

Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias

Sander Greenland

Abstract: It has long been known that stratifying on variables affected by the study exposure can create selection bias. More recently it has been shown that stratifying on a variable that precedes exposure and disease can induce confounding, even if there is no confounding in the unstratified (crude) estimate. This paper examines the relative magnitudes of these biases under some simple causal models in which the stratification variable is graphically depicted as a collider (a variable directly

affected by two or more other variables in the graph). The results suggest that bias from stratifying on variables affected by exposure and disease may often be comparable in size with bias from classical confounding (bias from failing to stratify on a common cause of exposure and disease), whereas other biases from collider stratification may tend to be much smaller. (EPIDEMIOLOGY 2003;14:300–306)

Key words: adjustment, causal diagrams, causal inference, odds ratio, relative risk, validity.

There is now an extensive theory of causal inference based on graphical probability models and their relation to models for potential outcomes.^{1,2} Most of this theory focuses on qualitative results; see Greenland *et al.*,³ Robins,⁴ Kaufman and Kaufman,⁵ Hernán *et al.*,⁶ Cole and Hernán⁷ and Greenland and Brumback⁸ for epidemiologic examples. A central point in this literature is that stratifying (conditioning) on a variable C will alter associations among its causes. For example, if X and Y are marginally independent (*ie*, unassociated before stratification), then they will be associated within at least one stratum of a variable that they both affect.^{1(pp17)}

The simplest such situation is represented by an “inverted fork” $X \rightarrow C \leftarrow Y$, where the arrow represents a direct effect of the tail variable on the head variable; C is then called a “collider” on the X - C - Y pathway in the graph, and stratifying on C will tend to change the association of X and Y . Depending on the relation of X , Y and C to the study exposure and disease, this change can in turn lead to biases in effect estimation; these

biases can be viewed as generalizations of Berkson’s bias⁹ beyond hospital-based studies. This paper examines the sizes of these induced biases under some basic models for studying the effect of an exposure E on a health outcome D . The results suggest that the biases may tend to be largest (comparable in size with classical confounding) if the collider C is affected by E and D , smaller if C is affected by E but not D , and even smaller if C is not affected by E or D .

For the most part I will argue informally and consider only binary variables whose effects can be well approximated by a constant odds ratio model; I discuss technical details and more general models in the Appendix. Parallel findings for linear structural models have recently been presented by Peter Spirtes.³³ All the biases discussed are discrepancies between marginal (“crude”) and conditional (stratum-specific) population associations, and hence correspond to large-sample (asymptotic) biases. Issues of measurement error and sampling variability will be addressed only briefly.

Classical Confounding and Berksonian Bias

Suppose we wish to estimate the causal effect on the frequency of disease ($D = 1$) of having everyone exposed ($E = 1$) *vs* everyone unexposed ($E = 0$) in a population that is both our source of study subjects and our target for inference. Suppose also that we have observations on some or all individuals in that population, but exposure is not randomized. If we see the E - D association change upon stratification by C , to what extent does this change represent removal (control) of bias or, instead, introduc-

From the Departments of Epidemiology and Statistics, University of California Los Angeles, Los Angeles, CA.

Address correspondence to: Sander Greenland, Departments of Epidemiology and Statistics, University of California Los Angeles, Los Angeles, CA 90095-1772; lesdomes@ucla.edu

This research was partially supported by grant R01 HD-39746 from the National Institute of Child Health and Human Development.

Submitted 27 March 2002; final version accepted 8 October 2002.

Copyright © 2003 by Lippincott Williams & Wilkins, Inc.

tion of a bias in estimating this effect? It has long been known that the answer depends on the direction of causal relations among E , C and D .

In simple situations, a change upon stratification by C corresponds to bias removal when C causally affects both E and D (most simply depicted by the “causal fork” $E \leftarrow C \rightarrow D$). Most intuitions about covariate control assume this situation, in which C is a “classical confounder” (causal confounder) of the E - D association.^{1(ch6),3,10(ch8),11} In accord with those intuitions, causal-diagram theory implies that the C -specific E - D associations will equal the C -specific effects of E on D when C is a classical confounder, assuming no other confounder or bias source is present.

