Biometrika (2017), **104**, 2, *pp.* 291–302
Printed in Great Britain

doi: 10.1093/biomet/asx009
Advance Access publication 17 April 2017

Instrumental variables as bias amplifiers with general outcome and confounding

By P. DING

Department of Statistics, University of California, 425 Evans Hall, Berkeley,
California 94720, U.S.A.
pengdingpku@berkeley.edu

T. J. VANDERWEELE and J. M. ROBINS

Department of Epidemiology, Harvard T. H. Chan School of Public Health, 677 Huntington
Avenue, Boston, Massachusetts 02115, U.S.A.
tvanderw@hsph.harvard.edu robins@hsph.harvard.edu

Summary

Drawing causal inference with observational studies is the central pillar of many disciplines. One sufficient condition for identifying the causal effect is that the treatment-outcome relationship is unconfounded conditional on the observed covariates. It is often believed that the more covariates we condition on, the more plausible this unconfoundedness assumption is. This belief has had a huge impact on practical causal inference, suggesting that we should adjust for all pretreatment covariates. However, when there is unmeasured confounding between the treatment and outcome, estimators adjusting for some pretreatment covariate might have greater bias than estimators that do not adjust for this covariate. This kind of covariate is called a bias amplifier, and includes instrumental variables that are independent of the confounder and affect the outcome only through the treatment. Previously, theoretical results for this phenomenon have been established only for linear models. We fill this gap in the literature by providing a general theory, showing that this phenomenon happens under a wide class of models satisfying certain monotonicity assumptions.

Some key words: Causal inference; Directed acyclic graph; Interaction; Monotonicity; Potential outcome.

1. Introduction

Causal inference from observational data is an important but challenging problem for empirical studies in many disciplines. Under the potential outcomes framework (Neyman, 1923 [1990]; Rubin, 1974), the causal effects are defined as comparisons between the potential outcomes under treatment and control, averaged over a certain population of interest. One sufficient condition for nonparametric identification of the causal effects is the ignorability condition (Rosenbaum & Rubin, 1983), that the treatment is conditionally independent of the potential outcomes given those pretreatment covariates that confound the relationship between the treatment and outcome. To make this fundamental assumption as plausible as possible, many researchers suggest that the set of collected pretreatment covariates should be as rich as possible. It is often believed that

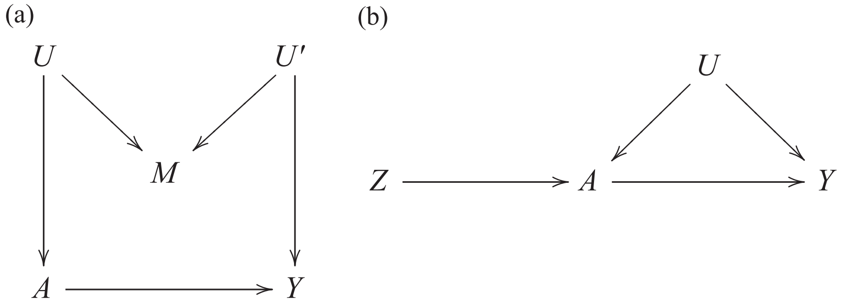

Fig. 1. Two directed acyclic graphs, where A is the treatment and Y is the outcome of interest. (a) Directed acyclic graph for M-bias, where U and U' are unobserved and M is observed. (b) Directed acyclic graph for Z-bias, where U is an unmeasured confounder and Z is an instrumental variable for the treatment-outcome relationship.

'typically, the more conditional an assumption, the more generally acceptable it is' (Rubin, 2009), and therefore 'in principle, there is little or no reason to avoid adjustment for a true covariate, a variable describing subjects before treatment' (Rosenbaum, 2002, p. 76).

Simply adjusting for all pretreatment covariates (d'Agostino, 1998; Rosenbaum, 2002; Hirano & Imbens, 2001), sometimes called the pretreatment criterion (VanderWeele & Shpitser, 2011), has a sound justification from the viewpoint of design and analysis of randomized experiments. Cochran (1965), citing Dorn (1953), suggested that the planner of an observational study should always ask himself the question, 'How would the study be conducted if it were possible to do it by controlled experimentation?' Following this classical wisdom, Rubin (2007, 2008a,b, 2009) argued that the design of observational studies should be in parallel with the design of randomized experiments, i.e., because we balance all pretreatment covariates in randomized experiments, we should also follow this pretreatment criterion and balance or adjust for all pretreatment covariates when designing observational studies.

However, this pretreatment criterion can result in increased bias under certain data-generating processes. We highlight two important classes of such processes for which the pretreatment criterion may be problematic. The first class is captured by an example of Greenland & Robins (1986), in which conditioning on a pretreatment covariate invalidates the ignorability assumption and thus a conditional analysis is biased, yet the ignorability assumption holds unconditionally, so an analysis that ignores the covariate is unbiased. Several researchers have shown that this phenomenon is generic when the data are generated under the causal diagram in Fig. 1(a), in which the ignorability assumption holds unconditionally but not conditionally on M (Pearl, 2000; Spirtes et al., 2000; Greenland, 2003; Pearl, 2009; Shrier, 2008, 2009; Sjölander, 2009; Ding & Miratrix, 2015). In Fig. 1(a), a pretreatment covariate M is associated with two independent unmeasured covariates U and U', but M does not itself affect either the treatment A or the outcome Y. Because the corresponding causal diagram looks like the English letter M, this phenomenon is called M-bias.

