Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2010;**39**:417–420

Advance Access publication 19 November 2009
doi:10.1093/ije/dyp334

METHODOLOGY

Illustrating bias due to conditioning on a collider

Stephen R Cole,[1]* Robert W Platt,[2] Enrique F Schisterman,[3] Haitao Chu,[4] Daniel Westreich,[1] David Richardson[1] and Charles Poole[1]

Accepted 1 October 2009

That conditioning on a common effect of exposure and outcome may cause selection, or collider-stratification, bias is not intuitive. We provide two hypothetical examples to convey concepts underlying bias due to conditioning on a collider. In the first example, fever is a common effect of influenza and consumption of a tainted egg-salad sandwich. In the second example, case-status is a common effect of a genotype and an environmental factor. In both examples, conditioning on the common effect imparts an association between two otherwise independent variables; we call this selection bias.

Keywords Bias, selection, methods, epidemiologic

Introduction

Epidemiologists, like researchers in many fields, give a great deal of attention to potential sources of bias in their study findings. As opposed to imprecision, biases (or systematic errors) abide as the sample size grows. Epidemiologists immediately understand why an uncontrolled common cause of exposure and outcome causes bias. We call this confounding.[1] In our experience, epidemiologists have a more difficult time understanding why conditioning (implicitly or explicitly) on, or controlling for, a common 'effect' of exposure and outcome may cause bias. We call this selection bias,[2] collider-stratification bias[3] or bias due to conditioning on a collider. Note that here the term 'conditioning' refers to restriction (by design or analysis), stratification or regression adjustment. Although some examples of collider-stratification bias have been published in the epidemiologic literature,[2,4–6] here we present two simple hypothetical examples that may help to convey some concepts underlying bias due to conditioning on a collider. First, we briefly introduce causal diagrams.

Diagrams have been used to encode knowledge about systems of variables in epidemiology for decades.[7] Recently, Pearl[8] formalized causal diagrams as directed acyclic graphs, providing investigators with a powerful tool for bias assessment, if the rules of causal diagrams are followed. Rules for working with causal diagrams are given by Greenland *et al*.[9] and succinctly in the Appendix of Hernán *et al*.[10] Causal diagrams link variables by single-headed (i.e. directed) arrows that represent direct causal effects. For a diagram to represent a causal system, all common causes of any pair of variables included on the diagram must also be included on the diagram. The absence of an arrow between two variables is a strong claim of no direct effect of the former variable on the latter. We denote conditioning by placing a box around the conditioned variable.

[1] Department of Epidemiology, UNC Gillings School of Global Public Health, Chapel Hill, NC, USA.
[2] Department of Biostatistics, McGill University, Montreal, Canada.
[3] Epidemiology Branch, *Eunice Kennedy Shriver* NICHD, NIH, Bethesda, MD, USA.
[4] Department of Biostatistics, UNC Gillings School of Global Public Health, Chapel Hill, NC, USA.
* Corresponding author. Department of Epidemiology, UNC Gillings School of Global Public Health, Chapel Hill, NC, USA. E-mail: cole@unc.edu

Table 1 Data illustrative of selection bias, due to conditioning on a collider

	Influenza				
	Yes	No	Total	Risk	Risk difference
Panel A					
Sandwich					
Chicken	5	45	50	0.1	0.0
Egg salad	5	45	50	0.1	
Panel B					
Fever					
Sandwich					
Chicken	5	0	5	1.0	0.9
Egg salad	5	45	50	0.1	
No fever					
Sandwich					
Chicken	0	45	45	0.0	NA
Egg salad	0	0	0	NA	

Example one

Suppose that you attend a meeting with 99 of your peers. Unbeknownst to you, 10 were pre-symptomatically infected with influenza upon arrival. We ignore transmission, the effects of which we would not see for a few days. Everyone at the meeting ate one of 50 chicken or 50 an egg-salad sandwiches in a boxed lunch, chosen at random. Without further information, what is the expected number of influenza cases in the 50 people who ate a chicken sandwich? The answer is five; and the same is true for the 50 people who ate an egg-salad sandwich. The data in Table 1 illustrate this example. In panel A of Table 1, the risk of influenza is 5/50 in both sandwich groups, and the risk difference is 0. Because sandwich type was randomized, we expect no association between pre-existing influenza and sandwich type. Ignoring chance imbalances due to the relatively small sample size, it is clear that the data in panel A of Table 1 represent what we would expect to see in an infinite sample (one can think of each of the 50 people as representing a much larger number of homogenous individuals).

That evening, you and 54 others develop a 102°F fever. Let's say that in our hypothetical world there are only two ways to get such a fever: influenza or consuming 1 of the 50 tainted egg-salad sandwiches. Among those individuals with a fever, therefore, all were exposed to either influenza or an egg-salad sandwich (or both). Put another way, restricting our attention to only those individuals with a fever or conditioning on (stratifying by) the variable fever, we have conditioned on a 'common effect' of both influenza and sandwich type.