Now suppose instead E and D both affect C , as when C is a collider on the E - C - D path (most simply depicted by $E \rightarrow C \leftarrow D$) or is affected by a collider between E and D . In this situation, we should expect at least one C -specific E - D association to differ from the marginal (“crude”) E - D association, so that it will appear as if C is a confounder when in fact it is not; thus, any change in the E - D association upon C -stratification represents bias introduction.¹² In a classic example, Berkson⁹ considered hospital-based studies of cholecystitis ($E = 1$) and diabetes ($D = 1$). Assuming a particular model for the joint E - D effect on hospital admission ($C = 1$), he showed that E and D would be negatively associated in the “hospitalized” ($C = 1$) stratum if E and D were independent in the original source population. Here, the induced bias must be away from the null (no association) because the true E - D effect (which here corresponds to the marginal E - D association) is null.

As another example, let E represent an unopposed-estrogen therapy indicator, D an endometrial-cancer indicator and C an indicator of uterine bleeding (or of endometrial dilation and curettage, which may be performed in response to bleeding problems). In the 1970s, some authors recommended estimating the effect of therapy on cancer among bleeders only or among women undergoing dilation and curettage ($C = 1$), to account for the effects of bleeding on ascertainment of estrogen use and cancer.¹³ It was shown, however, that the relative risk among these women would be much smaller than the true causal relative risk because of the strong effects of both the therapy and the cancer on bleeding risk, and this downward bias would be much larger than the upward bias attributable to differential disease ascertainment.^{14,15} This differed from Berkson’s example in that therapy had a large effect on endometrial cancer risk, and hence the downward bias induced by C -stratification was toward the null.

Quantifying the Biases

Many authors have described how bias attributable to classical confounding (bias attributable to ignoring C

when C is a cause of E and D) is limited by the strength of the association of C with D given E and the association of C with E .¹⁶⁻²¹ In particular, let R_{ED} , R_{CD} , and R_{CE} be the odds ratios for E - D given C , for C - D given E and for C - E given D (each assumed constant across strata of the given variable). Then the ratio of the marginal E - D odds ratio to the C -specific odds ratio R_{ED} will tend to be much closer to 1 (on the ratio scale) than either R_{CD} or R_{CE} ; for examples, see Breslow and Day, Table 3.4,¹⁹ which allows R_{CE} as large as 36, and Table 1 in Yanagawa.²⁰

The amount of classical confounding or Berksonian bias depends not only on the associations among C , E and D , but also on the distribution of C . This may be seen by noting that if $C = 1$ were rare (say, under 5%) in all E - D categories, the E - D odds ratio in the $C = 0$ stratum (which equals R_{ED}) would approximate the marginal E - D odds ratio, and hence bias from improperly ignoring or stratifying on C would be negligible; symmetrically, the bias would be negligible if $C = 0$ were always rare. These and further results follow from examining the ratio of the marginal to the C -specific E - D odds ratio (“crude”/ R_{ED}), which equals:

$$\text{Bias}(R_{CD}, R_{CE}, p) \equiv (R_{CD}R_{CE}p + 1 - p) / ([R_{CD}p + 1 - p][R_{CE}p + 1 - p]) \quad (1)$$

where $p = P(C = 1 | D = E = 0)$ (see Appendix²⁰). Given R_{CD} and R_{CE} , the ratio formula 1 approaches 1 as $P(C = 1)$ approaches 1 or 0. To apply formula 1 to measure confounding by C , one must assume $D = 1$ is rare in all C - E categories, or that the parameters refer to a case-cohort study in which the time of D is ignored. This limitation arises because formula 1 measures non-collapsibility of the study odds ratio, which does not equal confounding by C when $D = 1$ is common.¹¹

Let $G = (R_{CD}R_{CE})^{1/2}$ be the geometric mean of R_{CD} and R_{CE} . Given R_{CD} and R_{CE} , the log of formula 1 has its largest absolute value when $P = 1/(G + 1)$. Substitution of $1/(G + 1)$ for p in formula 1 yields

$$\text{Bias}_{\max}(R_{CD}, R_{CE}) \equiv (G + 1)^2 / (R_{CD} + 2G + R_{CE}) \quad (2)$$

for the maximal ratio of the marginal to the C -specific E - D odds ratio,^{20(corollary 1)} where “maximal” means maximally distant from 1 on the ratio scale. This maximum will be closer to 1 (in ratio terms) than either of R_{CD} and R_{CE} , and approaches 1 as either R_{CD} or R_{CE} approaches 1 with the other held fixed. Furthermore, for a fixed G this maximum is maximized when $R_{CD} = R_{CE} = R$, in which case formula 2 becomes $(R + 1)^2/4R$ (see Appendix). For $R_{CD} = R_{CE} = 2, 4, 8$ and 16 the respective maximal biases are only 1.13, 1.56, 2.53 and 4.52; these are also the maxima when $R_{CD} = R_{CE} = 1/2, 1/4, 1/8$ and $1/16$.