The second class of processes, which constitute the subject of this paper, are represented by the causal diagram in Fig. 1(b). Owing to confounding by the unmeasured common cause U of the treatment A and the outcome Y, both the analysis that adjusts and the analysis that fails to adjust for pretreatment measured covariates are biased. If the magnitude of the bias is larger when we adjust for a particular pretreatment covariate than when we do not, we refer to the covariate as a bias amplifier. It is of particular interest to determine the conditions under which an instrumental variable is a bias amplifier. An instrumental variable is a pretreatment covariate that is independent of the confounder U and has no direct effect on the outcome except through its effect on the treatment. The variable Z in Fig. 1(b) is an example. Heckman & Navarro-Lozano (2004), Bhattacharya & Vogt (2012) and Middleton et al. (2016) showed numerically that when the treatment and outcome are confounded, adjusting for an instrumental variable can result in greater bias than the unadjusted estimator. Wooldridge theoretically demonstrated this in linear models in a 2006 technical report finally published as Wooldridge (2016). Because instrumental variables are often denoted by Z as in Fig. 1(b), this phenomenon is called Z-bias.

The treatment assignment is a function of the instrumental variable, the unmeasured confounder and some other independent random error, which are the three sources of variation of the treatment. If we adjust for the instrumental variable, the treatment variation is driven more by the unmeasured confounder, which could result in increased bias due to this confounder. Seemingly paradoxically, without adjusting for the instrumental variable, the observational study is more like a randomized experiment, and the bias due to confounding is smaller. Although applied researchers (Myers et al., 2011; Walker, 2013; Brooks & Ohsfeldt, 2013; Ali et al., 2014) have confirmed through extensive simulation studies that this bias amplification phenomenon exists in a wide range of reasonable models, definite theoretical results have been established only for linear models. We fill this gap in the literature by showing that adjusting for an instrumental variable amplifies bias for estimating causal effects under a wide class of models satisfying certain monotonicity assumptions. However, we also show that there exist data-generating processes under which an instrumental variable is not a bias amplifier.

2. Framework and notation

We consider a binary treatment A, an instrumental variable Z, an unobserved confounder U, and an outcome Y, with the joint distribution depicted by the causal diagram in Fig. 1(b). Let $\perp\!\!\!\perp$ denote conditional independence between random variables. Then the instrumental variable Z in Fig. 1(b) satisfies $Z\perp\!\!\!\perp U$, $Z\perp\!\!\!\perp Y \mid (A, U)$ and $Z\not\perp\!\!\!\perp A$. We first discuss analysis conditional on observed pretreatment covariates X, and comment on averaging over X in § 6 and the Supplementary Material. We define the potential outcomes of Y under treatment a as $Y(a)$ $(a = 1, 0)$. The true average causal effect of A on Y for the population actually treated is

$$\text{ACE}_1^{\text{true}} = E\{Y(1) \mid A = 1\} - E\{Y(0) \mid A = 1\};$$

for the population who are actually in the control condition it is

$$\text{ACE}_0^{\text{true}} = E\{Y(1) \mid A = 0\} - E\{Y(0) \mid A = 0\};$$

and for the whole population it is

$$\text{ACE}^{\text{true}} = E\{Y(1)\} - E\{Y(0)\}.$$

Define $m_a(u) = E(Y \mid A = a, U = u)$ to be the conditional mean of the outcome given the treatment and confounder. As illustrated by Fig. 1(b), because U suffices to control confounding between A and Y, the ignorability assumption $A\perp\!\!\!\perp Y(a) \mid U$ holds for $a = 0$ and 1. Therefore, according to $Y = AY(1) + (1 - A)Y(0)$, we have

$$\text{ACE}_1^{\text{true}} = E(Y \mid A=1) - \int m_0(u)F(\mathrm{d}u \mid A=1),$$

$$\text{ACE}_0^{\text{true}} = \int m_1(u)F(\mathrm{d}u \mid A=0) - E(Y \mid A=0),$$

$$\text{ACE}^{\text{true}} = \int m_1(u)F(\mathrm{d}u) - \int m_0(u)F(\mathrm{d}u).$$

The unadjusted estimator is the naive comparison between the treatment and control means,

$$\text{ACE}^{\text{unadj}} = E(Y \mid A=1) - E(Y \mid A=0).$$

Define $\mu_a(z) = E(Y \mid A=a, Z=z)$ as the conditional mean of the outcome given the treatment and instrumental variable. Because the instrumental variable Z is also a pretreatment covariate unaffected by the treatment, the usual strategy to adjust for all pretreatment covariates suggests using the adjusted estimator for the population under treatment

$$\text{ACE}_1^{\text{adj}} = E(Y \mid A=1) - \int \mu_0(z)F(\mathrm{d}z \mid A=1),$$

for the population under control

$$\text{ACE}_0^{\text{adj}} = \int \mu_1(z)F(\mathrm{d}z \mid A=0) - E(Y \mid A=0),$$

and for the whole population

$$\text{ACE}^{\text{adj}} = \int \mu_1(z)F(\mathrm{d}z) - \int \mu_0(z)F(\mathrm{d}z).$$

Surprisingly, for linear structural equation models on (Z,U,A,Y), previous theory demonstrated that the magnitudes of the biases of the adjusted estimators are no smaller than the unadjusted ones (Pearl, 2010, 2011, 2013; Wooldridge, 2016). The goal of the rest of our paper is to show that this phenomenon exists in more general scenarios.

