The sign of the bias of unmeasured confounding

BY TYLER J. VANDERWEELE

Department of Health Studies, University of Chicago

5841 South Maryland Avenue, MC 2007, Chicago, IL 60637

vanderweele@uchicago.edu
Summary. Unmeasured confounding variables are a common problem in drawing causal inferences in observational studies. A theorem is given which in certain circumstances allows the researcher to draw conclusions about the sign of the bias of unmeasured confounding. Specifically, it is possible to determine the sign of the bias when monotonicity relationships hold between the unmeasured confounding variable and the treatment and between the unmeasured confounding variable and the outcome. Some discussion is given to the conditions under which the theorem applies and the strengths and limitations of using the theorem to assess the sign of the bias of unmeasured confounding.

Keywords. Bias; causal inference; observational studies; potential outcomes; unmeasured confounding.

1. Introduction

Causal inference in observational studies often proceeds by the assumption of no unmeasured confounding variables. The researcher engaged in data analysis hopes that sufficiently rich data has been collected so that this assumption is plausible. In many cases, however, the assumption is likely to be violated. Consequently, various sensitivity analysis techniques have been developed to assess the sensitivity of causal conclusions to the presence of unmeasured confounding variables (Cornfield et al., 1959; Rosenbaum and Rubin, 1983; Copas and Li, 1997; Lin, Psaty, and Kronmal, 1998; Robins, Scharfstein, and Rotnitzky, 2000; Brumback et al., 2004; Steenland and Greenland, 2004; McCandless, Gustafson and Levy, 2006). In this paper we show that if certain monotonicity properties hold between the unmeasured confounder and the treatment and the unmeasured confounder and the outcome then one can in some cases determine the sign of the bias of unmeasured confounding. If the estimate without control for the unmeasured confounding variable is of the opposite sign as the sign of the bias then the estimate itself is conservative and can be used to draw conclusions about the presence of a true causal effect. The theorem is given in the potential outcomes frame-
work and formalizes intuitions often drawn upon by epidemiologists and other researchers (e.g. Vander Stoep, Beresford and Weiss, 1999). Equally as important, however, are the circumstances in which intuition fails. We show that the theorem can essentially only be applied to binary treatment variables and that counterintuitive results may arise when this or other of the conditions of the theorem are violated.

The theorem is related to a result given by VanderWeele and Robins (2007) in the context of directed acyclic graphs. The theorem presented in this paper, however, is given within the potential outcomes framework. Furthermore, as discussed below, the theorem here weakens the monotonicity assumptions of VanderWeele and Robins (2007) but requires stronger assumptions concerning conditional independence. The remainder of the paper is organized as follows. Section 2 reviews the potential outcomes framework and the conditionally ignorable treatment assignment assumption necessary to draw conclusions about causal effects. Section 3 presents the result concerning the sign of the bias of unmeasured confounding. In section 4 two examples are given, one substantive and one numerical. Section 5 offers some concluding remarks concerning how the result can be applied to actual data analysis.

2. Potential outcomes and ignorable treatment assignment

We assume a deterministic potential outcomes framework (Rubin 1974, 1978, 1990). We will let \( \Omega \) denote the sample space of individuals in the population and we will use \( \omega \) for a particular sample point. We will use the notation \( A \perp \!
\!
\! \perp B \mid C \) to denote that \( A \) is independent of \( B \) given \( C \). Let \( A(\omega) \) denote the treatment received by individual \( \omega \). Let \( Y(\omega) \) denote some post-treatment outcome for individual \( \omega \). Let \( Y_a(\omega) \) denote the counterfactual value of \( Y \) for individual \( \omega \) if treatment \( A \) were set, possibly contrary to fact, to the value \( a \). Note that we assume that the counterfactual value \( Y_a(\omega) \) for individual \( \omega \) does not depend on the treatments received by other individuals. This assumption is sometimes referred to as SUTVA, the stable unit treatment value assumption (Rubin, 1990). We will use \( Y_a(\omega) \) and
\(Y_{A=a}(\omega)\) interchangeably. Mathematically, although it could be that \(Y_{A(\omega)}(\omega) \neq Y(\omega)\), we will require the "consistency" assumption that \(Y_{A(\omega)}(\omega) = Y(\omega)\) i.e. that the value of \(Y\) which would have been observed if \(A\) had been set to what it in fact was is equal to the value of \(Y\) which was in fact observed. Thus the only counterfactual outcome for individual \(\omega\) that is observed is the counterfactual outcome \(Y_{A(\omega)}\), the value of \(Y\) which would have been observed if \(A\) was set to what it in fact was.