Given a hypothesized value H for confounding by an unmeasured indicator C , we may equate H to formula 1 and see what combinations of p , R_{CD} and R_{CE} produce H . Solving $H = (R + 1)^2/4R$ for R yields the values for $R_{CD} = R_{CE} = R$ closest to 1 that could produce H . For example, the estrogen-endometrial cancer odds ratios initially observed were around 10. To find the R closest to 1 needed to completely explain those observations, we solve $H = 10 = (R + 1)^2/4R$, which yields $R = 1/38$ or 38; thus, a 10-fold bias requires at least one of R_{CD} or R_{CE} (or their inverses) be at least 38. Such computations provide a quick check on the plausibility of claims that an association is only attributable to an uncontrolled factor.

Formulas 1 and 2 measure the ratio of the marginal to the C -specific E - D odds ratio, regardless of any causal relations among the variables, and regardless of whether C stratification is removing or adding bias. To apply them to Berksonian bias, note that the C -specific odds ratio R_{ED} is the biased one, and so formulas 1 and 2 become the inverse and maximum inverse bias. When the biases are below 1, I will use these inverses as bias measures. Formula 2 inverts when one argument is inverted, eg, $\text{Bias}_{\max}(R_{CD}, 1/R_{CE}) = 1/\text{Bias}_{\max}(R_{CD}, R_{CE})$, and so inverting both arguments leaves the expression unchanged.

The above formulas can be used to quantify other biases. For example, consider again the diagram $E \leftarrow C \rightarrow D$ for classical confounding, but applied to a situation in which C is the correct (true) but unknown exposure status and E is an imperfect measure or surrogate for C . The marginal (unadjusted) C - D odds ratio is then the target parameter, but we observe the marginal E - D odds ratio in its place. In the diagram, E is not associated with D upon stratification by C (ie, $R_{ED} = 1$), which means the error in E as a measure of C is nondifferential with respect to D ; as a consequence, the E -specific C - D odds ratio R_{CD} equals the marginal C - D odds ratio. Because $R_{ED} = 1$ and R_{CD} is the target parameter, formula 1 is now the size of the marginal E - D odds ratio, rather than the bias; we may compare it with R_{CD} to measure the bias attributable to misclassification (ie, the bias attributable to using E in place of C). As before, formula 1 is closer to 1 than is R_{CD} , and approaches 1 as $P(C = 1)$ approaches 0 or 1; these properties are just the well-known results that misclassification of a binary exposure produces bias toward the null when the classification (E) is purely random given exposure (C), and that the bias worsens as the exposure prevalence approaches 0 or 1.²²

Bias Due to Adjustment for Exposure Effects

Suppose next that C is affected by E but not by D . If C is an intermediate, as in the diagram $E \rightarrow C \rightarrow D$, the C -specific associations will be biased for the net (total) effect of E on D , in part because the adjustment removes

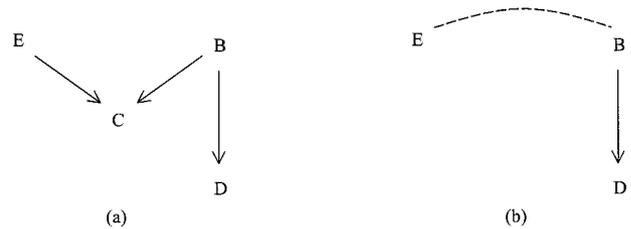


FIGURE 1. The study exposure E affects a covariate C that is also affected by an independent risk factor B for the disease D . Arrows are direct causal effects and *dashed line* is an induced association. (a), Relations before stratification on C (no confounding of the E - D association); (b), Relations after C stratification (E - D association now confounded by B).

some of the E effect.^{23,24} For example, to estimate the net impact of medical interventions ($E = 1$) on health outcomes, one must avoid adjusting for factors on the pathway from the intervention to the outcome. In this (as in the Berksonian) case, any change in the E - D association upon C -stratification represents a bias for estimating the net effect, and formula 1 is the inverse of the bias introduced. This problem is one basis for cautions against conventional adjustments for posttreatment variables in experiments²⁵ and in observational studies.^{10(ch8),26}