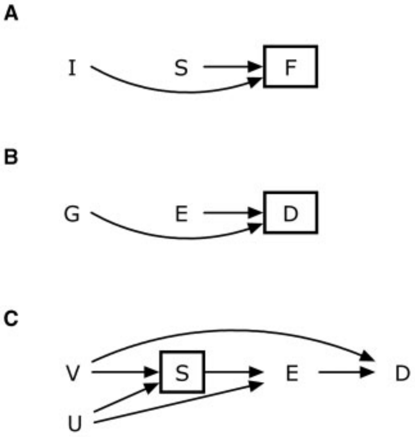

Figure 1 Causal diagrams depicting scenarios described in Example 1 (**A**), Example 2 (**B**) and in Discussion (**C**)

Therefore, knowing that you have a fever, if we ask you whether you ate an egg-salad sandwich and you respond 'no', then we know that your fever is due to having influenza. In panel B of Table 1, we see that, among those with a fever, the influenza risk among those who ate a chicken sandwich is $5/5 = 1$, compared with 5/50 for those who ate an egg-salad sandwich, yielding a risk difference of 0.9. (Conversely, all individuals without a fever were exposed to neither influenza nor tainted egg-salad.) This association is introduced by conditioning on a common effect, namely fever. Recall that sandwich type was randomly assigned. This association was not present before we knew about (and conditioned on) your fever status.

In general, we may introduce bias by conditioning on common effects of otherwise unrelated variables: we call this selection bias. Consider the variables influenza *I*, sandwich type *S* and fever *F*. A causal diagram[8] illustrating the associations as described previously is drawn in Figure 1A. A variable like *F*, where two arrowheads meet, is called a 'collider' on the 'path' *I–F–S*; but may not be a collider on other paths (if they existed). When we condition on a collider, we may introduce associations in one or more strata that were not present in the source population. One way to understand the bias caused by conditioning on a collider is to envision a connection made between *I* and *S* once *F* is conditioned upon; indeed, some methods for working with causal diagrams explicitly draw such connections.[9] The bias in this example is not in the association of exposure (*S*) with disease (*F*), but it is in the apparent *I–S* association, within one or both levels of *F*.

Example two

Another example of selection bias is the setting of a case-only study, which has become popular in the study of gene–environment interactions.[11] This design is used to measure departures from a multiplicative

Table 2 A source population for a *G–E* case-only study of a source population in which the RRs are not constant[a]

G	*E*	Risk	RR	Number	Cases	Non-cases
Yes	Yes	0.9	3.0	100	90	10
	No	0.3	1.	50	15	35
No	Yes	0.2	2.0	200	40	160
	No	0.1	1.	100	10	90

[a] *G* and *E* are unassociated in the population but associated in both levels of the collider, disease (i.e. among the cases and the non-cases). The case-only *G–E* OR of 1.5 equals the ratio of RRs.

model of a constant risk or rate ratio (RR) under assumptions that the environmental variable (E) and the genetic variable (G) are independently distributed in the source population and that the estimated effect of at least one of them on the disease (D) is not confounded. These conditions are present if G is subject to Mendelian randomization[12] and G does not affect E.

Let A_{GE} and N_{GE} represent the number of incident cases and the person-time at risk in each of the four combinations of $G=0$, 1 and $E=0$, 1. Then the rate is $R_{GE}=A_{GE}/N_{GE}$, and the ratio of RRs, $(R_{11}/R_{10})/(R_{01}/R_{00})$, may be written as a ratio of odds ratios (ORs) as $(((A_{11}/A_{10})/(A_{01}/A_{00}))/((N_{11}/N_{10})/(N_{01}/N_{00})))$. Given the independence of G and E in the source population, the denominator of this last expression is unity. The numerator of this last expression is the case-only G–E OR, or $\mathrm{OR}_{GE|D=1}$, and equals the ratio of RRs.

When E and G both affect D, a causal diagram is drawn in Figure 1B, and D is a collider on the path G–D–E. Unconditionally (i.e. in the source population for cases in the case-only study), E and G are not associated. However, when we condition on D, we expect to find an association between E and G within at least one stratum of D because of selection bias. In the case-only design, this conditioning is accomplished by restricting the study to the cases.

For illustration, it is useful to consider the full underlying cohort study and examine the G–E association in both levels of D (i.e. among the cases and non-cases). In the hypothetical data in Table 2, G and E are unassociated in the persons at risk ($\mathrm{OR}_{GE}=1.0$), but G and E are associated within levels of D. Among the non-cases, the $\mathrm{OR}_{GE|D=0}=0.16$. Among the cases, the $\mathrm{OR}_{GE|D=1}=1.5$, which equals the ratio of the RRs (i.e. $3.0/2.0=1.5$).