We say that treatment assignment \(A\) is ignorable (or weakly ignorable) given covariates \(X\) if \(Y_{a} \prod A | X\) for all \(a\) and if \(0 < P(A = a | X = x) < 1\) for all \(a\) and \(x\). This condition is sometimes also referred to as the assumption of no unmeasured confounders. It is easily verified that if treatment assignment \(A\) is ignorable given covariates \(X\) then for all \(a\), \(E(Y_a) = \sum_x E(Y | A = a, X = x)P(X = x)\). The right hand side of the equation, unlike the left hand size, is given entirely in observable quantities. When \(X\) contains numerous covariates then the quantity \(E(Y | A = a, X = x)\) is often estimated by means of regression or propensity score analysis. Note that in some of the initial derivations of the this result (e.g. Rosenbaum and Rubin, 1983) a stronger assumption than \(Y_{a} \prod A | X\) for all \(a\) was employed; namely, that \(\{Y_a\} \supp(A) \prod A | X\) where \(\supp(A)\) denotes the support of \(A\). This latter condition is sometimes referred to as strongly ignorable treatment assignment given covariates \(X\).

Ignorable treatment assignment given covariates \(X\) means that within strata of \(X\), treatment gives no information on the distribution of counterfactual outcomes. In practice, a researcher trying to draw causal inferences from observational data attempts to collect a sufficiently large set of variables so that any difference between the different treatment groups are attributable to factors in the set of measured variables \(X\). If this is achieved then, within strata of \(X\), the groups with different levels of treatment variable \(A\) are comparable except with regard to the treatment that they received. Under the assumption of conditionally ignorable treatment assignment, the researcher can use the observed outcomes within strata of \(X\), weighted by the probability of \(X\), as a valid estimate of average counterfactual
3. The sign of the bias of unmeasured confounding

The assumption that treatment assignment is ignorable given covariates $X$ is a strong one and one which is likely to be violated in many settings. In this section we present a theorem which allows the researcher to determine the sign of the bias that results in the presence of unmeasured confounding variables. The theorem requires that monotonicity relations hold between the unmeasured confounder(s) and treatment and also between the unmeasured confounder(s) and the outcome. For the proof of the theorem we will need the following lemma.

**Lemma 1** (Esary, Proschan, and Walkup, 1967, Theorem 2.1): Let $f$ and $g$ be functions with $n$ real-valued arguments such that both $f$ and $g$ are non-decreasing in each of their arguments. If $X = (X_1, ..., X_n)$ is a multivariate random variable with $n$ components such that each component is independent of the other components then $\text{cov}\{f(X), g(X)\} \geq 0$.

We can now state and prove the theorem concerning the sign of the bias of unmeasured confounding variables.

**Theorem 1**: Suppose that $A$ is binary and that (1) $Y_{A=a} \prod_A \{X, C\}$ for $a = 0, 1$ for some univariate $C$, (2) $E(Y|A = a, X = x, C = c)$ is non-decreasing in $c$ for all $a$ and $x$ and (3) $E(A|X = x, C = c)$ is non-decreasing in $c$ for all $x$ then $\sum_x E(Y|A = 1, X = x)P(X = x) \geq E(Y_{A=1})$ and $\sum_x E(Y|A = 0, X = x)P(X = x) \leq E(Y_{A=0})$. Furthermore, if $C$ is multivariate then the conclusion still holds if (4) the components of $C$ are conditionally independent given $X$.

**Proof**: Let $v_x(c) = \frac{P(A=1|X=x,C=c)}{P(A=1|X=x)} = \frac{E(A|X=x,C=c)}{P(A=1|X=x)}$. Then $v_x(c)$ is non-decreasing in $c$ since $E(A|X = x, C = c)$ is non-decreasing in $c$. Furthermore, by Bayes' Theorem,
The inequality follows because by letting confounding variables. If conditions (1)-(4) hold then the researcher to determine the sign of the bias that results in the presence of unmeasured and not controlling for the unmeasured confounding variable \( X \).