These cautions are often ignored, sometimes with the rationale that the intermediate-adjusted association represents the “direct effect” of E on D (ie, the effect of E on D outside the E - C - D pathway).²⁷ This would be so if D shared no causes with E or C ; formula 1 would then be the ratio of the net and direct effects of E on D . This interpretation could be applied to critically analyze the use of C as an “intermediate endpoint” in studies of effects of E on D (for example, if E , C and D were indicators of antihypertensive medication, hypertension and stroke): the fact that formula 1 tends to be much closer to 1 than are R_{CD} or R_{CE} shows how treatment efficacy (the net E - D effect) can be weak even if there are strong treatment effects on the intermediate (R_{CE} far from 1) and strong intermediate effects on the outcome (R_{CD} far from 1), and also shows how this problem worsens for rare or ubiquitous intermediates (ie, $P(C)$ near 0 or 1).

Unfortunately, the interpretation of the C -specific E - D associations as direct effects can be limited by the presence of other causes of the outcome. Even if E is randomized, adjustment for a C affected by E can generate confounding, in which case the change in the E - D association upon C -stratification will partly or wholly represent bias introduction.^{6,7,24} Consider Figure 1a. In such a causal graph, the absence of an arrow between two variables corresponds to absence of a direct effect between them; these absences (along with arrow directions) are central assumptions encoded in a causal

graph.¹ Similarly, absence of a *directed path* (a head-to-tail sequence of arrows) between two variables corresponds to absence of any effect between them. For example, Figure 1a asserts that *E* has no effect on *B* or *D*, *B* has no effect on *E* and *C* has no effect on *D*. Furthermore, because the only path from *B* to *E* passes through the collider *C*, *B* and *E* are marginally independent, and hence *B* cannot confound the *E-D* association.^{1,3,12} Nonetheless, because *B* and *E* both affect *C*, one should expect *E* and *B* to be associated in one or more *C* strata. Figure 1b illustrates that *C*-specific relation, using a dashed line to represent the noncausal *B-E* association; it shows that *B* will confound the *C*-specific *E-D* association, and hence a *C*-specific and *C*-adjusted association will be biased for the direct effect of *E* on *D*. Furthermore, because there is no *C* effect on *D*, there is no indirect effect of *E* on *D* through *C*; thus, the net and direct effects of *E* on *D* are equal, and so the *C*-adjusted association will also be biased for the net *E-D* effect.

Another way to describe the problem is that *B* is a classical confounder of the *C-D* association; thus, if one fails to adjust for *B*, the estimated indirect *E*-effect being “removed” by *C*-stratification will itself be confounded, and the resulting estimate of the direct *E-D* effect will be biased.^{4,6,24} This bias can occur whether or not *C* has an effect on *D*. For example, in a randomized trial of an appetite suppressant (*E* = 1) and mortality (*D* = 1) with weight loss (*C* = 1) as the intermediate, to estimate the direct drug effect one would have to control baseline factors that affect both weight and mortality (such as diabetes), despite the randomization of *E*.

To summarize, under Figure 1a the marginal *E-D* association is unconfounded, so it would properly reflect the lack of an *E* effect on *D* (ie, it would be null). There is no direct or indirect effect of *E* on *D*; thus, if stratification on *C* alone accomplished its goal of estimating the direct *E-D* effect, the *C*-specific *E-D* odds ratio R_{ED} should equal 1. But instead it becomes biased away from 1 in a two-step process: Because *B* and *E* are marginally independent and affect *C*, we should expect a *C*-specific *B-E* association, as in Berkson’s example; this association then couples with the direct *B-D* effect to make *B* a confounder of the *E-D* effect within *C* strata. The resulting bias can occur even if *E*, *B*, or both *E* and *B* are randomized.

Quantifying the Bias

Using odds ratios to measure the effects in Figure 1a, we can apply formula 2 repeatedly to bound the bias produced by controlling *C* alone. For example, suppose the causal odds ratios corresponding to the arrows in Figure 1 are $R_{CE} = R_{BC} = R_{BD} = R > 1$, and let R_{BE} be the (noncausal) *C*-specific *B-E* odds ratio. Because *B* and *E* are marginally independent, the marginal *B-E* odds ratio is 1; hence $1/R_{BE}$ has a maximum of $\text{Bias}_{\max}(R_{BC}, R_{CE}) = (R + 1)^2/4R$, and so R_{BE} has a minimum of

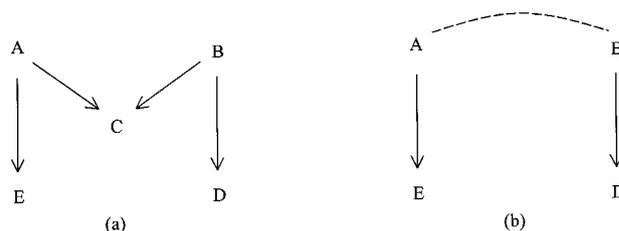


FIGURE 2. The relations of a nonconfounding covariate *C* to *E* and *D* are confounded by independent covariates *A* and *B*, respectively. (a), Relations before stratification on *C* (no confounding of the *E-D* association); (b), relations after stratification on *C* (*E-D* association now confounded by *B*).

