Award:

R305D150040 - PRESIDENT AND FELLOWS OF HARVARD COLLEGE

Paper Title:

Instrumental variables as bias amplifiers with general outcome and confounding

Authors:

Peng Ding, Tyler J. VanderWeele, and James M. Robins

Publication Date:

Jan 16, 2017 (arXiv)

April 17, 2017 (journal)

DOI:

https://doi.org/10.1093/biomet/asx009

Biometrika (2017), **xx**, x, *pp*. 1–25
Printed in Great Britain

Instrumental variables as bias amplifiers with general outcome and confounding

BY P. DING

Department of Statistics, University of California, Berkeley, California, USA.
pengdingpku@berkeley.edu

T. J. VANDERWEELE AND J. M. ROBINS

Departments of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA.
tvanderw@hsph.harvard.edu robins@hsph.harvard.edu

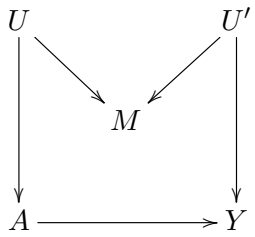
SUMMARY

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 without adjusting 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 in 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. We further show that when the treatment follows an additive or multiplicative model conditional on the instrumental variable and the confounder, these monotonicity assumptions can be interpreted as the signs of the arrows of the causal diagrams.

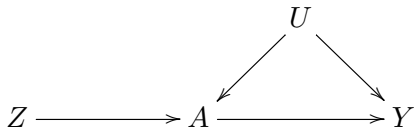
Some key words: Causal inference; Directed acyclic graph; Interaction; Monotonicity; Potential outcome

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 "typically, the more conditional an assumption, the more generally acceptable it is" (Ru-



(a) Directed Acyclic Graph for M-Bias. U and U' are unobserved, and M is observed.



(b) Directed Acyclic Graph for Z-Bias. U is an unmeasured confounder and Z is an instrumental variable for the treatment-outcome relationship.

Fig. 1: Two Directed Acyclic Graphs. A is the treatment, and Y is the outcome of interest.

bin, 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, pp. 76).

Simply adjusting for all pretreatment covariates (d'Agostino, 1998; Rosenbaum, 2002; Hirano & Imbens, 2001), or the pretreatment criterion (VanderWeele & Shpitser, 2011), has a sound justification from the view point 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 data generating 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 Figure 1(a). In this diagram, the ignorability assumption holds unconditionally but not conditionally (Pearl, 2000; Spirtes et al., 2000; Greenland, 2003; Pearl, 2009; Shrier, 2008, 2009; Sjölander, 2009; Ding & Miratrix, 2015). In Figure 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 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 Figure 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. Of particular interest is to determine the conditions under which an instrumental variable is a bias amplifier. An instrumental variables is a pretreamtnet 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 Figure 1(b) is an example. Heckman & Navarro-Lozano (2004) and Bhattacharya & Vogt (2012) 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 technical report in 2006, which was finally published as Wooldridge (2016). Because instrumental variables are often denoted by Z as in Figure 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 in 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. When the instrumental variable and the confounder have either no additive or no multiplicative interaction on the treatment, these assumptions can be interpreted as the signs of the arrows of the causal diagram (VanderWeele & Robins, 2010). 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 Figure 1(b). Let $\perp\!\!\!\perp$ denote conditional independence between random variables. Then the instrumental variable Z in Figure 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 Figure 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 Figure 1(b) with scalar Z and U, if*

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

then

$$\begin{pmatrix} \text{ACE}_1^{\text{adj}} \\ \text{ACE}_0^{\text{adj}} \\ \text{ACE}^{\text{adj}} \end{pmatrix} \geq \begin{pmatrix} \text{ACE}^{\text{unadj}} \\ \text{ACE}^{\text{unadj}} \\ \text{ACE}^{\text{unadj}} \end{pmatrix} \geq \begin{pmatrix} \text{ACE}_1^{\text{true}} \\ \text{ACE}_0^{\text{true}} \\ \text{ACE}^{\text{true}} \end{pmatrix}. \tag{1}$$

Inequalities among vectors as in (1) should be interpreted as component-wise relationships. Intuitively, the monotonicity in Condition (a) of Theorem 1 requires non-negative dependence structures on arrows $Z \to A$, $U \to A$ and $U \to Y$ in the causal diagram of Figure 1(b). Because

the dependence is in expectation, Condition (a) of Theorem 1 is weaker than the requirement of signed directed acyclic graphs (VanderWeele & Robins, 2010).

The monotonicity in Condition (b) of Theorem 1 reflects the collider bias caused by conditioning on A. As noted by Greenland (2003), in many cases, if Z and U affect A in the same direction, then the collider bias caused by conditioning on A is often in the opposite direction. Lemmas S5–S8 in the Supplementary Material show that, if Z and U are independent and have non-negative 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 which require 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 Figure 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 non-decreasing in* z*,* $\gamma(u)$ *is non-decreasing in* u*, and* $E(Y \mid A = a, U = u)$ *is non-decreasing 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 non-negative biases for the true average causal effects for the treatment, control and the whole populations. Furthermore, the adjusted estimators, either for the treatment, control or the whole populations, 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 Figure 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., $p_{11} \geq \max(p_{10}, p_{01})$ *and* $\min(p_{10}, p_{01}) \geq p_{00}$*, and* $E(Y \mid A = a, U = 1) \geq E(Y \mid A = a, U = 0)$ *for both* $a = 1$ *and* 0*.*

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

THEOREM 3. *In the causal diagram of Figure 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 Figure 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:

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

i.e., Z and U have non-positive 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 that 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 $f\{\max(w_1, w_2)\}f\{\min(w_1, w_2)\} \geq f(w_1)f(w_2)$, where $\max(w_1, w_2)$ and $\min(w_1, w_2)$ are component-wise 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

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

the adjusted estimator for the population under control is

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

and the adjusted estimator for the whole population is

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

When Z is scalar, then the above three formulas reduce to the ones in Section 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 need to take $U = \{Y(1), Y(0)\}$ as the ultimate confounder for the relationship of A on Y. This is not an assumption about U. 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 blocks all the back-door paths from A to Y (Pearl, 1995, 2000). We represent the causal structure in Figure 2.

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

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

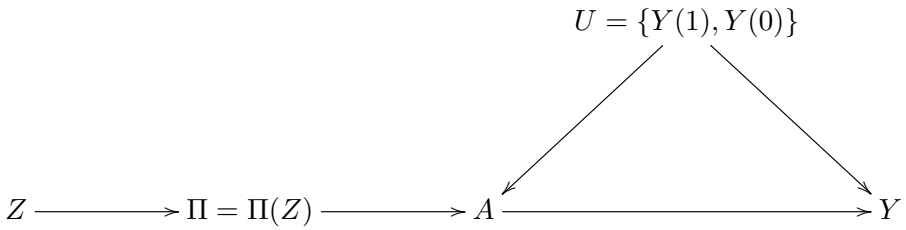


Fig. 2: Directed Acyclic Graph for Z-Bias With General Instrument and Confounder

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 $\text{cov}\{\Pi, \nu_a(\Pi)\} \leq 0$ 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 Figure 2. This measure of collider bias is more general than the one in Theorem 1. Analogous to Section 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., random variable V has a non-negative distributional association on random variable W, if the conditional distribution satisfies $\partial F(w \mid v)/\partial v \leq 0$ 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 Figure 2,* (1) *holds if*

(a) $\text{pr}(A = 1 \mid \Pi, U) = \Pi + \delta\{Y(1)\} + \eta\{Y(0)\}$ *with* $\delta(\cdot)$ *and* $\eta(\cdot)$ *being non-decreasing;*
(b) $\{Y(1), Y(0)\}$ *have non-negative distributional associations on each other, i.e.,* $\partial F(y_1 \mid y_0)/\partial y_0 \leq 0$ *and* $\partial F(y_0 \mid y_1)/\partial y_1 \leq 0$ *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 non-negative distributional association between $Y(1)$ and $Y(0)$ is equivalent to

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

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

COROLLARY 3. *In the causal diagram of Figure 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* $\delta, \eta, \theta \geq 0$*;*
(b) $\text{OR}_Y \geq 1$.

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 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 Figure 2,* (1) *holds if we replace Condition (a) of Theorem* 5 *by*

(a') $\text{pr}(A = 1 \mid \Pi, U) = \Pi\delta\{Y(1)\}\eta\{Y(0)\}$ *with* $\delta(\cdot)$ *and* $\eta(\cdot)$ *being non-decreasing.*

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

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

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 $\text{pr}(Z = 1) = 0.5$ 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 further 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 Bernoulli$(p = 0.5)$, and the confounder U is another independent Bernoulli$(\pi = 0.5)$. 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 or (2) holds, but Z-Bias still arises. Our conditions are only sufficient but not necessary. In Case 3, neither the condition in Theorem 1 or (2) holds, and Z-Bias does not arise.

