observational Epidemiology," *Statistical Science*, 25, 22–40. [937]

Didelez, V., and Sheehan, N. (2007), "Mendelian Randomization as an Instrumental Variable Approach to Causal Inference," *Statistical Methods in Medical Research*, 16, 309–330. [942]

Finkelstein, A. (2013), "Oregon Health Insurance Experiment Public Use Data," available at *http://www.nber.org/oregon/data.html* [933,943]

Flores, C. A., and Flores-Lagunes, A. (2013), "Partial Identification of Local Average Treatment Effects With an Invalid Instrument," *Journal of Business & Economic Statistics*, 31, 534–545. [941]

Greenland, S., Pearl, J., and Robins, J. M. (1999), "Causal Diagrams for Epidemiologic Research," *Epidemiology*, 10, 37–48. [938]

Heckman, J. J., and Vytlacil, E. J. (2001), *Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect*, New York: Springer. [942]

Hernán, M. A., and Robins, J. M. (2006), "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology*, 17, 360–372. [Errata in *Epidemiology*, 2014 Jan; 15(1): 164]. [934,941,942,944]

Huber, M. (2014), "Sensitivity Checks for the Local Average Treatment Effect," *Economics Letters*, 123, 220–223. [941]

Huber, M., Laffers, L., and Mellace, G. (2015), "Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations Under Endogeneity and Noncompliance," *Journal of Applied Econometrics*, 32, 56–79. [938,940,941,942]

Huber, M., and Mellace, G. (2015), "Testing Instrument Validity for Late Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411. [938,941]

Hudgens, M. G., Hoering, A., and Self, S. G. (2003), "On the Analysis of Viral Load Endpoints in HIV Vaccine Trials," *Statistics in Medicine*, 22, 2281–2298. [938]

Imai, K. (2008), "Sharp Bounds on the Causal Effects in Randomized Experiments With 'Truncation-by-Death'," *Statistics & Probability Letters*, 78, 144–149. [938]

Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. [940]

Kaufman, S., Kaufman, J. S., and MacLehose, R. F. (2009), "Analytic Bounds on Causal Risk Differences in Directed Acyclic Graphs Involving Three Observed Binary Variables," *Journal of Statistical Planning and Inference*, 139, 3473–3487. [941]

Kitagawa, T. (2009), "Identification Region of the Potential Outcome Distributions Under Instrument Independence," Technical Report, working paper CWP30/09, London, UK: CEMMAP. [937,942]

——— (2015), "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063. [938]

Kreider, B., and Hill, S. C. (2009), "Partially Identifying Treatment Effects With an Application to Covering the Uninsured," *Journal of Human Resources*, 44, 409–449. [943]

Manski, C. F. (1990), "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. [934,935,942]

——— (1997), "Monotone Treatment Response," *Econometrica: Journal of the Econometric Society*, 65, 1311–1334. [941]

——— (2003), *Partial Identification of Probability Distributions*, New York: Springer Science & Business Media. [937]

Manski, C. F., and Pepper, J. V. (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. [941,942]

——— (2015), "How do Right-to-Carry Laws Affect Crime Rates? Coping With Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 100, 232–244. [945]

McCarthy, I., Millimet, D. L., and Roy, M. (2015), "Bounding Treatment Effects: Stata Command for the Partial Identification of the Average Treatment Effect With Endogenous and Misreported Treatment Assignment," *Stata Journal*, 15, 411–436. [933]

Mealli, F., and Pacini, B. (2013), "Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance," *Journal of the American Statistical Association*, 108, 1120–1131. [943]

Mourifie, I., and Wan, Y. (2014), "Testing Local Average Treatment Effect Assumptions," *Review of Economics and Statistics*, 99, 305–313. [938]

Palmer, T. M., Ramsahai, R. R., Didelez, V., and Sheehan, N. A. (2011), "Nonparametric Bounds for the Causal Effect in a Binary Instrumental-Variable Model," *Stata Journal*, 11, 345–367. [933,942]

