Instruments for Causal Inference

An Epidemiologist's Dream?

Miguel A. Hernán and James M. Robins*†*

Abstract: The use of instrumental variable (IV) methods is attractive because, even in the presence of unmeasured confounding, such methods may consistently estimate the average causal effect of an exposure on an outcome. However, for this consistent estimation to be achieved, several strong conditions must hold. We review the definition of an instrumental variable, describe the conditions required to obtain consistent estimates of causal effects, and explore their implications in the context of a recent application of the instrumental variables approach. We also present (1) a description of the connection between 4 causal models—counterfactuals, causal directed acyclic graphs, nonparametric structural equation models, and linear structural equation models—that have been used to describe instrumental variables methods; (2) a unified presentation of IV methods for the average causal effect in the study population through structural mean models; and (3) a discussion and new extensions of instrumental variables methods based on assumptions of monotonicity.

(*Epidemiology* 2006;17: 360–372)

Can you guarantee that the results from your observational study are unaffected by unmeasured confounding? The only answer an epidemiologist can provide is "no." Regardless of how immaculate the study design and how perfect the measurements, the unverifiable assumption of no unmeasured confounding of the exposure effect is necessary for causal inference from observational data, whether confounding adjustment is based on matching, stratification, regression, inverse probability weighting, or g-estimation.

Now, imagine for a moment the existence of an alternative method that allows one to make causal inferences from observational studies even if the confounders remain unmeasured. That method would be an epidemiologist's dream. Instrumental variable (IV) estimators, as reviewed by Martens et al[1] and applied by Brookhart et al[2] in the previous issue of EPIDEMIOLOGY, were developed to fulfill such a dream.

Instrumental variables have been defined using 4 different representations of causal effects:

1. Linear structural equations models developed in econometrics and sociology[3,4] and used by Martens et al[1]
2. Nonparametric structural equations models[4]
3. Causal directed acyclic graphs[4–6]
4. Counterfactual causal models[7–9]

Much of the confusion associated with IV estimators stems from the fact that it is not obvious how these various representations of the same concept are related. Because the precise connections are mathematical, we will relegate them to an Appendix. In the main text, we will describe the connections informally.

Let us introduce IVs, or instruments, in randomized experiments before we turn our attention to observational studies. The causal diagram in Figure 1 depicts the structure of a double-blind randomized trial. In this trial, Z is the randomization assignment indicator (eg, 1 = treatment, 0 = placebo), X is the actual treatment received (1 = treatment, 0 = placebo), Y is the outcome, and U represents all factors (some unmeasured) that affect both the outcome and the decision to adhere to the assigned treatment. The variable Z is referred to as an instrument because it meets 3 conditions: (i) Z has a causal effect on X, (ii) Z affects the outcome Y only through X (ie, no direct effect of Z on Y, also known as the exclusion restriction), and (iii) Z does not share common causes with the outcome Y (ie, no confounding for the effect of Z on Y). Mathematically precise statements of these conditions are provided in the Appendix.

A double-blind randomized trial satisfies these conditions in the following ways. Condition (i) is met because trial participants are more likely to receive treatment if they were assigned to treatment, condition (ii) is ensured by effective double-blindness, and condition (iii) is ensured by the random assignment of Z. The intention-to-treat effect (the average causal effect of Z on Y) differs from the average treatment effect of X on Y when some individuals do not comply with the assigned treatment. The greater the rate of noncompliance (eg, the smaller the effect of Z on X on the risk-difference scale), the more the intention-to-treat effect and the average treatment effect will tend to differ. Because the average treatment effect reflects the effect of X under optimal conditions (full compliance) and does not depend on local conditions, it is often of intrinsic public health or scientific interest.

Submitted 30 January 2006; accepted 6 February 2006.
From the *Department of Epidemiology, Harvard School of Public Health and †Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts.
Editors' note: *A related article appears on page 373.*
Correspondence: Miguel A. Hernán, Department of Epidemiology. Harvard School of Public Health. 677 Huntington Ave. 02115 Boston, MA. E-mail: Miguel_hernan@post.harvard.edu.

ISSN: 1044-3983/06/1704-0360
DOI: 10.1097/01.ede.0000222409.00878.37

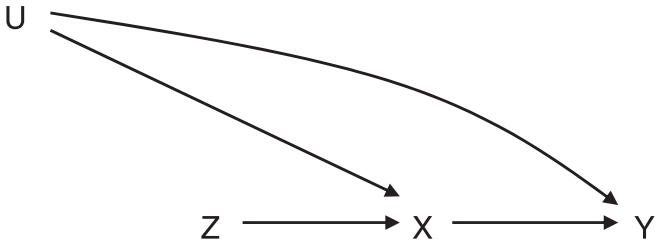

FIGURE 1. A double-blind randomized experiment with assignment *Z*, treatment *X*, outcome *Y*, and unmeasured factors *U*. *Z* is an instrument.

Unfortunately, the average effect of X on Y may be affected by unmeasured confounding.

Instrumental variables methods promise that if you collect data on the instrument Z and are willing to make some additional assumptions (see below), then you can estimate the average effect of X on Y, regardless of whether you measured the covariates normally required to adjust for the confounding caused by U. IV estimators bypass the need to adjust for the confounders by estimating the average effect of X on Y in the study population from 2 effects of Z: the average effect of Z on Y and the average effect of Z on X. These 2 effects can be consistently estimated without adjustment because Z is randomly assigned. For example, consider this well-known IV estimator: The estimated effect of X on Y is equal to an estimate of the ratio

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

of the effect of Z on Y divided by the effect of Z on X, all measured in the scale of difference of risks or means, where $\mathrm{E}[X|Z]=\Pr[X=1|Z]$ for the dichotomous variable X. (Martens et al[1] showed the derivation and a geometrical explanation of this IV estimator in the context of linear models, and Brookhart et al[2] applied it to pharmacoepidemiologic data.) To obtain the average treatment effect, one inflates the intention-to-treat effect in the numerator of the estimator by dividing by a denominator, which decreases as noncompliance increases. That is, the effect of X on Y will equal the effect of Z on Y when X is perfectly determined by Z (risk difference $\mathrm{E}[X|Z=1]-\mathrm{E}[X|Z=0]=1$). The weaker the association between Z and X (the closer the Z-X risk difference is to zero), the more the intention-to-treat effect will be inflated because of the shrinking denominator.

This instrumental variables estimator can also be used in observational settings. Investigators can estimate the average effect of an exposure X by identifying and measuring a Z-like variable that meets conditions (ii) and (iii) as well as a more general modified version of condition (i), which we designate as condition (i*). Under condition (i*), the instrument Z and exposure X are associated either because Z has a causal effect on X, or because X and Z share a common cause.[4,10] Martens et al[1] cite several articles that describe some instruments used in observational studies. As these examples show, the challenge of identifying and measuring an instrument in an observational study is not trivial. The goal of Brookhart et al[2] is to compare the effect of prescribing 2 classes of drugs (cyclooxygenase 2-[COX-2] selective and nonselective nonsteroidal antiinflammatory drugs [NSAIDs]) on gastrointestinal bleeding. The authors propose the "physician's prescribing preference" for drug class as the instrument, arguing that it meets conditions (i), (ii), and (iii). Because the proposed instrument is unmeasured, the authors replace it in their main analysis by the (measured) surrogate instrument "last prescription issued by the physician before current prescription."

Figure 2 shows a causal structure in which the instrument Z (here, "last prescription issued by the physician before current prescription") is a surrogate for another unmeasured instrument U^* (here, "physician's prescribing preference"). Both Z and U^* meet conditions (i*), (ii) and (iii) but, in contrast to U^*, Z does not satisfy the original condition (i). The original condition (i) is equivalent to the second assumption of Martens and colleagues[1] for the validity of an instrument. It follows that Martens et al's assumptions are too restrictive and do not recognize that Z can be used as an instrument. That is, under the Martens et al assumption that the equations are structural (as defined in the Appendix), their instrumental variables estimator is consistent for the effect of X on Y provided the instrument Z is uncorrelated with the error term E in the structural equation for the outcome Y (which implies no confounding for the causal effect of Z on Y), even when the instrument is correlated with the error term F in the structural equation for the treatment X (which implies confounding for the causal effect of Z on X).

The IV estimator described previously looks like an epidemiologist's dream come true: we can estimate the effect of the X on Y, even if there is unmeasured confounding for the effect of X on Y! Many sober readers, however, will suspect any claim that an analytic method solves one of the major problems in epidemiologic research. Indeed there are good reasons for skepticism—as Martens et al[1] explain, and as the example of Brookhart et al[2] illustrates. First, the IV effect

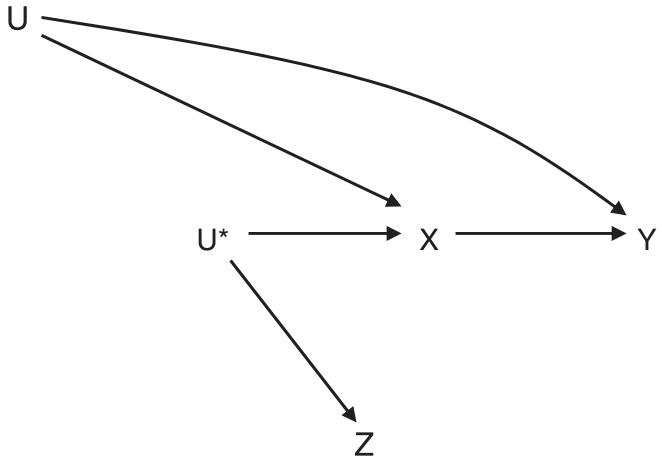

FIGURE 2. An observational study with unmeasured instrument *U**, exposure *X*, outcome *Y*, and unmeasured factors *U*. *Z* is a surrogate instrument.

estimate will be biased unless the proposed instrument meets conditions (ii) and (iii), but these conditions are not empirically verifiable. Second, any biases arising from violations of conditions (ii) and (iii), or from sampling variability, will be amplified if the association between instrument and exposure [condition (i*)] is weak. Third, our discussion so far may have appeared to suggest that conditions (i*), (ii), and (iii) are sufficient to guarantee that the IV estimate consistently estimates the average effect of X on Y. In fact, additional unverifiable assumptions are required, regardless of whether the data were generated from a randomized experiment or an observational study. Finally, most epidemiologic exposures are time-varying, which standard IV methods are poorly equipped to address.