Finally, for binary (Z, U, A, Y) we use Monte Carlo 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

Table 1: Examples for the presence and absence of Z-Bias, in which $Z \sim \text{Bernoulli}(0.5)$, $U \sim \text{Bernoulli}(0.5)$, the conditional probability of the treatment A is $p_{zu} = \text{pr}(A = 1 \mid Z = z, U = u)$, and the conditional probability of the outcome Y is $r_{au} = \text{pr}(Y = 1 \mid A = a, U = u)$.

Case	p_{11}	p_{10}	p_{01}	p_{00}	r_{11}	r_{10}	r_{01}	r_{00}	ACE^{true}	$\text{ACE}^{\text{unadj}}$	ACE^{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

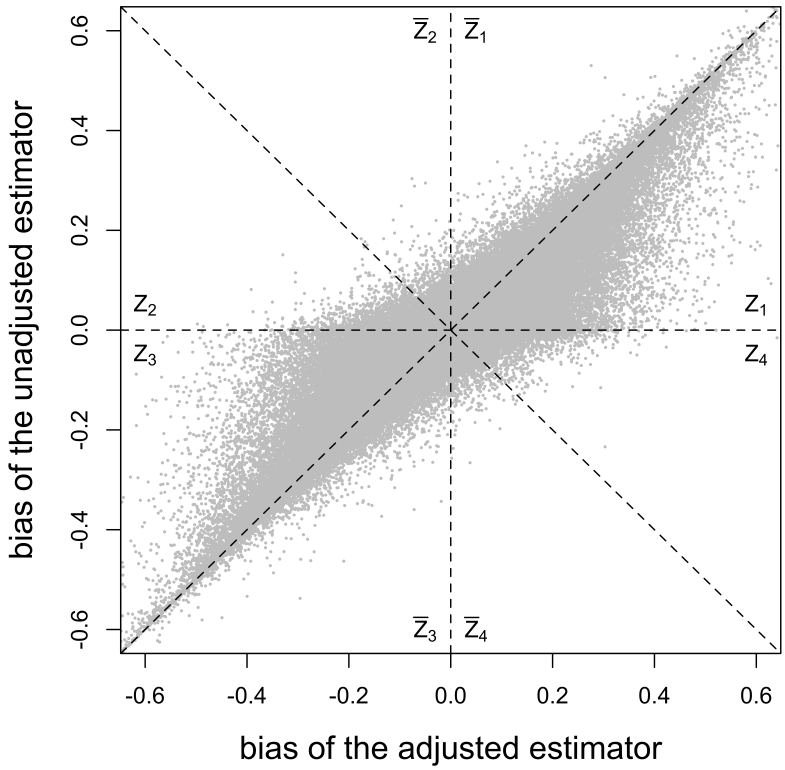


Fig. 3: Biases of the adjusted and unadjusted estimators over 10^6 draws of the probabilities. In areas (Z_1, Z_2, Z_3, Z_4) Z-Bias arises, and in areas $(\bar{Z}_1, \bar{Z}_2, \bar{Z}_3, \bar{Z}_4)$ Z-Bias does not arise.

bias than the unadjusted estimator. We randomly draw these ten probabilities from independent $\text{Uniform}(0, 1)$ 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 Figure 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 Monte Carlo 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 classroom in the third grade on test scores for reading. Their instrumental variable analysis gave point

Table 2: The example from Wooldridge (2010).

	point estimate	standard error	lower confidence limit	upper confidence limit
ACE^{true}	2.47	0.59	1.31	3.62
$\text{ACE}^{\text{unadj}}$	1.77	0.07	1.64	1.90
ACE^{adj}	1.76	0.07	1.64	1.89

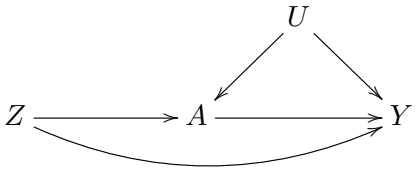


Fig. 4: Directed Acyclic Graph for Z-Bias Allowing for an Arrow from Z to Y

estimate 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 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 treatment A being a binary indicator for at least seven years of education, outcome Y being the number of living children, and instrumental variable Z being a binary indicator if the woman was born in the first half of the year. Although the original data set 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 for an Arrow from Z to Y*

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

THEOREM 7. *Consider the causal diagram of Figure 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*

(a') $\text{pr}(A = 1 \mid Z = z, U = u)$ *and* $E(Y \mid A = a, Z = z, U = u)$ *are non-decreasing 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 present. 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 non-increasing in z when $E(Y \mid A = a, U = u, Z = z)$ is non-decreasing 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 present as in Figure 4. This reflects the fact that the absence of an additive or multiplicative interaction of U and Π on A no longer implies $\text{cov}\{\Pi, \nu_a(\Pi)\} \leq 0$ 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 exits an unmeasured common cause of Z and Y on the causal diagram in Figure 1(b), 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, 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, Chapter 18.2) and Brookhart et al. (2010), 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.

ACKNOWLEDGMENTS

Peng Ding is partially supported by the U.S. Institute of Education Sciences, and Tyler J. VanderWeele by the U.S. National Institutes of Health. The authors thank the Associate Editor and two reviewers for detailed and helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes all 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 Services Research* **49**, 1074–1082.

BHATTACHARYA, J. & VOGT, W. B. (2012). Do instrumental variables belong in propensity scores? *Int. J. Stat. Econ.* **9**, 107–127.

BROOKHART, M. A., STÜRMER, T., GLYNN, R. J., RASSEN, J. & SCHNEEWEISS, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical Care* **48**, S114–S120.
BROOKS, J. M. & OHSFELDT, R. L. (2013). Squeezing the balloon: Propensity scores and unmeasured covariate balance. *Health Services Research* **48**, 1487–1507.
CHIBA, Y. (2009). The sign of the unmeasured confounding bias under various standard populations. *Biometrical Journal* **51**, 670–676.
COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society: Series A (General)* **128**, 234–266.
COX, D. & WERMUTH, N. (2003). A general condition for avoiding effect reversal after marginalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 937–941.
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. *Statistics in Medicine* **17**, 2265–2281.
DING, P. & MIRATRIX, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-Bias and Butterfly-Bias (with comments). *Journal of Causal Inference* **3**, 41–57.
DING, P. & VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27**, 368–377.
DORN, H. F. (1953). Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health* **43**, 677–683.
ESARY, J. D., PROSCHAN, F. & WALKUP, D. W. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics* **38**, 1466–1474.
GREENLAND, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology* **14**, 300–306.
GREENLAND, S. & ROBINS, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* **15**, 413–419.
HECKMAN, J. & NAVARRO-LOZANO, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* **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 Services and Outcomes Research Methodology* **2**, 259–278.
JU, C. & GENG, Z. (2010). Criteria for surrogate end points based on causal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 129–142.
KARLIN, S. & RINOTT, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *Journal of Multivariate Analysis* **10**, 467–498.
MA, Z., XIE, X. & GENG, Z. (2006). Collapsibility of distribution dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 127–133.
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. *American Journal of Epidemiology* **174**, 1213–1222.
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. *Statistical Science* **5**, 465–472.
PEARL, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika* **82**, 669–688.
PEARL, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
PEARL, J. (2009). Letter to the editor. *Statistics in Medicine* **28**, 1415–1416.
PEARL, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, P. Grunwald & P. Spirtes, eds. Association for Uncetainty in Artificial Intelligence, Corvallis, OR: 425–432.
PEARL, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology* **174**, 1223–1227.
PEARL, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference* **1**, 155–170.
PIEGORSCH, W. W., WEINBERG, C. R. & TAYLOR, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13**, 153–162.
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.
ROTHMAN, K. J., GREENLAND, S. & LASH, T. L. (2008). *Modern Epidemiology (3rd Edition)*. Lippincott Williams & Wilkins.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **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. *Statistics in Medicine* **26**, 20–36.
RUBIN, D. B. (2008a). Author's reply. *Statistics in Medicine* **27**, 2741–2742.
RUBIN, D. B. (2008b). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* **2**, 808–840.

RUBIN, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine* **28**, 1420–1423.
SHRIER, I. (2008). Letter to the editior. *Statistics in Medicine* **27**, 2740–2741.
SHRIER, I. (2009). Propensity scores. *Statistics in Medicine* **28**, 1315–1318.
SJÖLANDER, A. (2009). Propensity scores and M-structures. *Statistics in Medicine* **28**, 1416–1420.
SPIRTES, P., GLYMOUR, C. N. & SCHEINES, R. (2000). *Causation, Prediction, and Search*. Cambridge: MIT press, 2nd ed.
VANDERWEELE, T. J. (2008). The sign of the bias of unmeasured confounding. *Biometrics* **64**, 702–706.
VANDERWEELE, T. J. & ROBINS, J. M. (2010). Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 111–127.
VANDERWEELE, T. J. & SHPITSER, I. (2011). A new criterion for confounder selection. *Biometrics* **67**, 1406–1413.
WALKER, A. M. (2013). Matching on provider is risky. *Journal of Clinical Epidemiology* **66**, S65–S68.
WOOLDRIDGE, J. (2016). Should instrumental variables be used as matching variables? *Research in Economics* **70**, 232–237.
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. *Statistica Sinica* **18**, 1165–1183.
YANG, Q., KHOURY, M. J., SUN, F. & FLANDERS, W. D. (1999). Case-only design to measure gene-gene interaction. *Epidemiology* **10**, 167–170.