3. Scalar instrumental variable and scalar confounder

We first give a theorem for a scalar instrumental variable Z and a scalar confounder U.

Theorem 1. *In the causal diagram of Fig.* 1(b) *with scalar Z and U, suppose that*

(a) $\text{pr}(A=1 \mid Z=z)$ *is nondecreasing in z,* $\text{pr}(A=1 \mid U=u)$ *is nondecreasing in u, and* $E(Y \mid A=a, U=u)$ *is nondecreasing in u for both $a=0$ and* 1;
(b) $E(Y \mid A=a, Z=z)$ *is nonincreasing in z for both $a=0$ and* 1;

then

$$\begin{pmatrix} \text{ACE}_1^{\text{adj}} \\ \text{ACE}_0^{\text{adj}} \\ \text{ACE}^{\text{adj}} \end{pmatrix} \geqslant \begin{pmatrix} \text{ACE}^{\text{unadj}} \\ \text{ACE}^{\text{unadj}} \\ \text{ACE}^{\text{unadj}} \end{pmatrix} \geqslant \begin{pmatrix} \text{ACE}_1^{\text{true}} \\ \text{ACE}_0^{\text{true}} \\ \text{ACE}^{\text{true}} \end{pmatrix}. \qquad (1)$$

Inequalities among vectors as in (1) should be interpreted as componentwise relationships. Intuitively, the monotonicity in Condition (a) of Theorem 1 requires nonnegative dependence structures on arrows $Z \to A$, $U \to A$ and $U \to Y$ in the causal diagram of Fig. 1(b).

The monotonicity in Condition (b) of Theorem 1 reflects the collider bias caused by conditioning on A. As noted by Greenland (2003), if Z and U affect A in the same direction, then the collider bias caused by conditioning on A is often, although not always, in the opposite direction. Lemmas S5–S8 in the Supplementary Material show that if Z and U are independent and have nonnegative additive or multiplicative effects on A, then conditioning on A results in negative association between Z and U. This negative collider bias, coupled with the positive association between U and Y, further implies negative association between Z and Y conditional on A as stated in Condition (b) of Theorem 1.

For easy interpretation, we will give sufficient conditions for Z-bias when there is no interaction of Z and U on A. When A given Z and U follows an additive model, we have the following theorem.

THEOREM 2. *In the causal diagram of Fig.* 1(b) *with scalar Z and U,* (1) *holds if*

(a) $\text{pr}(A = 1 \mid Z = z, U = u) = \beta(z) + \gamma(u)$;

(b) *$\beta(z)$ is nondecreasing in z, $\gamma(u)$ is nondecreasing in u, and $E(Y \mid A = a, U = u)$ is nondecreasing in u for both $a = 1$ and 0;*

(c) *the essential supremum of U given $(A = a, Z = z)$ depends only on a.*

In summary, when A given Z and U follows an additive model and monotonicity of Theorem 2 holds, both unadjusted and adjusted estimators have nonnegative biases for the true average causal effects for the treatment, control and whole population. Furthermore, the adjusted estimators, for either the treatment, the control or the whole population, have larger biases than the unadjusted estimator, i.e., Z-bias arises.

When both the instrumental variable Z and the confounder U are binary, Theorem 2 has an even more interpretable form. Define $p_{zu} = \text{pr}(A = 1 \mid Z = z, U = u)$ for $z, u = 0$ and 1.

COROLLARY 1. *In the causal diagram of Fig.* 1(b) *with binary Z and U,* (1) *holds if*

(a) *there is no additive interaction of Z and U on A, i.e., $p_{11} - p_{10} - p_{01} + p_{00} = 0$;*

(b) *Z and U have monotonic effects on A, i.e., $p_{11} \geqslant \max(p_{10}, p_{01})$ and $\min(p_{10}, p_{01}) \geqslant p_{00}$, and $E(Y \mid A = a, U = 1) \geqslant E(Y \mid A = a, U = 0)$ for both $a = 1$ and 0.*

When A given Z and U follows a multiplicative model, we have the following theorem.

THEOREM 3. *In the causal diagram of Fig.* 1(b) *with scalar Z and U,* (1) *holds if we replace Condition* (a) *of Theorem* 2 *by*

(a′) $\text{pr}(A = 1 \mid Z = z, U = u) = \beta(z)\gamma(u)$.

When both the instrument Z and the confounder U are binary, Theorem 3 can be simplified.