$\min(R_{BE}) = 4R/(R + 1)^2 < 1$. Because *E* and *D* are marginally independent, the marginal *E-D* odds ratio equals 1 (which is the true *E-D* effect); hence the *C*-specific *E-D* odds ratio is purely bias attributable to ignoring *B* after stratifying on *C*, and a lower bound for this odds ratio is $\text{Bias}_{\max}(R_{BD}, \min[R_{BE}]) = \text{Bias}_{\max}(R, 4R/[R + 1]^2) = (3R + 1)^2/[R(R + 3)^2] < 1$ (see Appendix). This expression is near 1 unless the common effect *R* is large; eg, for *R* = 2, 4, 8 and 16 its inverse is only 1.02, 1.16, 1.55 and 2.41. The bound is thus much smaller than *R*, yet it is conservative: it assumes the factor prevalences could be such that the given bound is attained. With actual prevalences the bias will be smaller still; eg, with all the marginal prevalences set to 1/2, the actual inverse biases for *R* = 4, 8 and 16 are only 1.14, 1.43 and 1.93.

In general, the bias from controlling *C* will be closer to 1 (in ratio terms) than the responsible effects (R_{CE} , R_{BC} , R_{BD}), and, given those effects, it will approach 1 as either the confounder prevalence $P(B = 1)$ or the collider prevalence $P(C = 1)$ approaches 0 or 1. For estimating either the net or direct *E-D* effect under Figure 1a, large bias in the *C*-adjusted *E-D* association requires both a large *E-C* effect and substantial confounding of the *C-D* association. If Figure 1a is modified so that *C* affects *D* (by adding an arrow from *C* to *D*, so that *C* is an intermediate), the same conclusion applies to estimation of the direct *E-D* effect. For estimating the net *E-D* effect, however, the *C*-adjusted *E-D* association is further biased by the removal of the indirect ($E \rightarrow C \rightarrow D$) effect, although this further bias may add or subtract from the collider bias (which equals the confounding of the *E-D* association by *B* within *C* strata).

M Bias

The two collider-stratification biases described above can be avoided by not controlling variables affected by exposure or disease. Recent articles have emphasized that bias can also arise from stratifying on variables unaffected by *E* or *D*, including variables that temporally precede *E* and *D*.^{3,6,12} Figure 2a is a simple structure in

which this occurs, called an “M diagram” because of its shape when events are temporally ordered from top (earliest) to bottom (latest). Here, C is a collider on the “back-door” path from E to D passing through A, C and B (a *back-door path* from E to D is a path that begins with an arrow pointing to E; such paths are sources of confounding).^{3,12} This diagram is of special interest when A and B are not used for adjustment (perhaps because they were unmeasured, or because they were discarded by a variable-selection algorithm), but a variable C affected by A and B is used or affects subject selection. For example, if C = 1 represented study participation, A would be a factor that affected participation and exposure E but did not affect disease D except through E, and B would be an independent factor that affected participation and D but not E; in an environmental exposure study, candidates for A and B might include neighborhood and sex, respectively.

In Figure 2a, E and D share no cause or other source of covariation. Thus, the graph implies an absence of confounding of the E-D association, so that E and D (like A and B) will be marginally independent under this graph, properly reflecting the absence of an effect between them. Any stratification that moves the E-D association away from the null must be introducing bias for the true E effect on D, just as in Berkson’s hospital example.

One should expect to see a confounded E-D association within C strata. First, C is a collider on the A-C-B path; hence, A and B should be associated within some C strata, as in Figure 2b. Second, one should expect this C-specific A-B association to couple with the effects of A on E and B on D to produce confounding of the E-D association.¹² A more traditional way to explain the confounding within C strata views it as the result of a three-component process: if one does not control A, one should expect a spurious association of C and E attributable to confounding by A; if one does not control B, one should expect a spurious association of C and D attributable to confounding by B; finally, these spurious associations can make the C-specific E-D associations differ from the marginal E-D association (which is unconfounded). In the absence of data on A or B, the dual confounding (of C-E by A and of C-D by B) gives C the spurious appearance of being a classical confounder of the E-D association.