Supplementary material for "Instrumental variables as bias amplifiers with general outcome and confounding"

APPENDIX 1. LEMMAS AND THEIR PROOFS

In order to prove the main results, we need to invoke the following lemmas. Some of them are from the literature, and some of them are new and of independent interest.

Lemma S1 is from Esary et al. (1967, Theorem 2.1).

LEMMA S1. *Let $f(\cdot)$ and $g(\cdot)$ be functions with K real-valued arguments, which are both non-decreasing in each of their arguments. If $U = (U_1, \ldots, U_K)$ is a multivariate random variable with K mutually independent components, then* $\text{cov}\{f(U), g(U)\} \geq 0$.

Lemma S2 is from VanderWeele (2008), and Lemmas S3 and S4 are from Chiba (2009).

LEMMA S2. *For a univariate U or a multivariate U with mutually independent components, if for $a = 1$ and 0, $Y(a) \perp\!\!\!\perp A \mid U$, $E(Y \mid A = a, U = u)$ is non-decreasing in each component of u, and* $\text{pr}(A = 1 \mid U = u)$ *is non-decreasing in each component of u, then $E(Y \mid A = 1) \geq E\{Y(1)\}$ and $E(Y \mid A = 0) \leq E\{Y(0)\}$.*

LEMMA S3. *For a univariate U and a multivariate U with mutually independent components, if $Y(0) \perp\!\!\!\perp A \mid U$, $E(Y \mid A = 0, U = u)$ is non-decreasing in each component of u, and* $\text{pr}(A = 1 \mid U = u)$ *is non-decreasing in each component of u, then $E(Y \mid A = 0) \leq E\{Y(0) \mid A = 1\}$.*

LEMMA S4. *For a univariate U and a multivariate U with mutually independent components, if $Y(1) \perp\!\!\!\perp A \mid U$, $E(Y \mid A = 1, U = u)$ is non-decreasing in each component of u, and* $\text{pr}(A = 1 \mid U = u)$ *is non-decreasing in each component of u, then $E(Y \mid A = 1) \geq E\{Y(1) \mid A = 0\}$.*

Lemma S5, extending Rothman et al. (2008), states that under monotonicity, no additive interaction implies non-positive multiplicative interactions for both presence and absence of the outcome.

LEMMA S5. *If $p_{11} \geq \max(p_{10}, p_{01})$, $\min(p_{10}, p_{01}) \geq p_{00} > 0$, and $p_{11} - p_{10} - p_{01} + p_{00} = 0$, then*

$$\frac{p_{11}p_{00}}{p_{10}p_{01}} \leq 1, \quad \frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} \leq 1. \tag{S5}$$

Proof of Lemma S5. Define $\text{RR}_{11} = p_{11}/p_{00} \geq 1$, $\text{RR}_{10} = p_{10}/p_{00} \geq 1$ and $\text{RR}_{01} = p_{01}/p_{00} \geq 1$. Then $p_{11} - p_{10} - p_{01} + p_{00} = 0$ implies $\text{RR}_{11} = \text{RR}_{10} + \text{RR}_{01} - 1$, which further implies

$$\begin{aligned}
\frac{p_{11}p_{00}}{p_{10}p_{01}} &= \frac{\text{RR}_{11}}{\text{RR}_{10}\text{RR}_{01}} = 1 + \frac{1}{\text{RR}_{10}\text{RR}_{01}}(\text{RR}_{11} - \text{RR}_{10}\text{RR}_{01}) \\
&= 1 + \frac{1}{\text{RR}_{10}\text{RR}_{01}}(\text{RR}_{10} + \text{RR}_{01} - 1 - \text{RR}_{10}\text{RR}_{01}) \\
&= 1 - \frac{1}{\text{RR}_{10}\text{RR}_{01}}(\text{RR}_{10} - 1)(\text{RR}_{01} - 1) \leq 1.
\end{aligned}$$

The second inequality of (S5) follows from

$$\begin{aligned}
&\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} = 1 + \frac{(1-p_{11})(1-p_{00}) - (1-p_{10})(1-p_{01})}{(1-p_{10})(1-p_{01})} \\
&= 1 + \frac{1}{(1-p_{10})(1-p_{01})}\{(1-p_{11}-p_{00}+p_{11}p_{00}) - (1-p_{10}-p_{01}+p_{10}p_{01})\} \\
&= 1 + \frac{1}{(1-p_{10})(1-p_{01})}(p_{11}p_{00} - p_{10}p_{01}) \\
&= 1 + \frac{p_{10}p_{01}}{(1-p_{10})(1-p_{01})}\left(\frac{p_{11}p_{00}}{p_{10}p_{01}} - 1\right) \leq 1. \qquad \square
\end{aligned}$$

Lemma S5 is about interaction between two binary causes, and for our discussion we need to extend it to interaction between two general causes. Lemma S6 extends Piegorsch et al. (1994) and Yang et al. (1999) by relating the conditional association between two independent causes given the outcome to the interaction between the two causes on the outcome.

LEMMA S6. *If* $Z \perp\!\!\!\perp U$, *and* $\text{pr}(A = 1 \mid Z = z, U = u) = \beta(z) + \gamma(u)$ *with* $\beta(z)$ *and* $\gamma(u)$ *non-decreasing in* z *and* u, *then for both* $a = 1$ *and* 0 *and for all values of* u *and* z,

$$\frac{\partial F(u \mid A = a, Z = z)}{\partial z} \geq 0,$$

i.e., U *has non-positive distributional dependence on* Z, *given* A.

Proof of Lemma S6. For a fixed u and $z_1 > z_0$, we define

$$\begin{aligned}
p_{11} &= \text{pr}(A = 1 \mid U > u, Z = z_1) = \int_u^\infty \{\beta(z_1) + \gamma(u')\}F(\mathrm{d}u')/\{1 - F(u)\}, \\
p_{10} &= \text{pr}(A = 1 \mid U > u, Z = z_0) = \int_u^\infty \{\beta(z_0) + \gamma(u')\}F(\mathrm{d}u')/\{1 - F(u)\}, \\
p_{01} &= \text{pr}(A = 1 \mid U \leq u, Z = z_1) = \int_{-\infty}^u \{\beta(z_1) + \gamma(u')\}F(\mathrm{d}u')/F(u), \\
p_{00} &= \text{pr}(A = 1 \mid U \leq u, Z = z_0) = \int_{-\infty}^u \{\beta(z_0) + \gamma(u')\}F(\mathrm{d}u')/F(u),
\end{aligned}$$

following from the additive model of A and $Z \perp\!\!\!\perp U$.

Because $\beta(z_1) \geq \beta(z_0)$, it is straightforward to show that $p_{11} \geq p_{10}$ and $p_{01} \geq p_{00}$. Because $\gamma(u)$ is increasing in u, we have

$$p_{11} \geq \beta(z_1) + \gamma(u), \quad p_{10} \geq \beta(z_0) + \gamma(u), \quad p_{01} \leq \beta(z_1) + \gamma(u), \quad p_{00} \leq \beta(z_0) + \gamma(u),$$

which imply $p_{11} \geq p_{01}$ and $p_{10} \geq p_{00}$. We further have

$$\begin{aligned}
&p_{11} - p_{10} - p_{01} + p_{00} \\
&= \int_u^\infty \{\beta(z_1) - \beta(z_0)\}F(\mathrm{d}u')/\{1 - F(u)\} - \int_{-\infty}^u \{\beta(z_1) - \beta(z_0)\}F(\mathrm{d}u')/F(u) \\
&= 0.
\end{aligned}$$