COROLLARY 2. *In the causal diagram of Fig.* 1(b) *with binary Z and U,* (1) *holds if we replace Condition* (a) *of Corollary* 1 *by*

(a′) *there is no multiplicative interaction of Z and U on A, i.e., $p_{11}p_{00} = p_{10}p_{01}$.*

We invoke the assumptions of no additive and multiplicative interaction of Z and U on A in Theorems 2 and 3 for easy interpretation. They are sufficient but not necessary conditions for

Z-bias. In fact, we show in the proofs that Conditions (a) and (a′) in Theorems 2 and 3 and Corollaries 1 and 2 can be replaced by weaker conditions. For the case with binary Z and U, these conditions are particularly easy to interpret:

$$\frac{p_{11}p_{00}}{p_{10}p_{01}} \leqslant 1, \quad \frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} \leqslant 1, \tag{2}$$

i.e., Z and U have nonpositive multiplicative interaction on both the presence and absence of A. Even if Condition (a) or (a′) does not hold, one can show that half of the parameter space of $(p_{11}, p_{10}, p_{01}, p_{00})$ satisfies the weaker condition (2), which is only sufficient, not necessary. Therefore, even in the presence of additive or multiplicative interaction, Z-bias arises in more than half of the parameter space for binary (Z, U, A, Y).

4. GENERAL INSTRUMENTAL VARIABLE AND GENERAL CONFOUNDER

When the instrumental variable Z and the confounder U are vectors, Theorems 1–3 still hold if the monotonicity assumptions hold for each component of Z and U, and Z and U are multivariate totally positive of order two (Karlin & Rinott, 1980), including the case where the components of Z and U are mutually independent (Esary et al., 1967). A random vector W is multivariate totally positive of order two if its density $f(\cdot)$ satisfies $f\{\max(w_1,w_2)\}f\{\min(w_1,w_2)\} \geqslant f(w_1)f(w_2)$, where $\max(w_1,w_2)$ and $\min(w_1,w_2)$ are the componentwise maximum and minimum of the vectors w_1 and w_2. In the following, we will develop general theory for Z-bias without the total positivity assumption about the components of Z and U.

It is relatively straightforward to summarize a general instrumental variable Z by a scalar propensity score $\Pi = \Pi(Z) = \text{pr}(A = 1 \mid Z)$, because $Z \perp\!\!\!\perp A \mid \Pi(Z)$, as shown in Rosenbaum & Rubin (1983). We define $\nu_a(\pi) = E(Y \mid A = a, \Pi = \pi)$. The adjusted estimator for the population under treatment is

$$\text{ACE}_1^{\text{adj}} = E(Y \mid A = 1) - \int \nu_0(\pi)F(\mathrm{d}\pi \mid A = 1),$$

the adjusted estimator for the population under control is

$$\text{ACE}_0^{\text{adj}} = \int \nu_1(\pi)F(\mathrm{d}\pi \mid A = 0) - E(Y \mid A = 0),$$

and the adjusted estimator for the whole population is

$$\text{ACE}^{\text{adj}} = \int \nu_1(\pi)F(\mathrm{d}\pi) - \int \nu_0(\pi)F(\mathrm{d}\pi).$$

When Z is scalar and $\Pi(Z)$ is monotone in Z, then the above three formulas reduce to the ones in § 3.

Greenland & Robins (1986) showed that for the causal effect on the treated population, $Y(0)$ alone suffices to control for confounding; likewise, for the causal effect on the control population, $Y(1)$ alone suffices to control for confounding. If interest lies in all three of our average causal effects, then we take $U = \{Y(1), Y(0)\}$ as the ultimate confounder for the relationship of A on Y. Because $Y = AY(1) + (1-A)Y(0)$ is a deterministic function of A and $\{Y(1), Y(0)\}$, this implies that $U = \{Y(1), Y(0)\}$ satisfies the ignorability assumption (Rosenbaum & Rubin, 1983), or

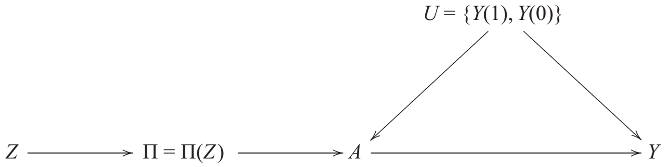

Fig. 2. Directed acyclic graph for Z-bias with general instrument and confounder.

blocks all the back-door paths from A to Y (Pearl, 1995, 2000). We represent the causal structure in Fig. 2.

We first state a theorem without assuming the structure of the causal diagram in Fig. 2.

THEOREM 4. *If for both* $a = 1$ *and* 0, $\text{pr}\{A = 1 \mid Y(a)\}$ *is nondecreasing in* $Y(a)$, *and* $\text{cov}\{\Pi, \nu_a(\Pi)\} \leqslant 0$, *then* (1) *holds.*

In a randomized experiment $A \perp\!\!\!\perp Y(a)$, so the dependence of $\text{pr}\{A = 1 \mid Y(a)\}$ on $Y(a)$ characterizes the self-selection process of an observational study. The condition $\text{cov}\{\Pi, \nu_a(\Pi)\} \leqslant 0$ in Theorem 4 is another measure of the collider-bias caused by conditioning on A, as $\nu_a(\pi) = E\{Y(a) \mid A = a, \Pi = \pi\}$ and $Y(a)$ is a component of U in Fig. 2. This measure of collider bias is more general than the one in Theorem 1. Analogous to § 3, we will present more transparent sufficient conditions for Z-bias to aid interpretation.