Under the design assumptions of the case-only study, the case-only G–E OR is an unbiased estimate of the ratio of RRs. The selection bias is not in the estimate of that measure, but in the G–E association within one or both levels of D as an estimate of the G–E association in the source population. In Table 2, OR_{GE} equals 1, but $\mathrm{OR}_{GE|D=0}\neq 1$ and $\mathrm{OR}_{GE|D=1}\neq 1$. When the ratio of RRs equals 1, $\mathrm{OR}_{GE|D=1}$ equals 1 as well and the overall G–E association is distorted only among the non-cases, $\mathrm{OR}_{GE|D=0}\neq 1$.

Discussion

This note is about the distortion of an association between two variables that occurs by conditioning on a common effect; we call this selection bias. Whether through restriction, stratification or regression adjustment, the result in this setting is a selection bias that is distinct from bias due to confounding.

Some confusion between confounding and selection bias may have resulted from definitions of confounding,[13] which were not explicit about confounding being a bias due to the existence of a common cause of exposure and outcome. By definition, a common cause must be 'temporally prior' to both exposure and outcome. However, a measured confounder may be temporally 'posterior' to exposure if it is on a causal pathway from the common cause to the outcome, or temporally posterior to 'both' exposure and outcome if it is a descendant of the common cause. In contrast, by definition, a common effect must be temporally posterior to both exposure and outcome. Selection bias results from conditioning on such a common effect. In the present examples the collider is the 'outcome' itself rather than a variable along a path that would cause bias in an exposure–outcome association. We chose such examples to minimize the number of variables needed. Selection bias may also occur when the selection happens temporally 'prior' to exposure, such as when we condition on a common effect of causes of exposure and outcome as shown in Figure 1C. In many real-data settings one cannot determine whether a purported bias is due to confounding or selection and the distinction may not matter. In addition to confounding and selection bias, causal diagrams have been used to depict direct and indirect effects,[14] bias due to overadjustment,[15] time varying confounding[16] and time modified confounding.[17]

In summary, when one has conditioned (by design or analysis) on a common effect of a pair of variables then there is likely to be a spurious association between this pair of variables, which is due to selection bias. Of course, our examples are contrived. In the first, we ask that you ignore transmission and believe in a world with only two causes of fever and in both examples we ask that you ignore sampling variability. Our only defense for these artificialities is that sometimes clarity is inversely related to the complexities of life.

Funding

Grants R03-AI-071763, R01-AA-01759 and P30-AI-50410 (to S.R.C.); Lineberger Cancer Center Core Grant CA16086 and P30-AI-50410 (to H.C.); Intramural Research Program of the *Eunice Kennedy Shriver* NICHD, NIH; and the American Chemical Council to (to E.S.); P30-ES-10126 (to C.P.).

Acknowledgements

The authors would like to thank the members of the Causal Inference Research Group at the UNC Gillings School of Global Public Health for thoughtful discussions.

Conflict of interest: None declared.

KEY MESSAGES

- Conditioning, by design or analysis, on a common effect of a pair of variables may cause a spurious association (i.e. selection bias) between this pair of variables.
- Intuition for such bias may be gained by studying simple examples, as presented herein.

References

[1] Greenland S, Morgenstern H. Confounding in health research. *Annu Rev Public Health* 2001;**22:**189–212.

[2] Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15:**615–25.

[3] Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003;**14:**300–6.

[4] Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001;**12:**313–20.

[5] Hernandez-Diaz S, Schisterman EF, Hernan MA. The birth weight paradox uncovered? *Am J Epidemiol* 2006; **164:**1115–20.

[6] VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol* 2007;**166:**1096–104.

[7] Breslow N. Design and analysis of case–control studies. *Annu Rev Public Health* 1982;**3:**29–54.

[8] Pearl J. Causal diagrams for empirical research. *Biometrika* 1995;**82:**669–710.

[9] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10:**37–48.

[10] Hernán MA, Hernandez-Diaz S, Werler MM *et al*. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;**155:**176–84.

[11] Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction:case-control studies with no controls! *Am J Epidemiol* 1996;**144:**207–13.

[12] Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32:**1–22.

[13] Last JM. *A Dictionary of Epidemiology*. 4th edn. New York: Oxford University Press, 2001.

[14] Cole SR, Hernán MA. Fallibility in estimating direct effects (with discussion). *Int J Epidemiol* 2002;**31:**163–65.

[15] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 2009;**20:**488–95.

[16] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;**11:**550–60.

[17] Platt RW, Schisterman EF, Cole SR. Time-modified confounding. *Am J Epidemiol* 2009;**170:**687–94.