The four probabilities $(p_{11}, p_{10}, p_{01}, p_{00})$ satisfy the conditions in Lemma S5, Therefore, (2) holds. Replacing the probabilities in (2) by their definitions above, we have

$$\begin{aligned}&\frac{\text{pr}(A=1 \mid U>u, Z=z_1)\text{pr}(A=1 \mid U\le u, Z=z_0)}{\text{pr}(A=1 \mid U>u, Z=z_0)\text{pr}(A=1 \mid U\le u, Z=z_1)} \le 1\\ \iff &\frac{\text{pr}(A=1 \mid U>u, z_1)}{\text{pr}(A=1 \mid U\le u, z_1)} \le \frac{\text{pr}(A=1 \mid U>u, z_0)}{\text{pr}(A=1 \mid U\le u, z_0)},\end{aligned}$$

and

$$\begin{aligned}&\frac{\text{pr}(A=0 \mid U>u, Z=z_1)\text{pr}(A=0 \mid U\le u, Z=z_0)}{\text{pr}(A=0 \mid U>u, Z=z_0)\text{pr}(A=0 \mid U\le u, Z=z_1)} \le 1\\ \iff &\frac{\text{pr}(A=0 \mid U>u, z_1)}{\text{pr}(A=0 \mid U\le u, z_1)} \le \frac{\text{pr}(A=0 \mid U>u, z_0)}{\text{pr}(A=0 \mid U\le u, z_0)}.\end{aligned}$$

Therefore, for both $a = 1$ and 0 and for all values of u,

$$\frac{\text{pr}(A=a \mid U>u, Z=z)}{\text{pr}(A=a \mid U\le u, Z=z)} \tag{S6}$$

is non-increasing in z. Because of the independence of Z and U, we have

$$\begin{aligned}&F(u \mid A=a, Z=z)\\ =&\frac{\text{pr}(U\le u, A=a \mid Z=z)}{\text{pr}(A=a \mid Z=z)}\\ =&\frac{\text{pr}(U\le u)\text{pr}(A=a \mid U\le u, Z=z)}{\text{pr}(U\le u)\text{pr}(A=a \mid U\le u, Z=z)+\text{pr}(U>u)\text{pr}(A=a \mid U>u, Z=z)}\\ =&\left\{1+\frac{\text{pr}(U>u)}{\text{pr}(U\le u)}\times\frac{\text{pr}(A=a \mid U>u, Z=z)}{\text{pr}(A=a \mid U\le u, Z=z)}\right\}^{-1}.\end{aligned}$$

Therefore, $F(u \mid A = a, Z = z)$ is a non-increasing function of (S6), and the conclusion holds.□

Lemmas S5 and S6 above hold under the assumption of no additive interaction, and the following two lemmas state similar results under the assumption of no multiplicative interaction.

LEMMA S7. *If* $p_{11} \ge \max(p_{10}, p_{01}), \min(p_{10}, p_{01}) \ge p_{00}$, *and* $p_{11}p_{00} = p_{10}p_{01}$, *then*

$$p_{11} - p_{10} - p_{01} + p_{00} \ge 0, \quad \frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} \le 1.$$

Proof of Lemma S7. Using the same notation in the proof of Lemma S5, $p_{11}p_{00} = p_{10}p_{01}$ implies $\text{RR}_{11} = \text{RR}_{10}\text{RR}_{01}$, with $\text{RR}_{10} \ge 1, \text{RR}_{01} \ge 1$, and $\text{RR}_{11} \ge 1$. Therefore,

$$p_{11} - p_{10} - p_{01} + p_{00} = p_{00}(\text{RR}_{10}\text{RR}_{01} - \text{RR}_{10} - \text{RR}_{01} + 1) = p_{00}(\text{RR}_{10} - 1)(\text{RR}_{01} - 1) \ge 0,$$

which further implies that

$$\begin{aligned}\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} &= 1 + \frac{1}{(1-p_{10})(1-p_{01})}\{(1-p_{11})(1-p_{00}) - (1-p_{10})(1-p_{01})\}\\ &= 1 - \frac{p_{11} - p_{10} - p_{01} + p_{00}}{(1-p_{10})(1-p_{01})} \le 1.\end{aligned}$$

□

LEMMA S8. *If* $Z \perp\!\!\!\perp U$, *and* $\text{pr}(A = 1 \mid Z = z, U = u) = \beta(z)\gamma(u)$ *with* $\beta(z) > 0$ *and* $\gamma(u) > 0$ *non-decreasing in* z *and* u, *then* $Z \perp\!\!\!\perp U \mid A = 1$, *and for all values of* u *and* z,

$$\frac{\partial F(u \mid A = 0, Z = z)}{\partial z} \geq 0,$$

i.e., U *has non-positive distributional dependence on* Z, *given* $A = 0$.

*Proof of Lemma S*8. For a fixed u and $z_1 > z_0$, we define

$$\begin{aligned}
p_{11} &= \text{pr}(A = 1 \mid U > u, Z = z_1) = \beta(z_1) \int_u^\infty \gamma(u')F(\mathrm{d}u')/\{1 - F(u)\}, \\
p_{10} &= \text{pr}(A = 1 \mid U > u, Z = z_0) = \beta(z_0) \int_u^\infty \gamma(u')F(\mathrm{d}u')/\{1 - F(u)\}, \\
p_{01} &= \text{pr}(A = 1 \mid U \leq u, Z = z_1) = \beta(z_1) \int_{-\infty}^u \gamma(u')F(\mathrm{d}u')/F(u), \\
p_{00} &= \text{pr}(A = 1 \mid U \leq u, Z = z_0) = \beta(z_0) \int_{-\infty}^u \gamma(u')F(\mathrm{d}u')/F(u),
\end{aligned}$$

following from the multiplicative model of A and $Z \perp\!\!\!\perp U$. Because $\beta(z_1) \geq \beta(z_0)$, we have $p_{11} \geq p_{10}$ and $p_{01} \geq p_{00}$. Because $\gamma(u)$ is increasing in u, we have

$$p_{11} \geq \beta(z_1)\gamma(u), \quad p_{10} \geq \beta(z_0)\gamma(u), \quad p_{01} \leq \beta(z_1)\gamma(u), \quad p_{00} \leq \beta(z_0)\gamma(u),$$

which imply $p_{11} \geq p_{01}$ and $p_{10} \geq p_{00}$. We can further verify $(p_{11}p_{00})/(p_{10}p_{01}) = 1$. Because the four probabilities $(p_{11}, p_{10}, p_{01}, p_{00})$ satisfy the conditions in Lemma S7, we have $\{(1 - p_{11})(1 - p_{00})\}/\{(1 - p_{10})(1 - p_{01})\} \leq 1$. Replacing the probabilities by their definitions, we have

$$\begin{aligned}
\frac{\text{pr}(A = 1 \mid U > u, Z = z_1)\text{pr}(A = 1 \mid U \leq u, Z = z_0)}{\text{pr}(A = 1 \mid U > u, Z = z_0)\text{pr}(A = 1 \mid U \leq u, Z = z_1)} &= 1, \\
\frac{\text{pr}(A = 0 \mid U > u, Z = z_1)\text{pr}(A = 0 \mid U \leq u, Z = z_0)}{\text{pr}(A = 0 \mid U > u, Z = z_0)\text{pr}(A = 0 \mid U \leq u, Z = z_1)} &\leq 1.
\end{aligned}$$

Following the same logic of the proof of Lemma S6, we can prove that $Z \perp\!\!\!\perp U \mid A = 1$, and Z has non-positive distributional association on U, given $A = 0$. □

Define $f = \text{pr}(A = 1)$ to be the proportion of the population under treatment. The average causal effect for the whole population can be written as a convex combination of the average causal effects for the treated and control populations:

$$\text{ACE}^{\text{true}} = E\{Y(1)\} - E\{Y(0)\} = f\text{ACE}_1^{\text{true}} + (1 - f)\text{ACE}_0^{\text{true}}.$$

Analogously, with a scalar instrumental variable, the adjusted estimator for the whole population can be written as

$$\text{ACE}^{\text{adj}} = \int \mu_1(z)F(\mathrm{d}z) - \int \mu_0(z)F(\mathrm{d}z) = f\text{ACE}_1^{\text{adj}} + (1 - f)\text{ACE}_0^{\text{adj}},$$

and with a general instrumental variable,

$$\text{ACE}^{\text{adj}} = \int \nu_1(\pi)F(\mathrm{d}\pi) - \int \nu_0(\pi)F(\mathrm{d}\pi) = f\text{ACE}_1^{\text{adj}} + (1 - f)\text{ACE}_0^{\text{adj}}.$$

LEMMA S9. *With a scalar instrumental variable Z, the differences between the adjusted and unadjusted estimators are*

$$\begin{aligned}
\text{ACE}_1^{\text{adj}} - \text{ACE}^{\text{unadj}} &= -\frac{\text{cov}\{\Pi(Z), \mu_0(Z)\}}{f(1-f)}, \\
\text{ACE}_0^{\text{adj}} - \text{ACE}^{\text{unadj}} &= -\frac{\text{cov}\{\Pi(Z), \mu_1(Z)\}}{f(1-f)}, \\
\text{ACE}^{\text{adj}} - \text{ACE}^{\text{unadj}} &= -\frac{\text{cov}\{\Pi(Z), \mu_0(Z)\}}{1-f} - \frac{\text{cov}\{\Pi(Z), \mu_1(Z)\}}{f}.
\end{aligned}$$

With a general instrumental variable Z, the above formulas hold if we replace $\Pi(Z)$ by Π and $\mu_a(Z) = E(Y \mid A = a, Z)$ by $\nu_a(\Pi) = E(Y \mid A = a, \Pi)$.