We now briefly review these 4 reasons for skepticism (see also Greenland[11]). To illustrate these ideas, we will take the study by Brookhart et al[2] as an example because one can indirectly validate their observational estimates by comparing them with the estimates from a previous randomized trial that addressed the same question. We will focus on the effect of prescribing selective versus nonselective NSAIDs on gastrointestinal bleeding over a period of 60 days in patients with arthritis. This effect was estimated to be −0.47 (in the scale of risk difference multiplied by 100) in the randomized trial.

Violation of the Unverifiable Conditions (ii) and (iii) Introduces Bias

Condition (ii), the absence of a direct effect of the instrument on the outcome, will not hold if, as discussed by Brookhart et al,[2] doctors tend to prescribe selective NSAIDs together with gastroprotective medications (eg, omeprazol). This direct effect of the instrument would introduce a downward bias in the estimate, that is, the effect of prescribing selective NSAIDs would look more protective than it really is. However, the assumption cannot be verified from the data: the unexpectedly strong inverse association between Z and Y (−0.35, Table 3 in Brookhart et al) is consistent with a violation of condition (ii) but also with a very strong protective effect of selective NSAIDs without a violation of condition (ii).

Brookhart and colleagues[2] also discuss the possibility that physicians who prescribe selective NSAIDs frequently see higher-risk patients. This potential violation of condition (iii) is the result of unmeasured confounding for the instrument and would introduce an upward bias in the estimate. To deal with this potential problem—consistent with the association between Z and the measured covariates (Table 2 in Brookhart et al)—the authors made the unverifiable assumption that, within levels of the measured covariates, there were no other common causes of the instrument and the outcome.

These violations of conditions (ii) and (iii) can be represented by including arrows from U* to Y and from U to Z, respectively (Fig. 3).

A Weak Condition (i*) Amplifies The Bias

An instrument weakly associated with exposure leads to a small denominator of the IV estimator. Therefore, biases that affect the numerator of the IV estimator (eg, unmeasured confounding for the instrument, a direct effect of the instrument) or small sample bias in the denominator will be greatly exaggerated, and may result in an IV estimate that is more biased than the unadjusted estimate. The exaggeration of the effect by IV estimators may occur even in large samples and in the absence of model misspecification. In the study by Brookhart et al,[2] the overall $Z - X$ risk difference was 0.228 (the corresponding number in patients with arthritis was not reported). Therefore, any bias affecting the numerator of the IV estimator would be multiplied by approximately 4.4 (1/0.228), which might explain why the IV effect estimate −1.81 was farther from the randomized estimate −0.47 than the unadjusted estimate 0.10. The IV method might have exaggerated the effect if the proposed instrument had a direct effect due to, say, concomitant prescription of gastroprotective drugs. Alternatively, the instrument Z may satisfy conditions (i*), (ii), and (iii). In that case, the difference between the IV and the randomized estimates might not be due to bias in the instrumental variable estimator but rather to sampling variability or (as suggested by Brookhart et al) to the different age distributions in the observational study and the randomized trial, along with strong effect-measure modification by age. The latter hypothesis could be assessed by conducting an analysis stratified by age.

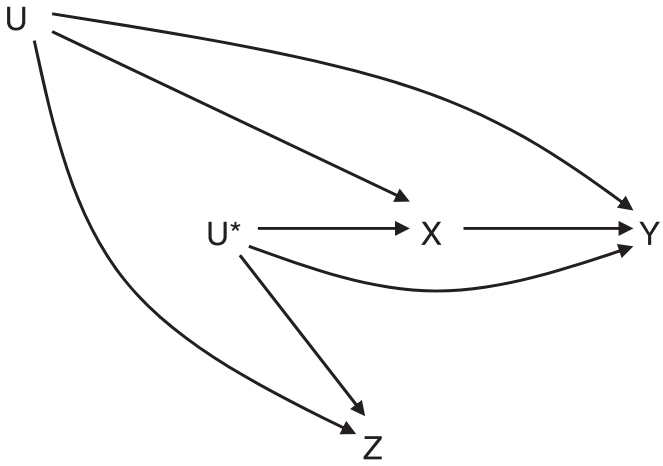

FIGURE 3. An observational study with exposure X, outcome Y, and unmeasured factors U in which the variables U* and Z do not qualify as instruments.

In the context of linear models, Martens et al[1] demonstrate that instruments are guaranteed to be weakly correlated with exposure in the presence of strong confounding because a strong association between X and U leaves little residual variation for X to be strongly correlated with the instrument U* in Figure 2. This problem may be compounded by the use of surrogate instruments Z.

When Treatment Effects Are Heterogenous, Conditions (i*) Through (iii) Are Insufficient to Obtain Effect Estimates

Even when an instrument is available, additional assumptions are required to estimate the average causal effect of X in the population. Examples of such assumptions are discussed in the following paragraphs as well as in the Appendix. Conditions

(i*), (ii), and (iii) allow one to compute upper and lower bounds, but not a point estimate, for the average causal effect. In a 1989 article, Robins[8] derived the bounds that can be computed under conditions (i*) and (ii) plus a weak version of condition (iii), as well as under different sets of other unverifiable assumptions. Subsequently, Manski[12] derived related results, and Balke and Pearl[13] derived narrower bounds under a stronger version of condition (iii) given in the Appendix; this holds, for example, when the instrument is a randomized assignment indicator. In a double-blind randomized trial, confidence intervals for the intention-to-treat effect of Z on Y that exceed zero by a wide margin show that a positive treatment effect is occurring in a subset of the population. However, if noncompliance is large (say, 50%), bounds for the average treatment effect may include the null hypothesis of zero. This would happen if, for example, the (unobserved) effect of treatment in the noncompliers were larger in magnitude and opposite in sign to that in the compliers.

However, Martens et al[1] and Brookhart et al[2] do present point estimates—not bounds—for the causal effect of X on Y. What other assumptions did the authors make either explicitly or implicitly? The linear structural equation model used by Martens et al assume that the effect of X on Y on the mean-difference scale is the same for all subjects. This assumption of no between-subject heterogeneity in the treatment effect combined with conditions (i*), (ii), and (iii) is sufficient to identify the effect of X on Y. (A causal effect is said to be identified if there exists an estimator based on the observed data [Z, X, Y] that converges to [is consistent for] the effect in large samples). This assumption will hold under the sharp null hypothesis that the exposure X has no effect on any subject's outcome (in contrast with the "nonsharp" null hypothesis in which the net effect is still zero but includes positive effects for some and negative for others). It follows that, when conditions (i*), (ii), and (iii) hold, the usual IV estimator will correctly estimate the average treatment effect of 0 whenever the sharp null hypothesis is true. However, when the sharp null is false, the assumption of no treatment effect heterogeneity is biologically implausible for continuous outcomes and logically impossible for dichotomous outcomes.

There is a weaker, more plausible assumption that, combined with conditions (i*), (ii) and (iii), still implies the effect of X on Y is the ratio of the effect of Z on Y to the effect of Z on X. This is the assumption that the X–Y causal risk difference is the same among treated subjects with $Z = 1$ as among treated subjects with $Z = 0$, and similarly among untreated subjects.[8,14] In other words, this assumes that there is no effect modification, on the additive scale, by Z of the effect of X on Y in the subpopulations of treated and untreated subjects (strictly speaking, any effect modification would be due to the causal instrument U*). The identifying assumption of no effect modification will not generally hold if the unmeasured factors U on Figure 2 interact with X on an additive scale to cause Y. Such effect modification would be expected in many studies, including that by Brookhart et al.[2] There might be effect modification, for example, if the risk difference for the effect of selective NSAIDs (X) on gastrointestinal bleeding (Y) was modified by past history of gastritis (U).