Pearl, J. (1995), "On the Testability of Causal Models With Latent and Instrumental Variables," in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, Burlington: Morgan Kaufmann, pp. 435–443. [942]

——— (2000), *Causality: Models, Reasoning and Inference*(Vol. 29), Cambridge: Cambridge University Press. [Errata Available at: *http://bayes.cs.ucla.edu/BOOK-09/causality2-errata-updated7_3_13.pdf*]. [938,939]

Ramsahai, R. R. (2012), "Causal Bounds and Observable Constraints for Non-Deterministic Models," *The Journal of Machine Learning Research*, 13, 829–848. [941,942]

Richardson, T. S., Evans, R. J., and Robins, J. M. (2010), "Transparent Parametrizations of Models for Potential Outcomes," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press, pp. 569–610. [943]

Richardson, T. S., and Robins, J. M. (2010), "Analysis of the Binary Instrumental Variable Model," in *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, eds. R. Dechter, H. Geffner, and J. Y. Halpern, London: London College Publications, pp. 415–444. [938,940]

——— (2013), "Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality," Working Paper 128, Center for the Statistics and the Social Sciences, University of Washington Series. [938,939]

——— (2014), "ACE Bounds; SEMs With Equilibrium Conditions," *Statistical Science*, 29, 363–366. [936,937,942]

Robins, J. (1986), "A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods – Applications to Control of the Healthy Worker Survivor Effect," *Mathematical Modeling*, 7, 1393–1512. [939]

——— (1989), *The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies*, Washington, DC: NCHRS, US Public Health Service. [934,935,941]

——— (1994), "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics-Theory and Methods*, 23, 2379–2412. [941,942]

Robins, J. M., and Greenland, S. (1996), "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association*, 91, 456–458. [937,943,945]

Robins, J. M., and Richardson, T. S. (2011), "Alternative Graphical Causal Models and the Identification of Direct Effects," in *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, eds. P. Shrout, K. Keyes, and K. Ornstein, Oxford: Oxford University Press, pp. 1–52. [939]

Siddique, Z. (2013), "Partially Identified Treatment Effects Under Imperfect Compliance: The Case of Domestic Violence," *Journal of the American Statistical Association*, 108, 504–513. [940,941]

Spirtes, P., Glymour, C., and Scheines, R. (1993), *Causation, Prediction and Search*(*Lecture Notes in Statistics*, Vol. 81), New York: Springer-Verlag. [938,939]

Sprecher, D. L., Schaefer, E. J., Kent, K. M., Gregg, R. E., Zech, L. A., Hoeg, J. M., McManus, B., Roberts, W. C., and Brewer, H. B. (1984), "Cardiovascular Features of Homozygous Familial Hypercholesterolemia: Analysis of 16 Patients," *The American Journal of Cardiology*, 54, 20–30. [942]

Swanson, S. A., Holme, Ø., Løberg, M., Kalager, M., Bretthauer, M., Hoff, G., Aas, E., and Hernán, M. A. (2015), "Bounding the Per-Protocol Effect in Randomized Trials: An Application to Colorectal Cancer Screening," *Trials*, 16, 1–11. [943]

Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K., and Finkelstein, A. N. (2014), "Medicaid Increases Emergency-Department Use: Evidence From Oregon's Health Insurance Experiment," *Science*, 343, 263–268. [933,943]

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., and Kraft, P. (2014), "Methodological Challenges in Mendelian Randomization," *Epidemiology*, 25, 427–435. [943]

Wang, L., and Tchetgen, E. T. (2018), "Bounded, Efficient and Multiply Robust Estimation of Average Treatment Effects Using Instrumental Variables," arXiv preprint arXiv:1611.09925. [941]

Zhang, J. L., and Rubin, D. B. (2003), "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "death"," *Journal of Educational and Behavioral Statistics*, 28, 353–368. [938]