Note that condition (4) is trivially satisfied if \( x \) are conditionally independent given \( E \). We can show that \( S_1 \geq E(Y_{A=1}) \) as follows:

\[
S_1 = \sum_x E(Y|A=1, X=x)P(X=x)
\]

\[
= \sum_x \{ \sum_c E(Y|A=1, X=x, C=c)P(C=c|A=1, X=x) \} P(X=x)
\]

\[
= \sum_x \{ \sum_c E(Y|A=1, X=x, C=c) v_x(c) P(C=c|X=x) \} P(X=x)
\]

\[
\geq \sum_x \{ \sum_c E_{F_C|X=x} \{ E(Y|A=1, X=x, C) v_x(C) \} P(X=x) \}
\]

\[
= \sum_x \{ \sum_c E(Y|A=1, X=x, C=c) P(C=c|X=x) \} P(X=x)
\]

\[
= \sum_x \{ \sum_c E(Y|A=1, X=x, C=c) P(C=c, X=x) \}
\]

\[
= E(Y_{A=1}).
\]

The inequality follows because by letting \( g(x,c) = E(Y|A=1, X=x, C=c) \) we have by Lemma 1 that \( E_{F_C|X=x} \{ g(x,C) v_x(C) \} - E_{F_C|X=x} \{ g(x,C) \} = Cov_{F_C|X=x} \{ g(x,C), v_x(C) \} \geq 0 \) since both \( g(x,c) = E(Y|A=1, X=x, C=c) \) and \( v_x(c) \) are non-decreasing in \( c \) and the components of \( C \) are conditionally independent given \( X \). The proof that \( S_0 = \sum_x E(Y|A=0, X=x) P(X=x) \leq E(Y_{A=0}) \) is similar.

Note that condition (4) is trivially satisfied if \( C \) is univariate. The theorem can allow the researcher to determine the sign of the bias that results in the presence of unmeasured confounding variables. If conditions (1)-(4) hold then \( \sum_x E(Y|A=1, X=x) P(X=x) - \sum_x E(Y|A=0, X=x) P(X=x) \geq E(Y_{A=1}) - E(Y_{A=0}) \). In words, under the conditions of Theorem 1, the prima facie estimate of the causal effect controlling only for \( X \) and not controlling for the unmeasured confounding variable \( C \) is greater than the true causal effect. Consequently, in this case, if \( \sum_x E(Y|A=1, X=x) P(X=x) - \sum_x E(Y|A=0, X=x) P(X=x) \geq E(Y_{A=1}) - E(Y_{A=0}) \).
If one of \( E(Y|A = a, X = x, C = c) \) or \( E(A|X = x, C = c) \) is non-decreasing in \( C \) and the other is non-increasing in \( C \) then the conclusions of the theorem apply with the inequality signs reversed so that the prima facie estimate of the causal effect controlling only for \( X \) and not controlling for the unmeasured confounding variable \( C \) is less than the true causal effect; and thus, in this case, if \( \sum_x E(Y|A = 1, X = x)P(X = x) - \sum_x E(Y|A = 0, X = x)P(X = x) \) is positive then it is conservative for \( E(Y_{A=1}) - E(Y_{A=0}) \). An example illustrating the application of the theorem is given in the following section. Theorem 1 can be easily applied to measures of effect other than the causal risk difference. For example, if the conditions of Theorem 1 hold and the outcome \( Y \) is binary then for the causal risk ratio one could conclude that

\[
\frac{\sum_x P(Y=1|A=1,X=x)P(X=x)}{P(Y=1|A=0,X=x)P(X=x)} \geq \frac{P(Y_{A=1}=1)}{P(Y_{A=0}=1)}
\]

and for the causal odds ratio one could conclude that

\[
\frac{\sum_x P(Y=1|A=1,X=x)P(X=x)}{\sum_x P(Y=0|A=1,X=x)P(X=x)} \geq \frac{P(Y_{A=1}=1)/P(Y_{A=1}=0)}{P(Y_{A=0}=1)/P(Y_{A=0}=0)}
\]

In the case of a binary outcome, binary exposure and categorical unmeasured confounder, Flanders and Khoury (1990) give a similar result for the risk ratio. Because of the simplifying assumptions that the outcome is binary and that the unmeasured confounder is categorical, Flanders and Khoury (1990) are also able to derive both upper and lowers bounds on the risk ratio. Theorem 1, however, is more general in that it applies to binary, ordinal or continuous outcomes and to multiple binary, ordinal or continuous unmeasured confounding variables. The results of Flanders and Khoury (1990) also require further simplifying assumptions when adjustment is also made for measured covariates \( X \).