Another assumption that is commonly combined with conditions (i*), (ii), and (iii) to identify the average effect of X on Y is the monotonicity assumption. In the context of the research by Brookhart et al,[2] with dichotomous Z and U*, monotonicity means that no doctor who prefers nonselective NSAIDs would prescribe selective NSAIDs to any patient unless all doctors who prefer selective NSAIDs would do so. Clearly, in the substantive setting of the study by Brookhart et al, monotonicity is unlikely to hold. In other settings, monotonicity may be more likely. The monotonicity assumption does not affect the bounds for the average effect of X on Y in the population (our target parameter so far).[8,13] However, in the Appendix, we extend a result by Imbens and Angrist[15] to show that, if the assumptions encoded by the DAG in Figure 2 and the assumption of monotonicity all hold, a particular causal effect is identified and the usual IV estimator based on Z consistently estimates this effect. The identified causal effect is the average effect of X on Y in the subset of the study population who would be treated (1) with selective NSAIDs by all doctors whose "prescribing preference" is for selective NSAIDs and (2) with nonselective NSAIDs by all doctors whose preference is for nonselective NSAIDs.[15] This subset of the study population can be labeled as the "compliers" because it is analogous to the subset of the population in randomized experiments (in which the instrument is treatment assignment) who would comply with whichever treatment is assigned to them. A problem with this causal effect is that we cannot identify the subset of the population (the "compliers") the effect estimate refers to. Further, this result requires that a doctor's unobserved "prescribing preference" U* can be assumed to be dichotomous. In the Appendix we argue that assumptions encoded by the DAG in Figure 2 are more substantively plausible if U* is a continuous rather than a dichotomous measure, although in that case a "complier" is no longer well defined and the interpretation of the IV estimator based on Z is different (see Appendix).

The assumptions of monotonicity and no effect modification by Z on an additive (risk difference) scale by no means exhaust the list of assumptions that serve to identify causal effects. Alternative identifying assumptions can result in estimators of the average effect of X that differ from the usual IV estimator. For example, in the Appendix, we show that the assumption of no effect modification by Z on an multiplicative (risk ratio) scale within both levels of X identifies the average causal effect.[8,10] However, under this assumption, the estimated ratio of the average effect of Z on Y to the average effect of Z on X is now biased (inconsistent) for the average causal effect of X on Y; in the Appendix we provide a consistent (asymptotically normal) estimator for the treatment effect.[8,10]

Because all identifying assumptions are unverifiable, Robins and Greenland[16] argued that it is useful to estimate upper and lower bounds for the effect, instead of (or in addition to) point estimates and confidence intervals obtained under various explicit unverifiable assumptions. Such estimates help to make clear "the degree to which public health

decisions are dependent on merging the data with strong prior beliefs." As noted above, the problem with bounds is that the resulting interval may be too wide and therefore not very informative. (Further, there will be 95% confidence intervals around the upper and the lower bound attributable to sampling variation.)

In addition, when it is necessary to condition on continuous (or many discrete) preinstrument covariates to try to insure that the effect of Z on Y is unconfounded, the validity of IV estimates based on parametric linear models for a binary response Y also requires as usual both a correctly specified functional form for the covariates effects and estimated conditional probabilities that lie between zero and one.

The Standard IV Methodology Deals Poorly With Time-Varying Exposures

Most epidemiologic exposures are time-varying. For example, Brookhart et al[2] compared the risks after prescription of either selective or nonselective NSAIDs, regardless of whether patients stayed on the assigned drug during the follow-up. In other words, the treatment variable was considered to not be time-varying, and the authors estimated an observational analog of the intention-to-treat effect commonly estimated from randomized experiments. However, in reality, patients may discontinue or switch their assigned treatment over time. When this lack of adherence to the initial treatment is not due to serious side effects, one could be more interested in comparing the risks had the patients followed their assigned treatment continuously during the follow-up.

In the presence of time-varying instruments, exposures, and confounders, Robins's g-estimation of nested structural models[10,17–19] can be used to estimate causal effects. Nested structural models achieve identification by assuming a nonsaturated model for the treatment effect at each time t (measured on either an additive or multiplicative scale) as a function of a subject's treatment, instrument, and covariate history through t. These models naturally allow the analyst (1) to obtain asymptotically unbiased point estimates of the treatment effect in the treated study population, (2) to characterize the effect on one's inference to violations of the model assumptions through sensitivity analysis, (3) to adjust for baseline and time-varying continuous and discrete confounders of the instrument-outcome association, (4) to include continuous and multivariate instruments and treatments, and (5) to use doubly-robust estimators. In the Appendix we show that the linear structural equations of Martens et al[1] are a simple case of a nested structural mean model. Robins's methods apply to continuous, count, failure time, and rare dichotomous responses but not to nonrare dichotomous responses.[20] For nonrare dichotomous responses, a new extension due to Van der Laan et al[21] can be used. For treatments and instruments that are not time-varying, Tan[22] has shown how to achieve many of properties (a) through (e) under a model that achieves identification of causal effects by assuming monotonicity.

CONCLUSION

We have reviewed how, in observational research, the use of instrumental variables methods replaces the unverifiable assumption of no unmeasured confounding for the treatment effect with other unverifiable assumptions such as "no unmeasured confounding for the effect of the instrument" and "no direct effect of the instrument." Hence, the fundamental problem of causal inference from observational data–the reliance on assumptions that cannot be empirically verified—is not solved but simply shifted to another realm. As always, investigators must apply their subject-matter knowledge to study design and analysis to enhance the plausibility of the unverifiable assumptions.

Further, when conditions (i*), (ii), and (iii) do not hold, the direction of bias of IV estimates may be counterintuitive for epidemiologists accustomed to conventional approaches for confounding adjustment. For example, Brookhart et al[2] found a much bigger effect estimate using IV methods (-1.81) than the effect estimated by the randomized trial (-0.47), whereas conventional methods were unable to detect a beneficial effect of selective NSAIDs. The conventional unadjusted and adjusted estimates were quite close (0.10 and 0.07, respectively), despite careful adjustment for most of the known indications and risk factors for the outcome. If the assumptions required for the validity of the usual IV estimator held and these differences were not the result of sampling variability, the aforementioned estimates would imply that the magnitude of unmeasured confounding (from 0.07 to -1.81) is much greater than the magnitude of the measured confounding (from 0.10 to 0.07). An alternative explanation is that the IV assumptions do not hold and the IV estimate is biased in the apparently counterintuitive direction of exaggerating the protective effect.

In summary, Martens et al[1] are right: IV methods are not an epidemiologist's dream come true. Nonetheless, they certainly deserve greater attention in epidemiology, as shown by the interesting application presented by Brookhart et al[2] But users of IV methods need to be aware of the limitations of these methods. Otherwise, we risk transforming the methodologic dream of avoiding unmeasured confounding into a nightmare of conflicting biased estimates.

APPENDIX

This appendix is organized in 5 sections. The first section describes 4 mathematical representations of causal effects—counterfactuals, causal directed acyclic graphs, nonparametric structural equation models, linear structural equations models—and their relations. The second section describes IV estimators that identify the average causal effect of X on Y in the population by using no interaction assumptions. We show that these estimators can be represented by parameters of particular structural mean models. The third section describes IV estimators that identify the average causal effect of X on Y in certain subpopulations by using monotonicity assumptions. The fourth section contains important extensions. The last section contains the proofs of the theorems presented in the first 3 sections.

1. Representations of Causal Effects

As mentioned in the main text, IV estimators have been defined using 4 different mathematical representations of

causal effects. We now briefly describe each of these representations:

1.1 Counterfactuals

A counterfactual random variable $Y(x, z)$ encodes the value that the variable Y would have if, possibly contrary to fact, the variable X were set to the value x and the variable Z set to z. The counterfactual variable $Y(x, z)$ is assumed to be well defined[23] in the sense that there is reasonable agreement as to the hypothetical intervention (ie, closest possible world) which sets X to x and Z to z.

Counterfactuals allow us to give precise mathematical definitions for conditions (ii) and (iii) in the definition of an instrument. Condition (ii), the exclusion restriction, is formalized under the counterfactual model by the assumption that for all subjects,

$$Y(x, z = 1) = Y(x, z = 0) = Y(x)$$

where $Y(x)$ is the counterfactual value of Y when X is set to x, but each subject's Z takes the same value as in the observed data.[7] The condition (iii) that there is no confounding for the effect of Z on Y is formalized by the 2 assumptions

$$Y(x = 1) \amalg Z \quad \text{and} \quad Y(x = 0) \amalg Z$$

where $A \amalg B$ is read as "A is independent of B". The average causal effect of X on Y is defined to be $\mathrm{E}[Y(x = 1)] - \mathrm{E}[Y(x = 0)]$ when X is dichotomous, which we also write as $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$ when no ambiguity will arise.

1.2 Causal Directed Acyclic Graphs (DAG)[4,5]

We define a DAG G to be a graph whose nodes (vertices) are M random variables $V = (V_1, \ldots, V_M)$ with directed edges (arrows) and no directed cycles. We use PA_m to denote the parents of V_m, ie, the set of nodes from which there is a direct arrow into V_m. The variable V_j is a descendant of V_m if there is a sequence of nodes connected by edges between V_m and V_j such that, following the direction indicated by the arrows, one can reach V_j by starting at V_m. For example, consider the causal DAG in Figure 2 that represents the causal structure of an observational study with a surrogate instrument Z. In this DAG, $M = 5$ and we can choose $V_1 = U$, $V_2 = U^*$, $V_3 = Z$, $V_4 = X$, $V_5 = Y$; the parents PA_4 of $V_4 = X$ are (U, U^*) and the nondescendants of X are (U, U^*, Z).

A causal DAG is a DAG in which (1) the lack of an arrow from node V_j to V_m can be interpreted as the absence of a direct causal effect of V_j on V_m (relative to the other variables on the graph) and (2) all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph. In Figure 2, the lack of a direct arrow between Z and Y indicates that treatment prescribed to the previous patient Z does not have a direct causal effect (causative or preventive) on the next patient's outcome Y. Also, the inclusion of the measured variables (Z, X, Y) implies that the causal DAG must also include their unmeasured common causes (U, U^*). Note a causal DAG model makes no reference to and is agnostic as to the existence of counterfactuals.