Proof of Lemma S9. The difference $\text{ACE}_1^{\text{adj}} - \text{ACE}^{\text{unadj}}$ is equal to

$$\begin{aligned}
&\text{ACE}_1^{\text{adj}} - \text{ACE}^{\text{unadj}} \\
&= E(Y \mid A = 0) - \int \mu_0(z) F(\mathrm{d}z \mid A = 1) \\
&= \int \mu_0(z) F(\mathrm{d}z \mid A = 0) - \int \mu_0(z) F(\mathrm{d}z \mid A = 1) \\
&= \frac{\int \mu_0(z)\{1 - \Pi(z)\} F(\mathrm{d}z)}{1-f} - \frac{\int \mu_0(z)\Pi(z) F(\mathrm{d}z)}{f} \\
&= \frac{1}{f(1-f)} \Big[E\{\mu_0(Z)(1-\Pi(Z))\} E\{\Pi(Z)\} - E\{\mu_0(Z)\Pi(Z)\} E\{1 - \Pi(Z)\} \Big] \\
&= \frac{1}{f(1-f)} \Big[E\{\mu_0(Z)\} E\{\Pi(Z)\} - E\{\mu_0(Z)\Pi(Z)\} \Big] \\
&= -\frac{\text{cov}\{\Pi(Z), \mu_0(Z)\}}{f(1-f)}.
\end{aligned}$$

Similarly, the difference $\text{ACE}_0^{\text{adj}} - \text{ACE}^{\text{unadj}}$ is equal to

$$\begin{aligned}
\text{ACE}_0^{\text{adj}} - \text{ACE}^{\text{unadj}} &= \int \mu_1(z) F(\mathrm{d}z \mid A = 0) - \int \mu_1(z) F(\mathrm{d}z \mid A = 1) \\
&= -\frac{\text{cov}\{\Pi(Z), \mu_1(Z)\}}{f(1-f)}.
\end{aligned}$$

Therefore, the difference $\text{ACE}^{\text{adj}} - \text{ACE}^{\text{unadj}}$ is equal to

$$\begin{aligned}
\text{ACE}^{\text{adj}} - \text{ACE}^{\text{unadj}} &= f(\text{ACE}_1^{\text{adj}} - \text{ACE}^{\text{unadj}}) + (1-f)(\text{ACE}_0^{\text{adj}} - \text{ACE}^{\text{unadj}}) \\
&= -\frac{\text{cov}\{\Pi(Z), \mu_0(Z)\}}{1-f} - \frac{\text{cov}\{\Pi(Z), \mu_1(Z)\}}{f}.
\end{aligned}$$

Analogously, we can prove the results for general instrumental variables. □

APPENDIX 2. PROOFS OF THEOREMS AND COROLLARIES IN THE MAIN TEXT

Proof of Theorem 1. Because $\Pi(z) = \text{pr}(A = 1 \mid Z = z)$ and $\text{pr}(A = 1 \mid U = u)$ are non-decreasing in z and u, and $E(Y \mid A = a, U = u)$ is non-decreasing in u for both $a = 0$ and 1, the unadjusted estimator, $\text{ACE}^{\text{unadj}}$, is larger than or equal to ACE^{true}, $\text{ACE}_1^{\text{true}}$ and $\text{ACE}_0^{\text{true}}$, according to Lemmas S2–S4.

Because $\Pi(Z)$ is non-decreasing and $\mu_a(Z)$ is non-increasing in Z for both $a = 0$ and 1, their covariance is non-positive according to Lemma S1, i.e., $\text{cov}\{\Pi(Z), \mu_a(Z)\} \leq 0$.

Because the differences between all the adjusted estimators, $\text{ACE}_1^{\text{adj}}$, $\text{ACE}_0^{\text{adj}}$ and ACE^{adj}, and the unadjusted estimator, $\text{ACE}^{\text{unadj}}$, are negative constants multiplied by $\text{cov}\{\Pi(Z), \mu_a(Z)\}$, according to Lemma S9 all of $\text{ACE}_1^{\text{adj}}$, $\text{ACE}_0^{\text{adj}}$, and ACE^{adj} are larger or equal to $\text{ACE}^{\text{unadj}}$. □

Proof of Theorem 2. The independence of Z and U implies that

$$\text{pr}(A = 1 \mid Z = z) = \int \text{pr}(A = 1 \mid Z = z, U = u)F(\text{d}u) = \beta(z) + E\{\gamma(U)\},$$
$$\text{pr}(A = 1 \mid U = u) = \int \text{pr}(A = 1 \mid Z = z, U = u)F(\text{d}z) = E\{\beta(Z)\} + \gamma(u)$$

are non-decreasing in z and u. Therefore, according to Theorem 1 we need only to verify that $E(Y \mid A = a, Z = z)$ in non-increasing in z for both $a = 0$ and 1.

Because $Z \perp\!\!\!\perp U$ and $\text{pr}(A = 1 \mid Z = z, U = u) = \beta(z) + \gamma(u)$ with non-decreasing $\beta(z)$ and $\gamma(u)$, we can apply Lemma S6, and conclude that $\partial F(u \mid A = a, Z = z)/\partial z \geq 0$.

Write the essential infimum and supremum of U given $(A = a, Z = z)$ as $\underline{u}(a, z)$ and $\overline{u}(a)$, with the later depending only on a according to Condition (c) of Theorem 2. Because $Y \perp\!\!\!\perp Z \mid (A, U)$, integration or summation by parts gives

$$\begin{aligned}
&E(Y \mid A = a, Z = z) \\
&= \int E(Y \mid A = a, Z = z, U = u)F(\text{d}u \mid A = a, Z = z) \\
&= \int m_a(u)F(\text{d}u \mid A = a, Z = z) \\
&= m_a(u)F(u \mid A = a, Z = z)|_{u=\underline{u}(a,z)}^{u=\overline{u}(a)} - \int \left\{\frac{\partial m_a(u)}{\partial u}\right\} F(u \mid A = a, Z = z)\,\text{d}u \\
&= m_a\{\overline{u}(a)\} - \int \left\{\frac{\partial m_a(u)}{\partial u}\right\} F(u \mid A = a, Z = z)\,\text{d}u.
\end{aligned}$$

Therefore, its derivative with respect to z,

$$\begin{aligned}
\frac{\partial E(Y \mid A = a, z)}{\partial z} &= -\frac{\partial}{\partial z}\int \left\{\frac{\partial m_a(u)}{\partial u}\right\} F(u \mid A = a, Z = z)\,\text{d}u \\
&= -\int \left\{\frac{\partial m_a(u)}{\partial u}\right\}\left\{\frac{\partial F(u \mid A = a, Z = z)}{\partial z}\right\}\text{d}u,
\end{aligned}$$

is smaller than or equal to zero, because $\partial m_a(u)/\partial u \geq 0$ for both $a = 0$ and 1 and for all u. □

Proof of Corollary 1. According to Theorem 1 we need only to verify that $\mu_a(z) = E(Y \mid A = a, Z = z)$ is non-increasing in z for both $a = 0$ and 1. Following Lemma S5, for binary and independent Z and U, monotonicity and no additive interaction imply (S5), which, according to Bayes' Theorem, is equivalent to

$$\frac{\text{pr}(A = 1 \mid Z = 1, U = 1)\text{pr}(A = 1 \mid Z = 0, U = 0)}{\text{pr}(A = 1 \mid Z = 1, U = 0)\text{pr}(A = 1 \mid Z = 0, U = 1)} = \text{OR}_{ZU|A=1} \leq 1, \tag{S7}$$
$$\frac{\text{pr}(A = 0 \mid Z = 1, U = 1)\text{pr}(A = 0 \mid Z = 0, U = 0)}{\text{pr}(A = 0 \mid Z = 1, U = 0)\text{pr}(A = 0 \mid Z = 0, U = 1)} = \text{OR}_{ZU|A=0} \leq 1. \tag{S8}$$

The above inequalities (S7) and (S8) state that Z and U have negative association given each level of A, and therefore $\text{pr}(U = 1 \mid A = a, Z = z)$ is non-increasing in z for both $a = 1$ and 0.