Note also that if some component \( C_i \) of \( C \) is such that \( E(Y|A = a, X = x, C = c) \) and \( E(A|X = x, C = c) \) are non-increasing rather than non-decreasing in \( C_i \) then the result may be employed by replacing \( C_i \) with its negation \(-C_i\). In any case, however, the monotonicity relations (2) and (3) must hold for the unmeasured confounding variable and so it is necessary for the researcher to have some sense as to what this unmeasured confounding variable is in order to evaluate the monotonicity relationships.
As noted above, if $C$ is univariate (i.e. there is only one unmeasured confounding variable) then assumption (4) of Theorem 1 is trivially satisfied. If $C$ is multivariate, the conditional independence of the components of $C$ given $X$ is likely to be difficult to assess. If some knowledge of the relationships between the components of $C$ is available, then directed acyclic graphs (Pearl, 1995; Greenland, Pearl and Robins, 1999) can be useful in assessing the conditional independence requirements of assumption (4). Otherwise it will be difficult to use the result in practice unless $C$ is univariate. As noted in the introduction, the theorem presented in this paper is related to a result given by VanderWeele and Robins (2007) in the context of directed acyclic graphs. The result given by VanderWeele and Robins (2007) made stronger monotonicity assumptions; namely, that $P(Y > y|A = a, X = x, C = c)$ and $P(A > a|X = x, C = c)$ were non-decreasing in $c$ for all $y, a$ and $x$ (i.e. monotonicity in distribution and not just expectation). However, these stronger monotonicity assumptions also allowed for conclusions about the sign of the bias of unmeasured confounding in settings in which the components of $C$ were not conditionally independent given $X$. Note that the monotonicity of $P(Y > y|A = a, X = x, C = c)$ and $P(A > a|X = x, C = c)$ in $c$ implies the monotonicity of $E(Y|A = a, X = x, C = c)$ and $E(A|X = x, C = c)$ in $c$.

4. Examples

In example 1 we give a sample study in which the result presented above may be of use.

Example 1: Androgen deprivation therapy (ADT) is the standard treatment for prostate cancer. However, the side effects of ADT include fatigue, weakness, falls and physiological frailty. Because of the negative side effects, physicians must decide whether or not, and when, to administer ADT by balancing the beneficial effects of ADT and the adverse effects. Older men who are already frail may be less likely to receive ADT. Younger men with aggressive disease will generally be more likely to receive ADT. Shahinian et al. (2005) conduct a cohort study to measure the effect of ADT on the occurrence fractures. Regression analyses
are used to control for age, race, grade of prostate cancer, other cancer treatments received and the occurrence of a fracture or the diagnosis of osteoporosis during the twelve months preceding the diagnosis of cancer. Let \( X \) denote the aforementioned covariates; let \( A \) denote ADT; let \( Y \) denote the occurrence of a fracture during the period from from one year after the diagnosis of cancer to the end of the study. Note that the study did not measure and control for physicians’ subjective assessment at baseline of the patients’ maneuverability \( C \) as they watched the patients attempt to move about the room. Those with a high level of maneuverability \( C \) will in general be less likely to be frail and to experience a fracture but possibly more likely to receive ADT. It thus seems likely that \( E(Y|A = a, X = x, C = c) \) is non-increasing in \( c \) and that \( E(A|X = x, C = c) \) is non-decreasing in \( c \). Based on number needed to harm data reported by Shahinian et al. (2005), it can be verified that 

\[
\sum_x E(Y|A = 1, X = x)P(X = x) - \sum_x E(Y|A = 0, X = x)P(X = x) = 0.036.
\]