Our causal DAGs are of no practical use unless we make some assumption linking the causal structure represented by the DAG to the statistical data obtained in an epidemiologic study. This assumption, referred to as the causal Markov assumption (CMA), states that the nondescendants of a given variable V_j are independent of V_j conditional on the parents (ie, direct causes) of V_j. The CMA is mathematically equivalent to the statement that the density $f(V)$ of the variables V in DAG G satisfies the Markov factorization

$$f(v) = \prod_{j=1}^{M} \mathrm{f}(v_j | pa_j).$$

1.3 Nonparametric Structural Equation Models (NPSEMs)[4]

An NPSEM is a causal model that both assumes the existence of counterfactual random variables and can be represented by a DAG. To provide a formal definition of an NPSEM represented by a DAG G, we shall use the following notation. For any random variable W, let $\mathcal{W}$ denote the support (ie, the set of possible values w) of W. For any $w_1, \ldots, w_m$, define $\bar{w}_m = (w_1, \ldots, w_m)$. Let R denote any subset of variables in V and let r be a value of R. Then $V_m(r)$ denotes the counterfactual value of V_m when R is set to r. We number the variables V so that for $j < i$ V_j is not a descendant of V_i.

An NPSEM represented by a DAG G with vertex set V assumes the existence of mutually independent unobserved random variables (errors) ε_m and deterministic unknown functions $f_m(pa_m, \varepsilon_m)$ such that $V_1 = f_1(\varepsilon_1)$ and the one-step ahead counterfactual $V_m(\bar{v}_{m-1}) \equiv V_m(pa_m)$ is given by $f_m(pa_m, \varepsilon_m)$, and both V_m and the counterfactuals $V_m(r)$ for any $R \subset V$ are obtained recursively from V_1 and the $V_m(\bar{v}_{m-1})$, $m > 1$. For example, $V_3(v_1) = V_3\{v_1, V_2(v_1)\}$ and $V_3 = V_3\{V_1, V_2(V_1)\}$. In Figure 2, $Y(z,x) = V_5(v_3, v_4) = f_5(V_1, v_4, \varepsilon_5) = f_Y(U, x, \varepsilon_Y)$ does not depend on z since Z is not a parent of Y or U, where we define $f_Y = f_5$, $\varepsilon_Y = \varepsilon_5$ since $Y = V_5$. In summary, only the parents of V_m have a direct effect on V_m relative to the other variables on G. A DAG G represented by an NPSEM is a causal DAG for which the CMA holds because the independence of the error terms ε_m both implies the CMA and is essentially equivalent to the requirement that all common causes of any variables on the graph are themselves on the causal DAG.

1.4 Linear Structural Equation Models (LSEMs)

A (causal) LSEM for the observed variables is the special case of an NPSEM in which for each observed V_m the deterministic functions $f_m(pa_m, \varepsilon_m)$ are linear in all the observed parents of V_m. For example, in Figure 2, an LSEM for Y assumes $Y = f_Y(X, U, \varepsilon_y) = \beta X + \Delta_Y$ is linear in X, where $\Delta_Y = \delta_Y(U, \varepsilon_y)$ is an unknown function of the unobservables (U, ε_y). Note this LSEM for Y implies that the treatment effect $Y(x = 1) - Y(x = 0) = \beta$ is the same constant β for all subjects, since according to the model $Y(x = 1) = \beta + \Delta_Y$ and $Y(x = 0) = \Delta_Y$. Linear structural equation modelers

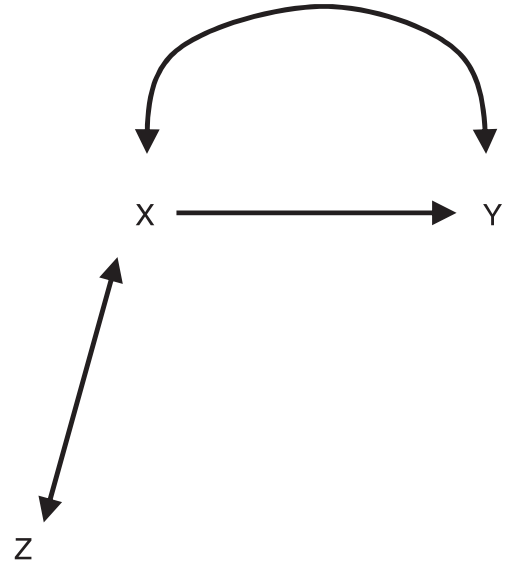

FIGURE 4. The observational study represented by Figure 2 with all unmeasured common causes replaced by bidirectional arrows.

would redraw the DAG in Figure 2 as the DAG in Figure 4, since they replace unmeasured common causes of 2 measured variables by bidirectional edges.

These 4 causal models are connected as follows. An LSEM is a special case of an NPSEM. An NPSEM is both a causal DAG model and a counterfactual model. For example, the NPSEM represented by the DAG in Figure 2 implies the counterfactual versions of conditions (ii) and (iii) previously. In fact an NPSEM implies a stronger version of condition (iii): joint independence of the counterfactuals and Z, represented as

$$\{Y(x = 1),\ Y(x = 0)\} \amalg Z$$

Although an NPSEM is a causal DAG, not all causal DAG models are NPSEMs. Indeed as mentioned above, a causal DAG model makes no reference to and is agnostic about the existence of counterfactuals. In this appendix we shall use counterfactuals freely to derive results. In Section 4, we briefly consider which of our results would remain true under a causal DAG agnostic about counterfactuals. All the results for NPSEMs described in this appendix actually hold under the slightly weaker assumptions encoded in a fully randomized causally interpreted structured tree graph (FRCISTG) model of Robins.[24,25] All NPSEMs are FRCISTGs but not all FRCISTGs are NPSEMs.[26]

2. IV Estimators and Effect Modification

In this section we show that the usual IV estimator estimates the parameter of a particular additive structural mean model: a counterfactual model for the effect of treatment on the treated. We then describe additional assumptions necessary for this estimator to also identify the average causal effect $E[Y(1)] - E[Y(0)]$ of X on Y in the entire study population. We end by contrasting these results with those obtained under a multiplicative structural mean model.

2.1 Additive Structural Mean Models (SMMs)

Additive and multiplicative SMMs were introduced by Robins[8] in 1989 and were treated more fully in his later work.[10] We first consider the special case in which X and Z are time-independent and dichotomous and there are no covariates (eg, measured confounders of the effect of Z on Y). The general time-independent case is treated in Section 4. See Robins[10,18] for time-varying treatments instruments and confounders. A nonparametric (saturated) additive SMM is

$$E[Y(1)|X = 1, Z] - E[Y(0)|X = 1, Z] = \gamma\{1, Z, \psi^*\}$$
$$\text{where } \gamma\{1, Z, \psi^*\} = \psi_0^* + \psi_1^* Z$$

or, equivalently,

$$E[Y|X, Z] - E[Y(0)|X, Z] = \gamma\{X, Z, \psi^*\} = X(\psi_0^* + \psi_1^* Z)$$

where $Y(1)$ and $Y(0)$ are shorthand for $Y(x = 1)$ and $Y(x = 0)$, respectively, and ψ_0^* and ψ_1^* are unknown parameters. The parameter ψ_0^* is the average causal effect of treatment among the treated subjects with $Z = 0$. Similarly $\psi_0^* + \psi_1^*$ is the average causal effect of treatment among the treated subjects ($X = 1$) with $Z = 1$. Thus, for the treated subjects, the parameter ψ_0^* is the main effect of treatment and ψ_1^* quantifies effect modification by Z on an additive scale. It immediately follows that an LSEM $Y = \beta X + \Delta_Y$ is an additive SMM without effect modification by Z with $\psi_0^* = \beta$ and $\psi_1^* = 0$.

We turn next to identification and estimation of the parameters of this additive SMM under the conditional mean independence assumption

$$E[Y(0)|Z = 1] = E[Y(0)|Z = 0] \quad (1)$$

which, by condition (iii), is satisfied by the NPSEM represented by the DAG in Figure 2 (but not by the DAG in Fig. 3). This assumption can be conveniently rewritten in the mathematically equivalent form

$$E[Y - X(\psi_0^* + \psi_1^*)|Z = 1] = E[Y - X\psi_0^*|Z = 0] \quad (2)$$

Let us first consider the case where we assume $\psi_1^* = 0$ a priori so there is no effect modification by Z among the treated. Then ψ_0^* is the only unknown parameter. Solving the aforementioned equation for ψ_0^* with $\psi_1^* = 0$, we have

$$\psi_0^* = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[X|Z = 1] - E[X|Z = 0]} \quad (3)$$

That is, ψ_0^* is exactly the usual IV estimand–the ratio of the average effect of Z on Y to the average effect of Z on X.[10] We conclude that the usual IV estimator is estimating the parameter ψ_0^* of our additive SMM.