In the following, we use the distributional association measure (Cox & Wermuth, 2003; Ma et al., 2006; Xie et al., 2008), i.e., a random variable V has a nonnegative distributional association on a random variable W if the conditional distribution satisfies $\partial F(w \mid v)/\partial v \leqslant 0$ for all v and w. If the random variables are discrete, then partial differentiation is replaced by differencing between adjacent levels (Cox & Wermuth, 2003).

If there is no additive interaction between Π and $\{Y(1), Y(0)\}$ on A, then we have the following results.

THEOREM 5. *In the causal diagram of Fig.* 2, (1) *holds if*

(a) $\text{pr}(A = 1 \mid \Pi, U) = \Pi + \delta\{Y(1)\} + \eta\{Y(0)\}$ *with* $\delta(\cdot)$ *and* $\eta(\cdot)$ *nondecreasing;*
(b) $\{Y(1), Y(0)\}$ *have nonnegative distributional associations on each other, i.e.,* $\partial F(y_1 \mid y_0)/\partial y_0 \leqslant 0$ *and* $\partial F(y_0 \mid y_1)/\partial y_1 \leqslant 0$ *for all* y_1 *and* y_0*;*
(c) *the essential supremum of* $Y(1)$ *given* $Y(0)$ *does not depend on* $Y(0)$, *and the essential supremum of* $Y(0)$ *given* $Y(1)$ *does not depend on* $Y(1)$.

Remark 1. If we impose an additive model $\text{pr}(A = 1 \mid \Pi, U) = h(\Pi) + \delta\{Y(1)\} + \eta\{Y(0)\}$, then independence of Π and U implies that $\text{pr}(A = 1 \mid \Pi) = h(\Pi) + E[\delta\{Y(1)\}] + E[\eta\{Y(0)\}] = \Pi$. Therefore, we must have $h(\Pi) = \Pi$ and $E[\delta\{Y(1)\}] + E[\eta\{Y(0)\}] = 0$.

When the outcome is binary, the distributional association between $Y(1)$ and $Y(0)$ becomes their odds ratio (Xie et al., 2008), and nonnegative distributional association between $Y(1)$ and $Y(0)$ is equivalent to

$$\text{OR}_Y = \frac{\text{pr}\{Y(1) = 1, Y(0) = 1\}\text{pr}\{Y(1) = 0, Y(0) = 0\}}{\text{pr}\{Y(1) = 1, Y(0) = 0\}\text{pr}\{Y(1) = 0, Y(0) = 1\}} \geqslant 1.$$

We can further relax the model assumption of A given Π and U by allowing for nonnegative interaction between $Y(1)$ and $Y(0)$ on A.

COROLLARY 3. *In the causal diagram of Fig.* 2 *with a binary outcome* Y*,* (1) *holds if*

(a) $\text{pr}(A = 1 \mid \Pi, U) = \alpha + \Pi + \delta Y(1) + \eta Y(0) + \theta Y(1)Y(0)$ *with* $\delta, \eta, \theta \geqslant 0$*;*
(b) $\text{OR}_Y \geqslant 1$.

Remark 2. If we have an additive model of A given Π and U, $\text{pr}(A = 1 \mid \Pi, U) = h(\Pi) + g(U)$, then the functional form $g(U) = \alpha + \delta Y(1) + \eta Y(0) + \theta Y(1)Y(0)$ imposes no restriction for a binary outcome. Furthermore, $\text{pr}(A = 1 \mid \Pi) = \Pi$ implies that $h(\Pi) = \Pi$ and $E\{g(U)\} = 0$, i.e., $\alpha = -\delta E\{Y(1)\} - \eta E\{Y(0)\} - \theta E\{Y(1)Y(0)\}$. Therefore, the additive model in Condition (a) of Corollary 3 is

$$\text{pr}(A = 1 \mid \Pi, U) = \Pi + \delta[Y(1) - E\{Y(1)\}] + \eta[Y(0) - E\{Y(0)\}] + \theta[Y(1)Y(0) - E\{Y(1)Y(0)\}].$$

If there is no multiplicative interaction of Π and $\{Y(1), Y(0)\}$ on Z, then we have the following results.

THEOREM 6. *In the causal diagram of Fig.* 2*,* (1) *holds if we replace Condition* (a) *of Theorem* 5 *by* $\text{pr}(A = 1 \mid \Pi, U) = \Pi\delta\{Y(1)\}\eta\{Y(0)\}$ *with* $\delta(\cdot)$ *and* $\eta(\cdot)$ *nondecreasing.*

COROLLARY 4. *In the causal diagram of Fig.* 2 *with a binary outcome* Y*,* (1) *holds if we replace Condition* (a) *of Corollary* 3 *by*

$$\text{pr}(A = 1 \mid \Pi, U) = \alpha\Pi\delta^{Y(1)}\eta^{Y(0)}\theta^{Y(1)Y(0)} \text{ with } \delta, \eta, \theta \geqslant 1.$$