Suppose that \( Y_{A=a} \prod A|\{X, C\} \). We apply an analogue of Theorem 1 that when \( E(Y|A = a, X = x, C = c) \) is non-increasing in \( c \) and when \( E(A|X = x, C = c) \) is non-decreasing in \( c \), if \( Y_{A=a} \prod A|\{X, C\} \) then 

\[
\sum_x E(Y|A = 1, X = x)P(X = x) \leq E(Y_{A=1}) \quad \text{and} \quad \sum_x E(Y|A = 0, X = x)P(X = x) \geq E(Y_{A=0}).
\]

Thus \( E(Y_{A=1}) - E(Y_{A=0}) \geq \sum_x E(Y|A = 1, X = x)P(X = x) - \sum_x E(Y|A = 0, X = x)P(X = x) = 0.036 \). The estimate of 0.036 for the average causal effect of ADT on the occurrence fractures is thus conservative.

The theorem presented above required that the treatment variable \( A \) be binary. The result also holds when \( A \) is not binary if \( A = 1 \) is replaced with the maximum value of \( A \) and if \( A = 0 \) is replaced with the minimum value of \( A \). However counterexamples can be constructed to demonstrate that the result cannot be generalized beyond the extreme values of the intervention variable. In Example 2 we numerically illustrate the application of Theorem 1 to the maximum and minimum of a treatment variable \( A \) and we also give a numerical illustration that shows that the theorem cannot be applied to intermediate values of the treatment variable.
Example 2: Suppose that $A$ is ternary and takes values in the set $\{1, 2, 3\}$ and suppose that a single binary variable $C$ confounds the relationship between $A$ and $Y$ so that $Y_a \prod A | C$. Suppose $P(C = 0) = P(C = 1) = 1/2$. Suppose $E(Y|A = 1, C = c) = 1/2 + c/5$ and $E(Y|A = 2, C = c) = 2/5 + c/5$ and $E(Y|A = 3, C = c) = 1/5 + c/5$; then $C$ satisfies monotonicity condition (2). Suppose the conditional distribution of $A$ given $C = 0$ is $P(A = 1|C = 0) = 8/10$; $P(A = 2|C = 0) = 1/10$; $P(A = 3|C = 0) = 1/10$ and that the conditional distribution of $A$ given $C = 1$ is $P(A = 1|C = 1) = 1/10$; $P(A = 2|C = 1) = 7/10$; $P(A = 3|C = 1) = 2/10$; then $E(A|C = 0) = 13/10$ and $E(A|C = 1) = 21/10$ so monotonicity condition (3) is satisfied. Clearly we also have $0 < P(A = a|C = c) < 1$ for all $a$ and $c$. We can also calculate conditional distribution of $C$ given $A$: $P(C = 1|A = 1) = 1/9$; $P(C = 1|A = 2) = 7/8$; $P(C = 1|A = 3) = 2/3$. And we can furthermore calculate the following: $E(Y|A = 1) = \sum_c E(Y|A = 1, C = c)P(C = c|A = 1) = \sum_c (1/2 + c/5)P(C = c|A = 1) = 47/90$ and $E(Y|A = 2) = \sum_c E(Y|A = 2, C = c)P(C = c|A = 1) = \sum_c (2/5 + c/5)P(C = c|A = 2) = 23/40$ and $E(Y|A = 3) = \sum_c E(Y|A = 3, C = c)P(C = c|A = 1) = \sum_c (1/5 + c/5)P(C = c|A = 3) = 1/3$. The true causal effects are given by: $E(Y_{A=1}) = \sum_c E(Y|A = 1, C = c)P(C = c) = \sum_c (1/2 + c/5)P(C = c) = 6/10$ and $E(Y_{A=2}) = \sum_c E(Y|A = 2, C = c)P(C = c) = \sum_c (2/5 + c/5)P(C = c) = 5/10$ and $E(Y_{A=3}) = \sum_c E(Y|A = 3, C = c)P(C = c) = \sum_c (1/5 + c/5)P(C = c) = 3/10$.