However if, as in most of the main text, our interest is in the average causal effect $E[Y(1)] - E[Y(0)]$ of X on Y in the study population, we are not yet finished because ψ_0^* does not

generally equal $E[Y(1)] - E[Y(0)]$. Rather, by definition of the model and the assumption of no effect modification by Z among the treated,

$$\psi_0^* = E[Y(1) - Y(0)|X = 1, Z = 1] = E[Y(1) - Y(0)|X = 1, Z = 0]$$

and thus $\psi_0^* = E[Y(1)|X = 1] - E[Y(0)|X = 1]$ is the effect of treatment on the treated ($X = 1$).[10,14] To conclude that $\psi_0^* = E[Y(1)] - E[Y(0)]$ and thus that $E[Y(1)] - E[Y(0)]$ is the usual IV estimand, we must assume or derive that the average effect of treatment on the treated and on the untreated are identical:

$$E[Y(1)|X = 1] - E[Y(0)|X = 1] = E[Y(1)|X = 0] - E[Y(0)|X = 0] \quad (4)$$

Equation 4 obviously holds when we assume an LSEM for Y since an LSEM implies the same treatment effect for all subjects regardless of their X. We can therefore conclude, as stated in the text, that assuming an LSEM for Y identifies the average causal effect as the ratio (3) irrespective of whether the denominator $E[X|Z = 1] - E[X|Z = 0]$ equals the causal effect of Z on X (as on the DAG in Fig. 1) or simply reflects the noncausal association between Z and X due to the presence of their common cause U* (as on the DAG in Fig. 2).

We now provide weaker, somewhat more plausible, assumptions than those imposed by an LSEM for Y under which (4) holds and thus (3) equals $E[Y(1)] - E[Y(0)]$. These weaker assumptions are mean independence of $Y(1)$ and Z as

$$E[Y(1)|Z = 1] = E[Y(1)|Z = 0] \quad (5)$$

and the assumption (6a) of no effect modification by Z within the untreated ($X = 0$). Consider the assumptions

$$E[Y(1) - Y(0)|X = 0, Z = 1] = E[Y(1) - Y(0)|X = 0, Z = 0], \quad (6a)$$

$$E[Y(1) - Y(0)|X = 1, Z = 1] = E[Y(1) - Y(0)|X = 1, Z = 0]. \quad (6b)$$

Assumption (6a) was called the assumption of no current treatment interaction with respect to Z in Robins (1994), and (6b) is a restatement of our assumption $\psi_1^* = 0$. Robins[8,10] noted that the conjunction of these 2 assumptions of no effect modification plus the counterfactual mean independence assumptions (1) and (5) implies

$$E[Y(1)] - E[Y(0)] = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[X|Z = 1] - E[X|Z = 0]} \quad (7)$$

Heckman[14] later derived the same identifying formula under the joint assumption that average treatment effect in those with $Z = 1$ did not further depend on X, and similarly for $Z = 0$. That is,

$$E[Y(1) - Y(0)|X = 1, Z = 1] = E[Y(1) - Y(0)|X = 0, Z = 1] \quad (8a)$$

$$E[Y(1) - Y(0)|X = 1, Z = 0] = E[Y(1) - Y(0)|X = 0, Z = 0] \quad (8b)$$

In fact, given Z and X are correlated and our 2 counterfactual mean independence assumptions, the assumptions (8) are equivalent to (6) as proved in Theorem 1 in the last section. Results closely related to Theorem 1 were discussed by Heckman.[14]

Furthermore, under the NPSEM represented by the DAG in Figure 2, we show in Theorem 2 of section 5 that sufficient conditions for equation (6b) [i.e., $\psi_1^* = 0$] are that, with probability one, either $Y(x = 1, U) - Y(x = 0, U)$ does not depend on U or $\frac{\Pr[X(z = 1, U) = x]}{\Pr[X(z = 0, U) = x]}$ does not depend on U for $x = 1$. Sufficient conditions for (6a) are identical except that we replace "for $x = 1$" by "for $x = 0$". However, under our NPSEM it is impossible for the above ratio not to depend on U for $x = 1$ and $x = 0$ simultaneously (see Theorem 3 in section 5). Thus whenever U and X interact on the additive scale to cause Y, it would not be reasonable to assume the usual IV estimand exactly equals the average effect $E[Y(1) - Y(0)]$. Finally note that the condition "$Y(x = 1, U) - Y(x = 0, U)$ does not depend on U" does not imply the treatment effect is the same for each individual as $Y(x = 1, U) - Y(x = 0, U) = f_y(x = 1, U, \varepsilon_y) - f_y(x = 0, U, \varepsilon_y)$ may depend on ε_y, although not on U.

2.2 Multiplicative SMM

We now show that if we assume a multiplicative (ie, log-linear) SMM without interaction and no multiplicative effect modification by Z given $X = 0$, $E[Y(1)] - E[Y(0)]$ remains identified (ie, depends on the distribution of the observed data), but no longer equals (7). The new identifying estimand is given in the next theorem.

Again we consider the special case of dichotomous X and Z and no covariates. The saturated multiplicative SMM is

$$E[Y(1)|X = 1, Z] = E[Y(0)|X = 1, Z]\ \gamma\{1, Z, \psi^*\} \quad \text{where } \gamma\{1, Z, \psi^*\} = \exp\{\psi_0^* + \psi_1^* Z\} \quad (9)$$

or, equivalently,

$$E[Y|X, Z] = E[Y_0|X, Z] \exp\{X\{\psi_0^* + \psi_1^* Z\}\}$$

For a dichotomous Y, $\exp\{\psi_0^*\}$ is the causal risk ratio in the treated subjects with $Z = 0$ and $\exp\{\psi_0^* + \psi_1^*\}$ is the causal risk ratio in the treated with $Z = 1$.

Theorem 4 in Section 5 shows that, when Equation (1) holds and $\psi_1^* = 0$, ie, no multiplicative effect modification by Z in the treated, then

$$\exp(-\psi_0^*) = 1 - \frac{E[Y|Z = 1] - E[Y|Z = 0]}{\left\{\begin{array}{l} E[Y|X = 1, Z = 1]\,E[X|Z = 1] \\ -\,E[Y|X = 1, Z = 0]\,E[X|Z = 0] \end{array}\right\}}$$

If, in addition, Equation 5 holds and there is no multiplicative effect modification by Z in the untreated, i.e.,

$$\frac{E[Y(1)|X=0, Z=1]}{E[Y(0)|X=0, Z=1]} = \frac{E[Y(1)|X=0, Z=0]}{E[Y(1)|X=0, Z=0]}$$

then $E[Y(1)]/E[Y(0)] = \exp(\psi_0^*)$, and the average causal effect is

$$E[Y(1)] - E[Y(0)] = E[Y|X=0] \times \{1 - E[X]\}\,[\exp(\psi_0^*) - 1] + E[X]\,E[Y|X=1] \quad (10)$$

Because whenever $E[Y(1)] \neq E[Y(0)]$, the expression for $E[Y(1)] - E[Y(0)]$ in Equation 10 differs from that in Equation 7, our estimate of $E[Y(1)] - E[Y(0)]$ will depend on whether we assume no effect modification by Z on an additive versus a multiplicative scale. Unfortunately, as shown by Robins,[10] it will not be possible to determine which, if either, assumption is true. The reason for this impossibility is that, even if we had an infinite sample size and Equations 1 and 5 hold, the only equality restriction on the joint distribution of the observed data is given by Equation 2 or by the mathematically equivalent expression

$$E[Y \exp\{-X\{\psi_0^* + \psi_1^*\}\}|Z=1] = E[Y \exp\{-X\psi_0^*\}|Z=0] \quad (11)$$

Thus we have only one restriction (ie, one equation) satisfied by the distribution of the observed data. This single restriction can be written in either of the 2 different but mathematically equivalent forms Equation 2 or Equation 11. Because with one equation it is not possible to solve for 2 parameters, one cannot test whether $\psi_1^* = 0$ either in the saturated additive model of Eq. (2) or in the saturated multiplicative model of Eq. (11). Further, one cannot solve for ψ_0^* in either model or estimate the average treatment in the total population or any subpopulation. Thus only bounds on $E[Y(1)] - E[Y(0)]$ are available.[8] Under an NPSEM's stronger version of condition (iii), which implies (but is not implied by) Equations 1 and 5, Balke and Pearl[13] showed that $E[Y(1)] - E[Y(0)]$ is identified in certain exceptional circumstances that are so unusual as to be curiosities.

Summarizing, we cannot identify causal effects using additive or multiplicative SMMs when we leave the functions $\gamma\{X, Z, \psi\}$ completely unspecified (saturated) as we then have more unknown parameters to estimate than equations to estimate them with. Thus, for identification, we must reduce the dimension of ψ through modeling assumptions, such as assuming certain interactions and/or main effects are absent.

An additional point is that although the assumptions encoded in the DAG in Figure 2 (ie, conditions (ii) and (iii) in the main text) are not empirically verifiable, they can, for certain data distributions, be empirically rejected.[13] More precisely, there exist empirical α-level tests of the composite assumptions encoded in the DAG in Figure 2 that, when they reject, the rejection can be taken as evidence against those assumptions. But, for most data distributions under which the assumptions encoded in the DAG are false, these tests will fail to reject at greater than level α even with an infinite sample size. That is, the tests are not consistent against all alternatives.