Because $m_a(1) \geq m_a(0)$ and

$$\begin{aligned}\mu_a(z) &= E(Y \mid A = a, Z = z) \\ &= \sum_{u=0,1} E(Y \mid A = a, Z = z, U = u)\text{pr}(U = u \mid A = a, Z = z) \\ &= m_a(1)\text{pr}(U = 1 \mid A = a, Z = z) + m_a(0)\{1 - \text{pr}(U = 1 \mid A = a, Z = z)\} \\ &= \{m_a(1) - m_a(0)\}\text{pr}(U = 1 \mid A = a, Z = z) + m_a(0),\end{aligned}$$

we know that $\mu_a(z)$ is non-decreasing in $\text{pr}(U = 1 \mid A = a, Z = z)$. Therefore, $\mu_a(z)$ is non-increasing in z for both $a = 1$ and 0. □

Proof of Theorem 3. Because of the independence of Z and U, we have $\text{pr}(A = 1 \mid Z = z) = \beta(z)E\{\gamma(U)\}$ and $\text{pr}(A = 1 \mid U = u) = E\{\beta(Z)\}\gamma(u)$ are non-decreasing in z and u. According to Lemma S8, the multiplicative model of A also implies that for both $a = 1$ and 0 and for all z and u, $\partial F(u \mid A = a, Z = z)/\partial z \geq 0$. Following exactly the same steps of the proof of Theorem 2, we can prove Theorem 3. □

Proof of Corollary 2. For binary and independent Z and U, monotonicity, no multiplicative interaction, and Lemma S7 imply

$$\frac{p_{11}p_{00}}{p_{10}p_{01}} = 1 \leq 1, \quad \frac{(1 - p_{11})(1 - p_{00})}{(1 - p_{10})(1 - p_{01})} \leq 1. \tag{S9}$$

With the above results in (S9), the rest of the proof is the same as the proof of Corollary 1. □

Proof of Theorem 4. First, we consider the treatment effect on the population under treatment. Taking $U = Y(0)$ in Lemma S3, we have $\text{ACE}^{\text{unadj}} \geq \text{ACE}_1^{\text{true}}$, because $A \perp\!\!\!\perp Y(0) \mid Y(0)$, $\text{pr}\{A = 1 \mid Y(0)\}$ is non-decreasing in $Y(0)$, and $E\{Y \mid A = 0, Y(0)\} = Y(0)$ is non-decreasing in $Y(0)$. The condition $\text{cov}\{\Pi, E(Y \mid A = 0, \Pi)\} \leq 0$ implies that $\text{ACE}_1^{\text{adj}} \geq \text{ACE}^{\text{unadj}}$ according to Lemma S9. Therefore, $\text{ACE}_1^{\text{adj}} \geq \text{ACE}^{\text{unadj}} \geq \text{ACE}_1^{\text{true}}$.

Second, we take $U = Y(1)$ in Lemma S4, and by a similar argument as above we have $\text{ACE}_0^{\text{adj}} \geq \text{ACE}^{\text{unadj}} \geq \text{ACE}_0^{\text{true}}$.

The conclusion holds because $\text{ACE}^{\text{true}} = f\text{ACE}_1^{\text{true}} + (1 - f)\text{ACE}_0^{\text{true}}$ and $\text{ACE}^{\text{adj}} = f\text{ACE}_1^{\text{adj}} + (1 - f)\text{ACE}_0^{\text{adj}}$. □

Proof of Theorem 5. Under the additive model of A given Π and $U = \{Y(1), Y(0)\}$, we have the following results. First, $\text{pr}(A = 1 \mid \Pi) = \Pi$ is increasing in Π. Second, $\Pi \perp\!\!\!\perp \{Y(1), Y(0)\}$ implies

$$\begin{aligned}\text{pr}\{A = 1 \mid \Pi, Y(1) = y_1\} &= \int \text{pr}(A = 1 \mid \Pi, U)F(\text{d}y_0 \mid y_1) \\ &= \int \{\Pi + \delta(y_1) + \eta(y_0)\}F(\text{d}y_0 \mid y_1) \\ &= \Pi + \delta(y_1) + \int \eta(y_0)F(\text{d}y_0 \mid y_1) \equiv \Pi + \widetilde{\delta}(y_1).\end{aligned}$$

Denote the infimum and supremum of $Y(0)$ given $Y(1) = y_1$ by $\underline{y}_0(y_1)$ and $\overline{y}_0$, with the later not depending on y_1 according to Condition (c) of Theorem 5. Applying integration or summation

by parts, we have

$$\begin{aligned}\widetilde{\delta}(y_1) &= \delta(y_1) + \eta(y_0)F(y_0 \mid y_1)|_{y_0=\underline{y}_0(y_1)}^{y_0=\overline{y}_0} - \int \left\{\frac{\mathrm{d}\eta(y_0)}{\mathrm{d}y_0}\right\} F(y_0 \mid y_1)\,\mathrm{d}y_0 \\ &= \delta(y_1) + \eta(\overline{y}_0) - \int \left\{\frac{\mathrm{d}\eta(y_0)}{\mathrm{d}y_0}\right\} F(y_0 \mid y_1)\,\mathrm{d}y_0.\end{aligned}$$

The function $\widetilde{\delta}(y_1)$ is non-decreasing in y_1, because

$$\frac{\mathrm{d}\,\widetilde{\delta}(y_1)}{\mathrm{d}y_1} = \frac{\mathrm{d}\delta(y_1)}{\mathrm{d}y_1} - \int \left\{\frac{\mathrm{d}\eta(y_0)}{\mathrm{d}y_0}\right\} \left\{\frac{\partial F(y_0 \mid y_1)}{\partial y_1}\right\} \mathrm{d}y_0 \geq 0.$$

Third, following the same reasoning as the second argument, we have $\mathrm{pr}\{A = 1 \mid \Pi, Y(1) = y_0\} = \Pi + \widetilde{\eta}(y_0)$, with $\widetilde{\eta}(y_0)$ being a non-decreasing function of y_0. Fourth, $\Pi \perp\!\!\!\perp Y(1)$ implies $\mathrm{pr}\{A = 1 \mid Y(1) = y_1\} = f + \widetilde{\delta}(y_1)$, which is non-decreasing in y_1. Fifth, $\Pi \perp\!\!\!\perp Y(0)$ implies $\mathrm{pr}\{A = 1 \mid Y(0) = y_0\} = f + \widetilde{\eta}(y_0)$, which is non-decreasing in y_0.

According the fourth and fifth arguments above, Condition (a) in Theorem 4 holds. Therefore, we need only to verify Condition (b) in Theorem 4 to complete the proof.

We have shown that $\mathrm{pr}\{A = 1 \mid \Pi, Y(1)\} = \Pi + \widetilde{\delta}\{Y(1)\}$, which is additive and non-decreasing in Π and $Y(1)$. According to Lemma S6, we know that

$$\frac{\partial \mathrm{pr}\{Y(1) \leq y_1 \mid A = 1, \Pi = \pi\}}{\partial \pi} \geq 0 \tag{S10}$$

for all y_1 and π. We have also shown that $\mathrm{pr}\{A = 1 \mid \Pi, Y(0)\} = \Pi + \widetilde{\eta}\{Y(0)\}$, which is additive and non-decreasing in Π and $Y(0)$. Again according to Lemma S6, we know that

$$\frac{\partial \mathrm{pr}\{Y(0) \leq y_0 \mid A = 0, \Pi = \pi\}}{\partial \pi} \geq 0 \tag{S11}$$

for all y_0 and π. According to Xie et al. (2008), the above negative distributional associations in (S10) and (S11) imply the negative associations in expectation between $Y(0)$ and Π given A, as required by condition (b) of Theorem 4. □

Proof of Corollary 3. As shown in the proof of Theorem 5, the conclusion follows immediately from the five ingredients. We will show that they hold even if there is non-negative interaction between binary $Y(1)$ and $Y(0)$. The following proof is in parallel with the proof of Theorem 5.

First, $\mathrm{pr}(A = 1 \mid \Pi) = \Pi$ is increasing in Π. Second,

$$\begin{aligned}&\mathrm{pr}\{A = 1 \mid \Pi, Y(1) = y_1\} \\ &= E\left[\mathrm{pr}\{A = 1 \mid \Pi, Y(1) = y_1, Y(0)\} \mid \Pi, Y(1) = y_1\right] \\ &= E\left\{\alpha + \Pi + \delta y_1 + \eta Y(0) + \theta y_1 Y(0) \mid \Pi, Y(1) = y_1\right\} \\ &= \alpha + \Pi + \delta y_1 + \eta \mathrm{pr}\{Y(0) = 1 \mid Y(1) = y_1\} + \theta y_1 \mathrm{pr}\{Y(0) = 1 \mid Y(1) = y_1\} \quad \text{(S12)} \\ &\equiv \Pi + \widetilde{\delta}[y_1 - E\{Y(1)\}]. \quad \text{(S13)}\end{aligned}$$

The last equation in (S13) follows from the fact that $Y(1)$ is binary and the functional form must be linear in y_1, where the coefficient is

$$\begin{aligned}\widetilde{\delta} &= \text{pr}\{A=1 \mid \Pi, Y(1)=1\} - \text{pr}\{A=1 \mid \Pi, Y(1)=0\} \\ &= \delta + \eta[\text{pr}\{Y(0)=1 \mid Y(1)=1\} - \text{pr}\{Y(0)=1 \mid Y(1)=0\}] + \theta \text{pr}\{Y(0)=1 \mid Y(1)=1\} \quad \text{(S14)} \\ &\geq \eta[\text{pr}\{Y(0)=1 \mid Y(1)=1\} - \text{pr}\{Y(0)=1 \mid Y(1)=0\}], \quad \text{(S15)}\end{aligned}$$

where (S14) follows from (S12), and (S15) follows from $\delta \geq 0$ and $\theta \geq 0$. Because $\text{OR}_Y \geq 1$, the potential outcomes have non-negative association, implying that their risk difference $\text{RD}_Y = \text{pr}\{Y(0)=1 \mid Y(1)=1\} - \text{pr}\{Y(0)=1 \mid Y(1)=0\} \geq 0$. Therefore, $\widetilde{\delta} \geq 0$, and $\text{pr}\{A=1 \mid \Pi, Y(1)\}$ is additive and non-decreasing in Π and $Y(1)$.