5. ILLUSTRATIONS

5·1. *Numerical examples*

Myers et al. (2011) simulated binary (Z, U, A, Y) to investigate Z-bias. They generated (Z, U) according to $\text{pr}(Z = 1) = 0{\cdot}5$ and $\text{pr}(U = 1) = \gamma_0$. The first set of their generative models is additive,

$$\text{pr}(A = 1 \mid U, Z) = \alpha_0 + \alpha_1 U + \alpha_2 Z, \quad \text{pr}(Y = 1 \mid U, A) = \beta_0 + \beta_1 U + \beta_2 A, \tag{3}$$

where the coefficients are all positive. The second set of their generative models is multiplicative,

$$\text{pr}(A = 1 \mid U, Z) = \alpha_0\alpha_1^U\alpha_2^Z, \quad \text{pr}(Y = 1 \mid U, A) = \beta_0\beta_1^U\beta_2^A, \tag{4}$$

where the coefficients in (3) and (4) are all positive. They use simulation to show that Z-bias arises under these models. In fact, in the above models, Z and U have monotonic effects on A without additive or multiplicative interactions, and U acts monotonically on Y, given A. Therefore, Corollaries 1 and 2 imply that Z-bias must occur. The qualitative conclusion follows immediately from our theory. However, our theory does not make statements about the magnitude of the bias, and for more details about the magnitude and finite-sample properties, see Myers et al. (2011).

We use three numerical examples to illustrate the role of the no-interaction assumptions required by Theorems 2 and 3 and Corollaries 1 and 2. Recall the conditional probability of the treatment A, $p_{zu} = \text{pr}(A = 1 \mid Z = z, U = u)$, and define the conditional probabilities of the outcome Y as $r_{au} = \text{pr}(Y = 1 \mid A = a, U = u)$, for $z, a, u = 0, 1$. Table 1 gives three examples where monotonicity on the conditional distributions of A and Y hold, and there are both additive and multiplicative interactions. In all cases, the instrumental variable Z is $\text{Ber}(p = 0{\cdot}5)$, and

Table 1. *Examples of the presence and absence of Z-bias, in which* $Z \sim \mathrm{Ber}(0{\cdot}5)$*,* $U \sim \mathrm{Ber}(0{\cdot}5)$*, the conditional probability of the treatment* A *is* $p_{zu} = \mathrm{pr}(A = 1 \mid Z = z, U = u)$*, and the conditional probability of the outcome* Y *is* $r_{au} = \mathrm{pr}(Y = 1 \mid A = a, U = u)$

Case	p_{11}	p_{10}	p_{01}	p_{00}	r_{11}	r_{10}	r_{01}	r_{00}	ACE$^{\text{true}}$	ACE$^{\text{unadj}}$	ACE$^{\text{adj}}$	Z-bias
1	0·8	0·6	0·2	0·1	0·08	0·06	0·02	0·01	0·0550	0·0574	0·0584	YES
2	0·3	0·2	0·3	0·1	0·03	0·02	0·03	0·01	0·0050	0·0076	0·0077	YES
3	0·5	0·4	0·4	0·1	0·04	0·04	0·04	0·01	0·0150	0·0173	0·0172	NO

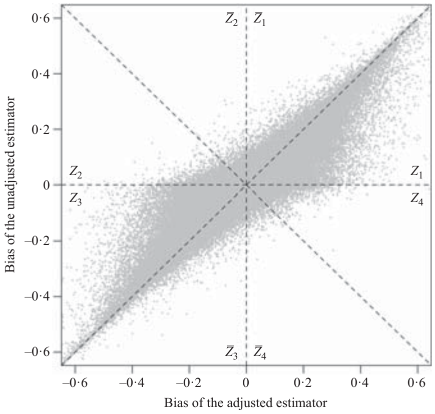

Fig. 3. Biases of the adjusted and unadjusted estimators over 10^6 random draws of the probabilities. In areas (Z_1, Z_2, Z_3, Z_4) Z-bias arises, and in areas $(\bar{Z}_1, \bar{Z}_2, \bar{Z}_3, \bar{Z}_4)$ Z-bias does not arise.

the confounder U is another independent $\mathrm{Ber}(\pi = 0{\cdot}5)$. In Case 1, the weaker condition (2) holds, and our theory implies that Z-bias arises. In Case 2, neither the condition in Theorem 1 nor (2) holds, but Z-bias still arises. Our conditions are only sufficient but not necessary. In Case 3, neither the condition in Theorem 1 nor (2) holds, and Z-bias does not arise.

Finally, for binary (Z, U, A, Y) we use Monte Carlo simulation to compute the volume of the Z-bias space, i.e., the parameter space of p, π, p_{zu}s and r_{au}s in which the adjusted estimator has higher bias than the unadjusted estimator. We randomly draw these ten probabilities from independent $\mathrm{Un}(0, 1)$ random variables, and for each draw of these probabilities we compute the average causal effect ACE^{true}, the unadjusted estimator $\text{ACE}^{\text{unadj}}$ and the adjusted estimator ACE^{adj}. We plot the joint values of the biases $(\text{ACE}^{\text{adj}} - \text{ACE}^{\text{true}}, \text{ACE}^{\text{unadj}} - \text{ACE}^{\text{true}})$ in Fig. 3. The volume of the Z-bias space can be approximated by the frequency that ACE^{adj} deviates more from ACE^{true} than $\text{ACE}^{\text{unadj}}$. With 10^6 random draws, our simulation gives an unbiased estimate for this volume as 0·6805 with estimated standard error 0·0005. Therefore, in about 68% of the parameter space, the adjusted estimator is more biased than the unadjusted estimator.