Additive and multiplicative SMM models were developed to provide a rigorous framework for identification and estimation via instrumental variables of the effects of a time-varying treatment or exposure. SMMs explicitly use counterfactuals (ie, potential outcomes) to characterize the consequences of between-subject heterogeneity in the treatment effect for instrumental variable estimation. For time-independent (but not for time-varying treatments) treatments, additive SMMs are somewhat related to the random coefficients model discussed by Heckman and Robb.[3,27] However, Heckman and Robb did not fully appreciate the usefulness of instrumental variable methods in these models. In particular, Heckman and Robb[3,27] and Heckman[28]–in contrast to Robins[10]–failed to recognize the value of instrumental variables for estimating average effect of treatment on the treated in the presence of heterogenous treatment effects (ie, random coefficients), as pointed out by Angrist, Imbens, and Rubin.[29]

3. IV Estimators Based on Monotonicity Assumptions

As discussed in the text, monotonicity assumptions are an alternative to the assumption of a nonsaturated model for $\gamma\{X, Z, \psi\}$ for obtaining identification.

When the causal instrument U^* is binary, we can define the compliers to be subjects for whom $X(u^* = 0) = 0$, $X(u^* = 1) = 1$. Imbens and Angrist[15] proved that the average causal effect in the compliers

$$E[Y(x=1) - Y(x=0)|X(u^*=0) = 0, X(u^*=1) = 1]$$

equals $\dfrac{E[Y|U^*=1] - E[Y|U^*=0]}{E[X|U^*=1] - E[X|U^*=0]}$ under the monotonicity assumption $X(u^* = 1) \geq X(u^* = 0)$ for all subjects. However, they considered a setting in which, in contrast to ours, data on the causal instrument U^* was available. In Theorem 5 of Section 5, we show that the average effect in the compliers is identified by the ratio (7) even when we only have data on a surrogate Z for the causal instrument U^*. This result depends critically on 2 assumptions: that Z is independent of X and Y given the causal instrument U^*, and that U^* is binary. However, we now argue that the independence assumption has little substantive plausibility unless U^* is continuous. To do so we need to provide a more precise operational definition of a physician's prescribing preference. We consider 2 possible definitions—one binary and one continuous.

Definition 1

Dichotomous prescribing preference: Let U^* be a dichotomous (0,1) *variable that takes the value* 1 *for a subject i if and only if at the time the physician treats subject i, he would treat more than* 50% *of all study subjects with selective NSAIDs.*

Definition 2

Continuous prescribing preference: Let U be a continuous variable whose value for subject i is the proportion of the study population that the subject's physician would treat with selective NSAIDs at the time the physician treats the subject i.*

Consider 2 physicians both with the continuous $U^* > 0.5$ (say, one with continuous U^* equal to 0.51 and the other equal to 0.95) and thus with discrete $U^* = 1$. Then if the last patient treated by subject i's physician received selective NSAIDs ($Z = 1$), it is more likely that the patient's physician had the higher continuous U^* and thus it is more likely that subject i will receive selective NSAIDs ($X = 1$). That is, X and Z will be correlated given the discrete U^* and the DAG in Figure 2 will not represent the data. However, Figure 2 remains plausible if we use the continuous definition of U^*. In that case, neither Theorem 5 nor its monotonicity assumption are relevant. Rather, for continuous U^* we define monotonicity as follows:

Definition of Monotonicity for Continuous *U**

If a physician with $U^* = u$ would treat patient i with selective NSAIDs, then all physicians with U^* greater than or equal to u would treat the patient with selective NSAIDs. Formally, $X(u^*)$ is a nondecreasing function of u^* on the support of U^*.

Note that, under the DAG in Figure 2, U^* satisfies $\Pr(X = 1|U^*) = U^*$, ie, among those patients whose physician would treat a fraction U^* of all patients, the fraction of patients who receive treatment is exactly U^*. That is, the continuous instrument U^* is the propensity score for treatment.

Let MTP(u^*) be the average treatment effect among those who would be treated by a physician who treats a fraction u^* of the study population but by no physician who treats less, ie, $\text{MTP}(u^*) = E[Y(1) - Y(0)|X(U^* = u^*) = 1, \{X(U^* = v) = 0; v < u^*\}]$. Heckman and Vytlacil[30] and Angrist et al[31] show that under the assumptions encoded in DAG 2 and continuous monotonicity, MTP (u^*) equals the derivative $\partial\{E[Y|U^* = u^*]\}/\partial u^*$. Thus, were data on U^* available, MTP (u^*) would be identified. In Theorem 6 (see section 5) we show that, regardless of whether data on U^* are available, the estimand (7) based on Z is a particular weighted average of $\partial\{E[Y|U^* = u^*]\}/\partial u^*$ and thus of MTP (u^*). Specifically,

$$\frac{\text{E}[Y|Z=1] - \text{E}[Y|Z=0]}{\text{E}[X|Z=1] - \text{E}[X|Z=0]} = \int \left\{ \frac{\partial}{\partial U^*} \text{E}[Y|U^*] \right\} w(U^*)\, dU^*,$$

$$w(U^*) = \frac{S(U^*|Z=1) - S(U^*|Z=0)}{\int_{I_{low}}^{I_{up}} \{S(U^*|Z=1) - S(U^*|Z=0)\} dU^*} = \frac{S(U^*|Z=1) - S(U|Z=0)}{\text{E}[U^*|Z=1] - \text{E}[U^*|Z=0]}$$

where $S(\cdot)$ is the survival function.

4. Extensions

4.1 SMM With Covariates

We now present a more general additive SMM that allows for continuous or multivariate exposures X, instruments Z, and preinstrument covariates C. A general additive SMM assumes

$$\text{E}[Y|X, Z, C] - \text{E}[Y_0|X, Z, C] = \gamma\{X, Z, C, \psi^*\}$$

where $\gamma\{X, Z, C, \psi\}$ is a known function, ψ^* is an unknown parameter vector and $\gamma\{0, Z, C, \psi\} = \gamma\{X, Z, C, 0\} = 0$. That is, an additive SMM is a model for the average causal effect of treatment level X compared with baseline level 0 among the subset of subjects at level Z of the instrument and level C of the confounders whose observed treatment is precisely X.

We turn next to identification and estimation of the parameters of a general additive SMM under the conditional counterfactual mean independence assumption

$$E[Y(0)|Z = 1, C] = E[Y(0)|Z = 0, C] \qquad (12)$$

Now according to the model $E[Y(0)|X, Z, C] = E[Y - \gamma\{X, Z, C, \psi^*\}|X, Z, C]$. Hence, averaging over X within levels of (Z, C), we have $E[Y(0)|Z, C] = E[Y - \gamma\{X, Z, C, \psi^*\}|Z, C]$. Thus, by the assumed counterfactual mean independence assumption (12),

$$E[Y - \gamma\{X, Z, C, \psi^*\}|Z, C] = E[Y - \gamma\{X, Z, C, \psi^*\}|C]$$

This implies that $\sum_{i=1}^{n} U_i(\psi)$ has mean zero when $\psi = \psi^*$ with

$$U(\psi) = [Y - \gamma\{X, Z, C, \psi\}]\, b(C)\, (Z - E[Z|C])$$

where $b(C)$ is a user supplied vector function of C of the dimension of ψ^* (as one needs one equation per unknown parameter). Thus we would expect that the solution $\hat{\psi}$ to $\sum_{i=1}^{n} U_i(\psi) = 0$ will be consistent and asymptotically normal for ψ^* provided the square matrix $E[\partial U(\psi)/\partial \psi^T]$ of expected partial derivatives is invertible, which can only happen when ψ^* is identified. Conditions for identification are discussed by Robins[10] and in Section 2 for the special case of X and Z dichotomous and C absent. Note in a randomized trial $E[Z|C]$ will be a known function of the randomization probabilities. In most trials, $E[Z|C] = 1/2$ for all C. In observational studies $E[Z|C]$ will have to be estimated from the data, often by regression. The estimator given here is neither efficient nor doubly robust. Chamberlain[32] and Robins[10] discuss efficient estimators. Robins[33] discusses doubly robust estimators. G-estimation of nested additive and multiplicative SMMs extend the aforementioned IV methods for time-independent treatments to time-dependent treatments with time-varying confounders.[10]

Analogously, a more general multiplicative SMM assumes $E[Y|X, Z, C] = \text{E}[Y_0|X, Z, C]\exp(\gamma\{X, Z, C, \psi^*\})$ where $\gamma\{X,Z,C,\psi\}$ is a known function and $\gamma\{0,Z,C,\psi\}) = \gamma\{X,Z,C,0\} = 0$. Estimation proceeds as for an additive SMM

except $U(\psi)$ is redefined to be $Y \exp[-\gamma\{X,Z,C,\psi\}] \times b(C)(Z - E[Z|C])$.

4.2 Causal DAGs Without Counterfactuals

Dawid[6,34] has strenuously argued that any results obtained using counterfactual causal models that cannot also be obtained using causal DAG models without counterfactuals are suspect. He particularly criticized instrumental variable methods that obtain identification of the effect of treatment in the compliers by assuming monotonicity. He argued that joint counterfactuals such as $X(u^* = 0)$ and $X(u^* = 1)$ are not well defined. He therefore concluded that compliers are not a well-defined subset of the population and thus it is meaningless to speak of the causal effect among compliers. However he claimed that important instrumental variables results could still be obtained without counterfactuals and he backed up this claim by rederiving without counterfactuals the bounds for the average treatment effect that Balke and Pearl[13] had previously derived under a counterfactual model. This leaves unanswered the question of whether the identifying assumptions of no effect modification by Z within all levels of treatment X can be meaningfully expressed in a causal DAG model without counterfactuals. Elsewhere we show that it can be.