Third, similar to the second argument, we have $\text{pr}\{A=1 \mid \Pi, Y(0)=y_0\} = \Pi + \widetilde{\eta}[y_0 - E\{Y(0)\}]$ with $\widetilde{\eta} \geq 0$. Therefore, $\text{pr}\{A=1 \mid \Pi, Y(0)\}$ is additive and non-decreasing in Π and $Y(0)$. Fourth, $\Pi \perp\!\!\!\perp Y(1)$ implies that $\text{pr}\{A=1 \mid Y(1)\} = f + \widetilde{\delta} Y(1)$ is increasing in $Y(1)$. Fifth, $\Pi \perp\!\!\!\perp Y(0)$ implies that $\text{pr}\{A=1 \mid Y(0)\} = f + \widetilde{\eta} Y(0)$ is increasing in $Y(0)$.

With these five ingredients, the rest of the proof is exactly the same as the proof of Theorem 5. □

Proof of Theorem 6. First, $\text{pr}(A=1 \mid \Pi) = \Pi$ is non-decreasing in Π. Second,

$$\text{pr}\{A=1 \mid \Pi, Y(1)=y_1\} = \Pi\delta(y_1) \int \delta(y_0) F(\mathrm{d}y_0 \mid y_1) \equiv \Pi\widetilde{\delta}(y_1)$$

is multiplicative and non-decreasing in Π and y_1, following the same argument as the proof of Theorem 5. Third, $\text{pr}\{A=1 \mid \Pi, Y(0)=y_0\} = \Pi\widetilde{\eta}(y_0)$ is multiplicative and non-decreasing in Π and y_0. Fourth, $\text{pr}\{A=1 \mid Y(1)=y_1\} = f\widetilde{\delta}(y_1)$ is non-decreasing in y_1. Fifth, $\text{pr}\{A=1 \mid Y(0)=y_0\} = f\widetilde{\eta}(y_0)$ is non-decreasing in y_0.

The multiplicative models and Lemma S8 imply that for all π, y_1 and y_0,

$$\frac{\partial \text{pr}\{Y(1) \leq y_1 \mid A=1, \Pi=\pi\}}{\partial \pi} = 0 \leq 0, \quad \frac{\partial \text{pr}\{Y(0) \leq y_0 \mid A=0, \Pi=\pi\}}{\partial \pi} \geq 0. \quad \text{(S16)}$$

The rest part is the same as the proof of Theorem 5. □

Proof of Corollary 4. First, $\text{pr}(A=1 \mid \Pi) = \Pi$ is non-decreasing in Π. Second,

$$\text{pr}\{A=1 \mid \Pi, Y(1)=y_1\} = \alpha\Pi\delta^{y_1} E\{\eta^{Y(0)}\theta^{y_1 Y(0)} \mid Y(1)=y_1\} \equiv \alpha\Pi\widetilde{\delta}^{Y(1)},$$

where the functional form must be multiplicative because of binary $Y(0)$, and the parameter $\widetilde{\delta}$ is

$$\begin{aligned}\widetilde{\delta} &= \frac{\text{pr}\{A=1 \mid \Pi, Y(1)=1\}}{\text{pr}\{A=1 \mid \Pi, Y(1)=0\}} \\ &= \delta \times \frac{E\{\eta^{Y(0)}\theta^{Y(0)} \mid Y(1)=1\}}{E\{\eta^{Y(0)} \mid Y(1)=0\}} \\ &= \delta \times \frac{\eta\theta \text{pr}\{Y(0)=1 \mid Y(1)=1\} + \text{pr}\{Y(0)=0 \mid Y(1)=1\}}{\eta \text{pr}\{Y(0)=1 \mid Y(1)=0\} + \text{pr}\{Y(0)=0 \mid Y(1)=0\}} \\ &= \delta \times \frac{(\eta\theta - 1)\text{pr}\{Y(0)=1 \mid Y(1)=1\} + 1}{(\eta - 1)\text{pr}\{Y(0)=1 \mid Y(1)=0\} + 1}.\end{aligned}$$

Because $\text{OR}_Y \geq 1$, we have $\text{pr}\{Y(0)=1 \mid Y(1)=1\} \geq \text{pr}\{Y(0)=1 \mid Y(1)=0\}$, which implies that $\widetilde{\delta} \geq 1$. Therefore, $\text{pr}\{A=1 \mid \Pi, Y(1)\}$ is multiplicative and non-decreasing in Π

and $Y(1)$. Third, we can similarly show that $\text{pr}\{A=1 \mid \Pi, Y(0)\}$ is multiplicative and non-decreasing in Π and $Y(0)$. Fourth, $\text{pr}\{A=1 \mid Y(1)=y_1\} = \alpha f \widetilde{\delta}^{y_1}$ is non-decreasing in y_1. Fifth, $\text{pr}\{A=1 \mid Y(0)=y_0\} = \alpha f \widetilde{\eta}^{y_0}$ is non-decreasing in y_0.

The rest part is the same as the proof of Theorem 6. □

Proof of Theorem 7.. In Figure 4, Z and U are two independent confounders for the relationship between A and Y. Because $\text{pr}(A=1 \mid Z=z, U=u)$ and $E(Y \mid A=a, Z=z, U=u)$ are non-decreasing in z and u for both $a=0$ and 1, Lemmas S2–S4 imply that the unadjusted estimator, $\text{ACE}^{\text{unadj}}$, is larger than or equal to ACE^{true}, $\text{ACE}_1^{\text{true}}$ and $\text{ACE}_0^{\text{true}}$.

The independence between Z and U implies $\text{pr}(A=1 \mid Z=z) = \int \text{pr}(A=1 \mid Z=z, U=u)F(\text{d}u)$, and the monotonicity of $\text{pr}(A=1 \mid Z=z, U=u)$ in z implies that $\text{pr}(A=1 \mid Z=z)$ is non-decreasing in z. The rest of the proof is identical to the proof of Theorem 1. □

APPENDIX 3. EXTENSIONS TO OTHER CAUSAL MEASURES

Appendix 3·1. Distributional Causal Effects

Sometimes we are also interested in estimating the distributional causal effects (Ju & Geng, 2010) for the treatment, control and whole populations:

$$\begin{aligned}
\text{DCE}_1^{\text{true}}(y) &= \text{pr}\{Y(1)>y \mid A=1\} - \text{pr}\{Y(0)>y \mid A=1\},\\
\text{DCE}_0^{\text{true}}(y) &= \text{pr}\{Y(1)>y \mid A=0\} - \text{pr}\{Y(0)>y \mid A=0\},\\
\text{DCE}^{\text{true}}(y) &= \text{pr}\{Y(1)>y\} - \text{pr}\{Y(0)>y\}.
\end{aligned}$$

The unadjusted estimator is

$$\text{DCE}^{\text{unadj}}(y) = \text{pr}(Y>y \mid A=1) - \text{pr}(Y>y \mid A=0).$$

The adjusted estimators for the treatment, control and whole populations are

$$\begin{aligned}
\text{DCE}_1^{\text{adj}}(y) &= \text{pr}(Y>y \mid A=1) - \int \text{pr}(Y>y \mid A=0, z)F(\text{d}z \mid A=1),\\
\text{DCE}_0^{\text{adj}}(y) &= \int \text{pr}(Y>y \mid A=1, z)F(\text{d}z \mid A=0) - \text{pr}(Y>y \mid A=0),\\
\text{DCE}^{\text{adj}}(y) &= \int \text{pr}(Y>y \mid A=1, z)F(\text{d}z) - \int \text{pr}(Y>y \mid A=0, z)F(\text{d}z).
\end{aligned}$$

If the outcome is binary, then the distributional causal effects at $y<1$ are the average causal effects, and zero at $y \geq 1$. All results about distributional causal effects reduce to average causal effects for binary outcome. For a general outcome, the distributional causal effects are the average causal effects on the dichotomized outcome $I_y = I(Y>y)$. Therefore, if we replace the outcome Y by I_y in Theorems 1–3, the results about Z-Bias hold for distributional effects. For instance, the condition that $\text{pr}(Y>y \mid A=a, U=u)$ is non-decreasing in u for all a is the same as requiring a non-negative sign on the arrow $U \to Y$, according to the theory of signed directed acyclic graphs (VanderWeele & Robins, 2010). The following theorem states the results analogous to Theorems 4–6.