5·2. *Real data examples*

Bhattacharya & Vogt (2012) presented an example about the treatment effect of small classrooms in the third grade on test scores for reading, where the instrumental variable is the random assignment to small classrooms. Their instrumental variable analysis gave a point estimate of 8·73 with standard error 2·01. Without adjusting for the instrumental variable in the propensity score model, the point estimate was 6·00 with estimated standard error 1·34; adjusting for the instrumental variable, the point estimate was 2·97 with estimated standard error 1·84. The

Table 2. *The example from Wooldridge (2010)*

	Point estimate	Standard error	Lower confidence limit	Upper confidence limit
ACE$^{\text{true}}$	2·47	0·59	1·31	3·62
ACE$^{\text{unadj}}$	1·77	0·07	1·64	1·90
ACE$^{\text{adj}}$	1·76	0·07	1·64	1·89

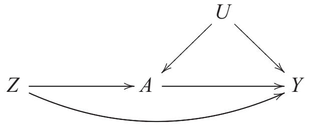

Fig. 4. Directed acyclic graph for Z-bias allowing an arrow from Z to Y

difference between the adjusted estimator and the instrumental variable estimator is larger than that between the unadjusted estimator and the instrumental variable estimator.

Wooldridge (2010, Example 21.3) discusses estimating the effect of attaining at least seven years of education on fertility, with the treatment A being a binary indicator for at least seven years of education, the outcome Y being the number of living children, and the instrumental variable Z being a binary indicator of whether the woman was born in the first half of the year. Although the original dataset of Wooldridge (2010) contains other variables, most of them are posttreatment variables, so we do not adjust for them in our analysis. The instrumental variable analysis gives point estimate 2·47 with estimated standard error 0·59. The unadjusted analysis gives point estimate 1·77 with estimated standard error 0·07. The adjusted analysis gives point estimate 1·76 with estimated standard error 0·07. Table 2 summarizes the results. In this example, the adjusted and unadjusted estimators give similar results.

6. Discussion

6·1. *Allowing an arrow from Z to Y*

When the variable Z has an arrow to the outcome Y as illustrated by Fig. 4, the following generalization of Theorem 1 holds.

Theorem 7. *Consider the causal diagram of Fig. 4 with scalar Z and U, where $Z \perp\!\!\!\perp U$ and $A \perp\!\!\!\perp Y(a) \mid (Z, U)$ for $a = 0$ and 1. The result in* (1) *holds if we replace Condition* (a) *of Theorem 1 by*

> pr$(A = 1 \mid Z = z, U = u)$ *and* $E(Y \mid A = a, Z = z, U = u)$ *are nondecreasing in z and u for $a = 0$ and 1.*

However, when there is an arrow from Z to Y, Theorem 7 is of little use in practice without strong substantive knowledge about the size of the direct effect of Z on Y. In particular, neither Theorem 2 nor Theorem 3 is true when an arrow from Z to Y is added to Fig. 1(b). This reflects the fact that neither the absence of an additive nor the absence of a multiplicative interaction of Z and U on A is sufficient to conclude that $E(Y \mid A = a, Z = z)$ is nonincreasing in z when $E(Y \mid A = a, U = u, Z = z)$ is nondecreasing in z and u.

With a general instrumental variable and a general confounder, Theorem 4 holds without any assumptions on the underlying causal diagram, and therefore it holds even if the variable Z affects the outcome directly. However, Theorems 5 and 6 no longer hold if an arrow from Z to Y is added to Fig. 2. This reflects the fact that the absence of an additive or multiplicative interaction of U

and Π on A no longer implies $\mathrm{cov}\{\Pi, \nu_a(\Pi)\} \leqslant 0$ when Z has a direct effect on Y, even if the remaining conditions of Theorems 5 and 6 hold. Analogously, Theorems 5 and 6 no longer hold if there exists an unmeasured common cause of Z and Y on the causal diagram in Fig. 2, even if Z has no direct effect on Y.

6·2. *Extensions*

In §§ 2–4, we discussed Z-bias for the average causal effects. We can extend the results to distributional causal effects for general outcomes (Ju & Geng, 2010) and causal risk ratios for binary or positive outcomes. Moreover, the results in §§ 2–4 are conditional on or within the strata of observed covariates. Similar results hold for causal effects averaged over observed covariates. We give more details in the Supplementary Material. In this paper we have given sufficient conditions for the presence of Z-bias; future work could consider sufficient conditions for the absence of Z-bias.