5. Theorems and Proofs

Theorem 1

Given Z and X dependent and Equations 5 and 1, a) Equations 8 hold $\Leftrightarrow$ Equations 6 hold, and b) both Equations 8 and 6 imply Equation 4 and thus 7

Proof.

a) $\Rightarrow$Let $Y(1)-Y(0)=\Delta$. $E[\Delta|X=1,Z]=E[\Delta|X=0,Z]$

$\Rightarrow E[\Delta|X=1,Z]=E[\Delta|X=0,Z]=E[\Delta|Z]=E[\Delta]$ where the last equality uses Equations 5 and 1

a) $\Leftarrow$Conversely define $\pi(Z) = E[X|Z]$.

Then $E[\Delta]=E[\Delta|Z=1]$

$=E[\Delta|X=1,Z=0]\pi(0)+E[\Delta|X=0,Z=0]\{1-\pi(0)\}$

$=E[\Delta|X=1,Z=1]\pi(1)+E[\Delta|X=0,Z=1]\{1-\pi(1)\}$

$=E[\Delta|X=1,Z=0]\pi(1)+E[\Delta|X=0,Z=0]\{1-\pi(1)\}$

where the last equality is by the premise Eqs (6).

Thus $\{E[\Delta|X=1,Z=0]-E[\Delta|X=0,Z=0]\}\pi(0)+E[\Delta|X=0,Z=0]$

$=\{E[\Delta|X=1,Z=0]-E[\Delta|X=0,Z=0]\}\pi(1)+E[\Delta|X=0,Z=0]$.

Thus $0=\{E[\Delta|X=1,Z=0]-E[\Delta|X=0,Z=0]\}\{\pi(1)-\pi(0)\}$. Since $\{\pi(1) - \pi(0)\} \neq 0$ by assumption, we conclude $E[\Delta|X=1,Z=1]=E[\Delta|X=0,Z=1]$. A symmetric argument shows $E[\Delta|X=1,Z=0]=E[\Delta|X=0,Z=0]$

b) From the proof of a) $\Rightarrow$ above, $E[\Delta|X=1,Z]=E[\Delta|X=0,Z]$ $=E[\Delta]$. Hence $E[\Delta] = E[\Delta|X=1]=E[\Delta|X=0]$ ■

Theorem 2

Consider an NPSEM represented by the DAG in Figure 2. $E[Y(1) - Y(0)|X = x, Z = z]$ does not depend on Z if, with probability 1, either (*i*) $Y(x = 1, U) - Y(x = 0, U)$ does not depend on U or (*ii*) $\dfrac{\Pr[X(z = 1,U) = x]}{\Pr[X(z = 0,U) = x]}$ does not depend on U.

Proof.

$E[Y(1) - Y(0)|X,Z] = \int E[Y(1) - Y(0)|X,Z,U]dF(U|X,Z)$.

But $E[Y(x)|X,Z,U] = E[f_y(x,U,\varepsilon_y)|X,Z,U]$

$= E[f_y(x,U,\varepsilon_y)|U] = E[Y(x,U)|U]$. Hence if (*i*) holds, $E[Y(1) - Y(0)|X = x, Z = z] = E[Y(1) - Y(0)]$ since $\int dF(U|X,Z)=1$.

If (*ii*) holds and $X = x$,

$$f(U|X,Z) = \frac{f(X|U,Z)f(U|Z)}{\int f(X|U,Z)dF(U|Z)} = \frac{f(X|U,Z)f(U)}{\int f(X|U,Z)dF(U)}$$

$$= \frac{\dfrac{f(X|U,Z)}{f(X|U,Z=0)}f(X|U,Z=0)f(U)}{\displaystyle\int \frac{f(X|U,Z)}{f(X|U,Z=0)}f(X|U,Z=0)dF(U)} = \frac{f(X|U,Z=0)f(U)}{\int f(X|U,Z=0)dF(U)}$$

which does not depend on Z since, under the NPSEM, $f(x|U,Z = z) = \Pr[X(z, U) = x]$. But $E[Y(1) - Y(0)|X,Z] = E[Y(1, U) - Y(1, U)|U]$ also does not depend on Z. ■

Theorem 3

On the NPSEM represented by the DAG in Figure 2, suppose U and X are dependent given Z. Then $\dfrac{\Pr[X(z = 1,U) = x]}{\Pr[X(z = 0,U) = x]}$ depends on U for either $x = 1$ or $x = 0$.

Proof.

By contradiction. Assume the lemma is false. Let $\dfrac{\Pr[X(z = 1,U) = x]}{\Pr[X(z = 0,U) = x]} = r(x)$.

Then $r(x) = \dfrac{1 - r(1-x)\Pr[X(z=0,U)=1-x]}{1 - \Pr[X(z = 0,U) = 1-x]}$

Hence $\Pr[X(z = 0,U) = 1 - x] = \dfrac{1 - r(x)}{[r(1 - x) - 1]}$. So $\Pr[X = 1|Z = 0,U] = \Pr[X = 1|Z = 0]$.

By symmetry $\dfrac{1}{r(x)} = \dfrac{1 - \Pr[X(z = 0,U) = 1 - x]}{1 - \Pr[X(z = 1,U) = 1 - x]}$

$= \dfrac{1 - \dfrac{1}{r(1 - x)}\Pr[X(z = 1,U) = 1 - x]}{1 - \Pr[X(z = 1,U) = 1 - x]}$ so $\Pr[X=1|Z=1, U] = \Pr[X = 1|Z = 1]$. Hence U and X are independent given Z, which is a contradiction. ■

Theorem 4 (*Robins 1989*)[8]

Assume Z and X are dichotomous and dependent, and Equation 1 holds. Further assume model (9) holds with $\psi_1^* = 0$, ie, no multiplicative effect modification by Z in the treated. Then ψ_0^* is identified and

$$\exp(-\psi_0^*) = 1 - \frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[Y|X=1, Z=1]\,\mathrm{E}[X|Z=1] - \mathrm{E}[Y|X=1, Z=0]\,\mathrm{E}[X|Z=0]} \quad (13)$$

If, in addition, Equation 5 holds and there is no multiplicative effect modification by Z in the untreated, ie,

$$\frac{\mathrm{E}[Y(1)|X=0, Z=1]}{\mathrm{E}[Y(0)|X=0, Z=1]} = \frac{\mathrm{E}[Y(1)|X=0, Z=0]}{\mathrm{E}[Y(0)|X=0, Z=0]} \quad (14)$$

then $\mathrm{E}[Y(0)]$,$\mathrm{E}[Y(1)]$,$\mathrm{E}[Y(1)]/\mathrm{E}[Y(0)]$ and $\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)]$ are identified.

$\mathrm{E}[Y(1)]/\mathrm{E}[Y(0)]=\exp(\psi_0^*)$,
$\mathrm{E}[Y(0)]=\mathrm{E}[Y|X=0]\{1-\mathrm{E}[X]\}+\mathrm{E}[X]\mathrm{E}[Y|X=1]\exp(-\psi_1^*)$,
$\mathrm{E}[Y(1)]=\mathrm{E}[Y|X=0]\{1-\mathrm{E}[X]\}\exp(\psi_0^*)+\mathrm{E}[X]\mathrm{E}[Y|X=1]$,
and $\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)]$
$=\mathrm{E}[Y|X=0]\{1-\mathrm{E}[X]\}[\exp(\psi_0^*)-1]+\mathrm{E}[X]\mathrm{E}[Y|X=1]$

Proof.
From (9), $\mathrm{E}[Y\exp\{-X\{\psi_0^* + \psi_1^* Z\}\}|X, Z] = \mathrm{E}[Y(0)|X, Z]$ and thus $\mathrm{E}[Y\exp\{-X\{\psi_0^* + \psi_1^* Z\}\}|Z]=\mathrm{E}[Y(0)|Z]=\mathrm{E}[Y(0)]$. Therefore $\mathrm{E}[Y\exp\{-X\{\psi_0^*+\psi_1^*\}\}|Z=1]=\mathrm{E}[Y\exp\{-X\{\psi_0^*+\psi_1^* Z\}\}|Z=0]$ and hence
$\mathrm{E}[Y\exp\{-X\{\psi_0^*+\psi_1^*\}\}|Z=1]$
$=\mathrm{E}[Y\exp\{-X\psi_0^*\}|Z=0]$. Putting $\psi_1^*=0$ we obtain
$\mathrm{E}[Y\exp\{-X\psi_0^*\}|Z=1] = \mathrm{E}[Y\exp\{-X\psi_0^*\}|Z=0]$.
So $[\exp(-\psi_0^*)-1]\mathrm{E}[YX|Z=1]+\mathrm{E}[Y|Z=1]$
$=[\exp(-\psi_0^*)-1]\mathrm{E}[YX|Z=0]+\mathrm{E}[Y|Z=0]$
from which (13) follows.
Now by $\psi_1^*=0$, $\mathrm{E}[Y(0)|X=1]=\mathrm{E}[Y|X=1]\exp(-\psi_0^*)$.
Hence $\mathrm{E}[Y(0)]=\mathrm{E}[Y|X=0]\{1-\mathrm{E}[X]\}+\mathrm{E}[X]\mathrm{E}[Y|X=1]\exp(-\psi_0^*)$. By $\psi_1^*=0$ and (14),$\mathrm{E}[Y(1)]/\mathrm{E}[Y(0)]=\exp(\psi_0^*)$, allowing us to calculate $\mathrm{E}[Y(1)]$ and thus $\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)]$ ■

Theorem 5

Suppose we have an NPSEM represented by the DAG in Figure 2. Further assume the causal instrument U* is binary and that the following monotonicity assumption holds

$$X(u^*=0) = 1 \text{ implies } X(u^*=1) = 1$$

Define the compliers to be subjects for whom $X(u^*=0)=0$, $X(u^*=1)=1$. Then the average causal effect in the compliers $\mathrm{E}[Y(x=1) - Y(x=0)|\,X(u^*=0)=0, X(u^*=1)=1]$ is identified from the data (X, Z, Y) and equals the ratio (7).