If we had data only on $Y$ and $A$ but knew that the monotonicity conditions (2) and (3) held we could calculate $E(Y|A = 3) - E(Y|A = 1) = 1/3 - 47/90 = -17/90$. We would know from Theorem 1 that $E(Y|A = 3) \geq E(Y_{A=3})$ and $E(Y|A = 1) \leq E(Y_{A=1})$ and thus that $E(Y_{A=3}) - E(Y_{A=1}) \leq E(Y|A = 3) - E(Y|A = 1)$. We could then conclude that the estimate $-17/90$ was an overestimate, that the true causal effect was in fact lower than this estimate and thus that the estimate $-17/90$ was conservative. In fact $E(Y_{A=3}) - E(Y_{A=1}) = 3/10 - 6/10 = -3/10 = -27/90$.

Now consider $A = 2$. In this numerical example, $E(Y|A = 2) = 23/40 > 5/10 =
Suppose now that the conditional distribution of $A$ given $C = 0$ had been $P(A = 1|C = 0) = 1/10$; $P(A = 2|C = 0) = 8/10$; $P(A = 3|C = 0) = 1/10$ and that the conditional distribution of $A$ given $C = 1$ had been $P(A = 1|C = 1) = 1/10$; $P(A = 2|C = 1) = 1/10$; $P(A = 3|C = 1) = 8/10$; then $E(A|C = 0) = 20/10$ and $E(A|C = 1) = 27/10$ so monotonicity condition (3) would again satisfied. However, this time $P(C = 1|A = 2) = 1/9$ and $E(Y|A = 2) = \sum_c E(Y|A = 2, C = c)P(C = c|A = 1) = \sum_c (2/5 + c/5)P(C = c|A = 2) = 19/45 < 5/10 = E(Y_{A=2})$. Thus under a different conditional distribution of $A$ given $C$ we would have the conditions of Theorem 1 holding but $E(Y|A = 2) < E(Y_{A=2})$ instead of $E(Y|A = 2) > E(Y_{A=2})$ as was the case with the first distribution. For intermediate values of the intervention variable we thus see that the bias when control for confounding is inadequate may be in either direction even when conditions (1)-(4) are satisfied.

5. Concluding remarks

Although Theorem 1 does not hold for intermediate values of the intervention variable, the result may still be useful in settings in which the intervention variable $A$ is ordinal or continuous. A non-binary intervention variable may be dichotomized at various cut-off points. The analysis may proceed with this dichotomized intervention variable with the theorem employed to assess the sign of the bias at the various points of dichotomization and conclusions can then be drawn from the resulting analyses.

Although there are many existing sensitivity analysis techniques to address problems of unmeasured confounding (Rosenbaum and Rubin, 1983; Copas and Li, 1997; Lin et al., 1998; Robins et al., 2000; Brumback et al., 2004; Steenland and Greenland, 2004; McCandless, Gustafson and Levy, 2006) often these techniques cannot be easily implemented by standard statistical software and require special programming. Other results provide bounds for causal effects in the presence of unmeasured confounding (Cornfield et al., 1959; Yanagawa, 1984;
Manski, 1990; Balke and Pearl, 1997; MacLehose et al., 2005) but these bounds are sometimes
too wide to be of particular use. The theorem we have provided in this paper allows the
researcher to easily and intuitively draw conclusions about the sign of the bias of unmeasured
confounding. If the bias is of the opposite sign as the sign of the estimate without control
for the unmeasured confounding variable(s) then this estimate with inadequate control for
confounding is conservative for the true causal effect. The limitations of the theorem
presented above are that it essentially applies only to binary treatment variables, that it
requires monotonicity assumptions (i.e. that the researcher have knowledge of the direction
of the exposure-confounder and disease-confounder associations) and that in the case of a
multivariate confounding variable \( C \) it requires certain conditional independence assumptions
which may be difficult to assess in practice. Many other sensitivity analysis techniques
impose other assumptions and do not make monotonicity assumptions (e.g. Rosenbaum
and Rubin, 1983; Steenland and Greenland, 2004; McCandless, Gustafson and Levy, 2006).
Nevertheless, when a single unmeasured confounding variable is present and the researcher
has some sense as to what the unmeasured confounding variable might be then the conditions
of the theorem given above will often hold and conclusions about the sign of the bias can
then be drawn.

References

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect

analyses for unmeasured confounding assuming a marginal structural model for repeated

Statistical Society, Series B* 59, 55-95.