6·3. *Conclusion*

It is often suggested that we should adjust for all pretreatment covariates in observational studies. However, we show that in a wide class of models satisfying certain monotonicity conditions, adjusting for an instrumental variable actually amplifies the impact of the unmeasured treatment-outcome confounding, which results in more bias than the unadjusted estimator. In practice, we may not be sure about whether a covariate is a confounder, for which one needs to control, or perhaps instead an instrumental variable, for which control would only increase any existing bias due to unmeasured confounding. Therefore, a more practical approach, as suggested by Rosenbaum (2010, Ch. 18.2), Brookhart et al. (2010) and Pimentel et al. (2016), may be to conduct analysis both with and without adjusting for the covariate. If two analyses give similar results, as in the example in Table 2, then we need not worry about Z-bias; otherwise, we need additional information and analysis before making decisions.

Acknowledgement

Ding was partially supported by the U.S. Institute of Education Sciences, and VanderWeele by the U.S. National Institutes of Health. VanderWeele and Robins are also affiliated with the Department of Biostatistics at the Harvard T. H. Chan School of Public Health. The authors thank the associate editor and two reviewers for detailed and helpful comments.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs and extensions.

References

Ali, M. S., Groenwold, R. H. & Klungel, O. H. (2014). Propensity score methods and unobserved covariate imbalance: Comments on “Squeezing the balloon”. *Health Serv. Res.* **49**, 1074–82.

Bhattacharya, J. & Vogt, W. B. (2012). Do instrumental variables belong in propensity scores? *Int. J. Statist. Econ.* **9**, 107–27.

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J. & Schneeweiss, S. (2010). Confounding control in healthcare database research: Challenges and potential approaches. *Med. Care* **48**, S114–20.

Brooks, J. M. & Ohsfeldt, R. L. (2013). Squeezing the balloon: Propensity scores and unmeasured covariate balance. *Health Serv. Res.* **48**, 1487–507.

Cochran, W. G. (1965). The planning of observational studies of human populations (with Discussion). *J. R. Statist. Soc.* A **128**, 234–66.

Cox, D. & Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *J. R. Statist. Soc.* B **65**, 937–41.

D'Agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statist. Med.* **17**, 2265–81.

Ding, P. & Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-Bias and Butterfly-Bias (with comments). *J. Causal Infer.* **3**, 41–57.

Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *Am. J. Public Health Nations Health* **43**, 677–83.

Esary, J. D., Proschan, F. & Walkup, D. W. (1967). Association of random variables, with applications. *Ann. Math. Statist.* **38**, 1466–74.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology* **14**, 300–6.

Greenland, S. & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* **15**, 413–9.

Heckman, J. & Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Rev. Econ. Statist.* **86**, 30–57.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.* **2**, 259–78.

Ju, C. & Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *J. R. Statist. Soc.* B **72**, 129–42.

Karlin, S. & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Mult. Anal.* **10**, 467–98.

Ma, Z., Xie, X. & Geng, Z. (2006). Collapsibility of distribution dependence. *J. R. Statist. Soc.* B **68**, 127–33.

Middleton, J. A., Scott, M. A., Diakow, R. & Hill, J. L. (2016). Bias amplification and bias unmasking. *Polit. Anal.*, **24**, 307–23.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M. & Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am. J. Epidemiol.* **174**, 1213–22.

Neyman, J. (1923 [1990]). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated by D. M. Dabrowska and T. P. Speed. *Statist. Sci.* **5**, 465–72.

Pearl, J. (1995). Causal diagrams for empirical research (with Discussion). *Biometrika* **82**, 669–88.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Pearl, J. (2009). Letter to the editor. *Statist. Med.* **28**, 1415–6.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In *Proc. 26th Conf. Uncert. Artif. Intel. (UAI 2010)*, P. Grunwald & P. Spirtes, eds. Corvallis, Oregon: Association for Uncetainty in Artificial Intelligence, pp. 425–32.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *Am. J. Epidemiol.* **174**, 1223–7.

Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *J. Causal Infer.* **1**, 155–70.

Pimentel, S. D., Small, D. S. & Rosenbaum, P. R. (2016). Constructed second control groups and attenuation of unmeasured biases. *J. Am. Statist. Assoc.*, **111**, 1157–67.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer, 2nd ed.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statist. Med.* **26**, 20–36.

Rubin, D. B. (2008a). Author's reply. *Statist. Med.* **27**, 2741–2.

Rubin, D. B. (2008b). For objective causal inference, design trumps analysis. *Ann. Appl. Statist.* **2**, 808–40.

Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statist. Med.* **28**, 1420–3.

Shrier, I. (2008). Letter to the editor. *Statist. Med.* **27**, 2740–1.

Shrier, I. (2009). Propensity scores. *Statist. Med.* **28**, 1315–8.

Sjölander, A. (2009). Propensity scores and M-structures. *Statist. Med.* **28**, 1416–20.

Spirtes, P., Glymour, C. N. & Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge: MIT Press, 2nd ed.

VanderWeele, T. J. & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics* **67**, 1406–13.

Walker, A. M. (2013). Matching on provider is risky. *J. Clin. Epidemiol.* **66**, S65–8.

Wooldridge, J. (2016). Should instrumental variables be used as matching variables? *Res. Econ.* **70**, 232–7.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press, 2nd ed.

Xie, X., Ma, Z. & Geng, Z. (2008). Some association measures and their collapsibility. *Statist. Sinica* **18**, 1165–83.

[*Received on* 24 *June* 2016. *Editorial decision on* 22 *January* 2017]