Proof.
$\mathrm{E}[Y|Z=1]-\mathrm{E}[Y|Z=0]$
$=\mathrm{E}[Y|U^*=1,Z=1]\mathrm{E}[U^*=1|Z=1]+\mathrm{E}[Y|U^*=0,Z=1]\{1-\mathrm{E}[U^*=1|Z=1]\}-\mathrm{E}[Y|U^*=1,Z=0]\mathrm{E}[U^*=1|Z=0]+\mathrm{E}[Y|U^*=0,Z=0]\{1-\mathrm{E}[U^*=1|Z=0]\}=$
$\mathrm{E}[Y|U^*=1]\mathrm{E}[U^*=1|Z=1]+\mathrm{E}[Y|U^*=0]\{1-\mathrm{E}[U^*=1|Z=1]\}-\mathrm{E}[Y|U^*=1]\mathrm{E}[U^*=1|Z=0]+\mathrm{E}[Y|U^*=0]\{1-\mathrm{E}[U^*=1|Z=0]\}$
$=\{\mathrm{E}[Y|U^*=1]-\mathrm{E}[Y|U^*=0]\}\{\mathrm{E}[U^*=1|Z=1]-\mathrm{E}[U^*=1|Z=0]\}$.
Similarly,
$\mathrm{E}[X|Z=1]-\mathrm{E}[X|Z=0]$
$=[\mathrm{E}[X|U^*=1]-\mathrm{E}[X|U^*=0]]\{\mathrm{E}[U^*=1|Z=1]-\mathrm{E}[U^*=1|Z=0]\}$.
Thus

$$\frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[X|Z=1] - \mathrm{E}[X|Z=0]} = \frac{\mathrm{E}[Y|U^*=1] - \mathrm{E}[Y|U^*=0]}{\mathrm{E}[X|U^*=1] - \mathrm{E}[X|U^*=0]}.$$

The theorem then follows from Imbens and Angrist. [15] ■

Theorem 6

Suppose the NPSEM represented by the DAG in Figure 2 and the monotonicity assumption for continuous U^* hold, that $\Pr(X=1|U^*) = U^*$, and that $\mathrm{E}[Y|U^*]$ is differentiable on the support $[I_{low},I_{up}] \subseteq [0,1]$ of U^*. Then

$$\frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[X|Z=1] - \mathrm{E}[X|Z=0]} = \int \left\{ \frac{\partial}{\partial U^*} \mathrm{E}[Y|U^*] \right\} w(U^*)\, dU^*,$$

$$w(U^*) = \frac{S(U^*|Z=1) - S(U^*|Z=0)}{\int_{I_{low}}^{I_{up}} \{S(U^*|Z=1) - S(U^*|Z=0)\}\, dU^*} = \frac{S(U^*|Z=1) - S(U^*|Z=0)}{\mathrm{E}[U^*|Z=1] - \mathrm{E}[U^*|Z=0]}$$

Proof.[30,31]

$$\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0] = \int_{I_{low}}^{I_{up}} \mathrm{E}[Y|U^*]\{f(U^*|Z=1) - f(U^*|Z=0)\} =$$
$$\mathrm{E}[Y|U^*]\{F(U^*|Z=1) - F(U^*|Z=0)\}|_{I_{low}}^{I_{up}} - \int_{I_{low}}^{I_{up}} \left\{ \frac{\partial}{\partial U^*} \mathrm{E}[Y|U^*] \right\} \{F(U^*|Z=1) - F(U^*|Z=0)\}\, dU^* =$$
$$\int_{I_{low}}^{I_{up}} \left\{ \frac{\partial}{\partial U^*} \mathrm{E}[Y|U^*] \right\} \{S(U^*|Z=1) - S(U^*|Z=0)\}\, dU^*$$

Similarly,

$$\mathrm{E}[X|Z=1] - \mathrm{E}[X|Z=0] = \int_{I_{low}}^{I_{up}} \mathrm{E}[X|U^*][f(U^*|Z=1) - f(U^*|Z=0)] =$$
$$\int_{I_{low}}^{I_{up}} \mathrm{U}^*[f(U^*|Z=1) - f(U^*|Z=0)] =$$
$$\mathrm{E}[U^*|Z=1] - \mathrm{E}[U^*|Z=0] =$$

$$\int_{I_{low}}^{I_{up}} \{S(U^*|Z = 1) - S(U^*|Z = 0)\}\, dU^* \blacksquare$$

REFERENCES

1. Martens E, Pestman W, de Boer A, et al. Instrumental variables: applications and limitations. *Epidemiology.* 2006;17:260–267.
2. Brookhart MA, Wang P, Solomon DH, et al. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology.* 2006;17:268–275.
3. Heckman J, Robb R. Alternative methods for estimating the impact of interventions. In: Heckman J, Singer B, eds. *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press; 1985: 156–245.
4. Pearl J. *Causality: Models, Reasoning, and Inference.* New York: Cambridge University Press; 2000.
5. Spirtes P, Glymour C, Scheines R. *Causation, Prediction and Search.* 2nd ed. Cambridge, MA: MIT Press; 2000.
6. Dawid AP. Causal inference using influence diagrams: the problem of partial compliance. In: Green PJ, Hjort NL, Richardson S, eds. *Highly Structured Stochastic Systems*. New York: Oxford University Press; 2003.
7. Holland PW. Causal inference, path analysis, and recursive structural equation models. In: Clogg C (ed). Sociological Methodology. Washington, DC: American Sociological Association; 1988:449–484.
8. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Services Research Methodology: A Focus on AIDS.* NCHRS, U.S. Public Health Service; 1989:113–59.
9. Angrist J, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91:444–455.
10. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat*. 1994;23:2379–412.
11. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29:722–729.
12. Manski C. Nonparametric bounds on treatment effects. *Am Econ Rev*. 1990;80:319–323.
13. Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. *J Am Stat Assoc*. 1997;92:1171–1176.
14. Heckman J. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *J Human Resources*. 1997;32:441–462.
15. Imbens GW, Angrist J. Identification and estimation of local average treatment effects. *Econometrica*. 1994;62:467–475.
16. Robins J, Greenland S. Comment on "Identification of causal effects using instrumental variables" by Angrist, Imbens and Rubin. *J Am Stat Assoc*. 1996;91:456–8.
17. Robins JM. Analytic methods for estimating HIV treatment and cofactor effects. In: Ostrow DG, Kessler R, eds. *Methodological Issues of AIDS Mental Health Research*. New York: Plenum Publishing; 1993:213–290.
18. Robins JM. Optimal structural nested models for optimal sequential decisions. In: Lin DY, Heagerty P, eds. *Proceedings of the Second Seattle Symposium on Biostatistics*. New York: Springer; 2003.
19. Robins JM. Comment on "Covariance adjustment in randomized experiments and observational studies" by Paul Rosenbaum. *Stat Sci*. 2002; 17:286–327.
20. Robins JM, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*. 2004;91:763–783.
21. Van der Laan MJ, Hubbard A, Jewell N. Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome. UC Berkeley Division of Biostatistics Working Paper Series 2004; Working Paper 157.
22. Tan Z. Estimation of causal effects using instrumental variables. *J Am Stat Assoc.* in press.
23. Robins JM, Greenland S. Comment on "Causal inference without counterfactuals" by A.P. Dawid. *J Am Stat Assoc*. 2000;95:431–5.
24. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods–Application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986;7:1393–1512 (errata in *Computers and Mathematics with Applications.* 1987;14: 917–921.
25. Robins JM. Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods." *Computers and Mathematics with Applications.* 1987;14:923–945 (errata in *Computers and Mathematics with Applications.* 1987;18:477.
26. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green P, Hjort NL, Richardson S, eds. *Highly Structured Stochastic Systems*. New York: Oxford University Press; 2003:70–81.
27. Heckman J, Robb R. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In: Wainer H, ed. *Drawing Inferences from Self-Selected Samples*. Berlin: Springer Verlag; 1986.
28. Heckman J. Randomization and Social Policy Evaluation. Technical Working Paper 107. National Bureau of Economic Research; 1991.
29. Angrist J, Imbens GW, Rubin DB. Rejoinder to comments on "Identification of causal effects using instrumental variables." *J Am Stat Assoc*. 1996;91:468–472.
30. Heckman JJ, Vytlacil EJ. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proc Natl Acad Sci USA*. 1999;96:4730–4734.
31. Angrist JD, Graddy K, Imbens GW. The interpretation of instrumental variable estimators in simultaneous equations models with an application to the demand for fish. *Rev Econ Stud*. 2000;67:499–527.
32. Chamberlain G. Asymptotic efficiency in estimation with conditional moment estrictions. *J Econometrics*. 1987;34:305–334.
33. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: 1999 *Proceedings of the Section on Bayesian Statistical Science.* Alexandria, VA: American Statistical Association; 2000:6:–10.
34. Dawid AP. Causal inference without counterfactuals. *J Am Stat Assoc*. 2000;95:407–424.