COROLLARY S5. *In the causal diagram of Figure 2, if for all y and for both $a=1$ and 0,*

(a) $\text{pr}\{Y(a)>y \mid A=1\} \geq \text{pr}\{Y(a)>y \mid A=0\}$*;*
(b) $\text{cov}\{\Pi, \text{pr}(Y>y \mid A=a, \Pi)\} \leq 0$*;*

then

$$\begin{pmatrix} \mathrm{DCE}_1^{\mathrm{adj}}(y) \\ \mathrm{DCE}_0^{\mathrm{adj}}(y) \\ \mathrm{DCE}^{\mathrm{adj}}(y) \end{pmatrix} \geq \begin{pmatrix} \mathrm{DCE}^{\mathrm{unadj}}(y) \\ \mathrm{DCE}^{\mathrm{unadj}}(y) \\ \mathrm{DCE}^{\mathrm{unadj}}(y) \end{pmatrix} \geq \begin{pmatrix} \mathrm{DCE}_1^{\mathrm{true}}(y) \\ \mathrm{DCE}_0^{\mathrm{true}}(y) \\ \mathrm{DCE}^{\mathrm{true}}(y) \end{pmatrix}. \tag{S17}$$

Under the conditions of Theorems 5 *and* 6, (S17) *holds.*

*Proof of Corollary S*5. Condition (a) of Corollary S5 is equivalent to $\mathrm{pr}\{A = 1 \mid I_y(a) = 1\} \geq \mathrm{pr}\{A = 1 \mid I_y(a) = 0\}$, and Condition (b) of Corollary S5 is equivalent to $\mathrm{cov}\{\Pi, E(I_y \mid A = a, \Pi)\} \leq 0$. Therefore, the conclusion follows from Theorem 4.

According to the proofs of Theorems 5 and 6, we have

$$\begin{aligned} &\mathrm{pr}\{A = 1 \mid I_y(a) = 1\} = \mathrm{pr}\{A = 1 \mid Y(a) > y\} \geq \mathrm{pr}\{A = 1 \mid Y(a) = y\} \\ \geq\ &\mathrm{pr}\{A = 1 \mid Y(a) \leq y\} = \mathrm{pr}\{A = 1 \mid I_y(a) = 0\}, \end{aligned}$$

because of monotonicity of $\mathrm{pr}\{A = 1 \mid Y(a)\}$ in $Y(a)$. Therefore, Condition (a) of Theorem S5 holds. Under the conditions of Theorems 5 and 6, we have also shown in (S10)–(S16) that for all a, y and π, $\partial \mathrm{pr}(Y \leq y \mid A = a, \Pi = \pi)/\partial\pi \geq 0$, which implies that $E(I_y \mid A = a, \Pi = \pi)$ is non-increasing in π. Therefore, Condition (b) of Theorem S5 holds. The proof is complete. □

Appendix 3·2. Ratio Measures

In many applications with binary or positive outcomes, we are also interested in assessing causal effects on the ratio scale for the treatment, control and whole populations, defined as

$$\mathrm{RR}_1^{\mathrm{true}} = \frac{E\{Y(1) \mid A = 1\}}{E\{Y(0) \mid A = 1\}}, \quad \mathrm{RR}_0^{\mathrm{true}} = \frac{E\{Y(1) \mid A = 0\}}{E\{Y(0) \mid A = 0\}}, \quad \mathrm{RR}^{\mathrm{true}} = \frac{E\{Y(1)\}}{E\{Y(0)\}}.$$

The unadjusted estimator on the ratio scale is

$$\mathrm{RR}^{\mathrm{unadj}} = \frac{E(Y \mid A = 1)}{E(Y \mid A = 0)}.$$

The adjusted estimators on the ratio scale for the treatment, control and whole populations are

$$\begin{aligned} \mathrm{RR}_1^{\mathrm{adj}} &= \frac{E(Y \mid A = 1)}{\int E\{Y \mid A = 0, Z = z\} F(\mathrm{d}z \mid A = 1)}, \\ \mathrm{RR}_0^{\mathrm{adj}} &= \frac{\int E\{Y \mid A = 1, Z = z\} F(\mathrm{d}z \mid A = 0)}{E(Y \mid A = 0)}, \\ \mathrm{RR}^{\mathrm{adj}} &= \frac{\int E\{Y \mid A = 1, Z = z\} F(\mathrm{d}z)}{\int E\{Y \mid A = 0, Z = z\} F(\mathrm{d}z)}. \end{aligned}$$

With a general instrumental variable Z, we can replace Z by Π in the definitions of the adjusted estimators.

COROLLARY S6. *All the theorems and corollaries in* §§3 *and* 4 *hold on the ratio scale, i.e., under their conditions,*

$$\begin{pmatrix} \mathrm{RR}_1^{\mathrm{adj}} \\ \mathrm{RR}_0^{\mathrm{adj}} \\ \mathrm{RR}^{\mathrm{adj}} \end{pmatrix} \geq \begin{pmatrix} \mathrm{RR}^{\mathrm{unadj}} \\ \mathrm{RR}^{\mathrm{unadj}} \\ \mathrm{RR}^{\mathrm{unadj}} \end{pmatrix} \geq \begin{pmatrix} \mathrm{RR}_1^{\mathrm{true}} \\ \mathrm{RR}_0^{\mathrm{true}} \\ \mathrm{RR}^{\mathrm{true}} \end{pmatrix}.$$

Proof of Corollary S6. First, $\mathrm{RR}^{\mathrm{true}}$ is a convex combination of $\mathrm{RR}_1^{\mathrm{true}}$ and $\mathrm{RR}_0^{\mathrm{true}}$, and $\mathrm{RR}^{\mathrm{adj}}$ is a convex combination of $\mathrm{RR}_1^{\mathrm{adj}}$ and $\mathrm{RR}_0^{\mathrm{adj}}$, which are formally stated in Ding & VanderWeele (2016, eAppendix). Then the conclusion follows from the proofs of the theorems above. □

Appendix 3·3. Average Over Observed Covariates

In practice, we need to adjust for the observed covariates X that are confounders affecting both the treatment and outcome. The discussion in previous sections is conditional on or within strata of observed covariates X, and the causal effects and their estimators are given X. For example,

$$\begin{aligned}
\mathrm{ACE}^{\mathrm{true}}(x) &= E\{Y(1) \mid X = x\} - E\{Y(0) \mid X = x\},\\
\mathrm{ACE}^{\mathrm{unadj}}(x) &= E(Y \mid A = 1, X = x) - E(Y \mid A = 0, X = x),\\
\mathrm{ACE}^{\mathrm{adj}}(x) &= \int E(Y \mid A = 1, Z = z, X = x) F(\mathrm{d}z \mid X = x)\\
&\quad - \int E(Y \mid A = 0, Z = z, X = x) F(\mathrm{d}z \mid X = x),
\end{aligned}$$

and other conditional quantities can be analogously defined. If the conditions in the theorems and corollaries in §§3 and 4 hold within each level of X, then the conclusions in (1) and (S17) hold not only within each level of X but also averaged over X. For example, for the average causal effects, we have

$$\begin{aligned}
\begin{pmatrix} \int \mathrm{ACE}_1^{\mathrm{adj}}(x) F(\mathrm{d}x \mid A = 1) \\ \int \mathrm{ACE}_0^{\mathrm{adj}}(x) F(\mathrm{d}x \mid A = 0) \\ \int \mathrm{ACE}^{\mathrm{adj}}(x) F(\mathrm{d}x) \end{pmatrix} &\geq \begin{pmatrix} \int \mathrm{ACE}^{\mathrm{unadj}}(x) F(\mathrm{d}x \mid A = 1) \\ \int \mathrm{ACE}^{\mathrm{unadj}}(x) F(\mathrm{d}x \mid A = 0) \\ \int \mathrm{ACE}^{\mathrm{unadj}}(x) F(\mathrm{d}x) \end{pmatrix}\\
&\geq \begin{pmatrix} \int \mathrm{ACE}_1^{\mathrm{true}}(x) F(\mathrm{d}x \mid A = 1) \\ \int \mathrm{ACE}_0^{\mathrm{true}}(x) F(\mathrm{d}x \mid A = 0) \\ \int \mathrm{ACE}^{\mathrm{true}}(x) F(\mathrm{d}x) \end{pmatrix}.
\end{aligned}$$