I will use the term “M bias” to refer to the bias in C-specific or C-adjusted E-D associations arising from an M pattern within the underlying causal structure (in which all or part of the C-E association arises from shared causes A of C and E, and all or part of the C-D association arises from shared causes B of C and D). Like earlier examples of collider-stratification bias, M-bias arises from adjustment for a variable C that numerically behaves like a classical confounder (in that the effect

estimate changes upon adjustment for C). Unlike those examples, however, the bias attributable to C-adjustment may not be apparent from the time order of the events, for C may be determined before E or D; hence, one may be led to adjust for C (and thus introduce bias) if one uses traditional confounder-selection criteria, even if one takes care to not adjust for variables affected by E or D. It thus seems imperative to ask how large M bias can be relative to the causal effects that produce it.

Quantifying the Bias

Using odds ratios to measure the effects in Figure 2a, we may apply formula 2 repeatedly to bound the bias produced by controlling C alone. For example, suppose the causal odds ratios corresponding to the arrows in Figure 2a are $R_{AE} = R_{AC} = R_{BC} = R_{BD} = R > 1$; then, because C is independent of E given A, the marginal C-E odds ratio is pure confounding by A and so has a maximum of $(R + 1)^2/4R$. For parallel reasons the maximum marginal C-D odds ratio is also $(R + 1)^2/4R$. Because E and D are marginally independent, the marginal E-D odds ratio equals 1 (which is the true E-D effect); hence the C-specific E-D odds ratio is purely bias from stratifying on C, and an approximate lower bound for this odds ratio is $1/\text{Bias}_{\max}([R + 1]^2/4R, [R + 1]^2/4R) = 16R(R + 1)^2/(R^2 + 6R + 1)^2$ (see Appendix). This expression is very close to 1 unless the common effect R is very large; eg, for R = 2, 4, 8 and 16 its inverse is only 1.003, 1.05, 1.23 and 1.68. The bound is thus much smaller than R, yet it is conservative; it assumes the factor prevalences could be such that the given bound is attained. With actual prevalences the bias will be smaller still; eg, setting all the marginal prevalences to 1/2, the actual inverse biases for R = 4, 8 and 16 are only 1.04, 1.16 and 1.39. More generally, under Figure 2a, large bias in the C-adjusted E-D association requires enormous confounding of both the C-E and C-D associations. Thus, M bias is much closer to 1 than the responsible effects, and, given those effects, will approach 1 as $P(A = 1)$, $P(B = 1)$ or $P(C = 1)$ approaches 0 or 1.

Discussion

A traditional belief about classical confounding is that “large values for relative risks are unlikely to be explained by some uncontrolled variable.”¹⁸ This belief might be rephrased as “it is unlikely some composite C of all uncontrolled variables is at once strongly associated with E and with D given E, and also neither too infrequent nor ubiquitous.” Only successful randomization of E makes the “unlikely” in these sentences an objective probability, because given enough subjects it will rarely create an allocation with a large degree of confounding.²⁸ Otherwise, “unlikely” is a subjective judgment,

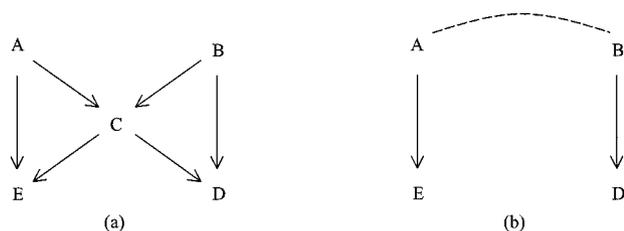


FIGURE 3. The relations of a confounder C to E and D are confounded by independent covariates A and B , respectively. (a), relations before stratification on C (E - D association confounded by multiple paths); (b), relations after stratification on C (E - D association now confounded by only one path).

and often more speculative than acknowledged (eg, failure to imagine an uncontrolled confounder is not evidence that there is little or no uncontrolled confounding, any more than imagining a possible uncontrolled confounder is evidence that there is uncontrolled confounding). Hence, algebraic evaluations of uncontrolled confounding will be most convincing when applied to specific unmeasured variables (eg, smoking) for which there is background information about their prevalence and associations with E and D .

Although an evaluation of collider bias can employ algebra developed for confounder analysis, it is somewhat less speculative insofar as the bias depends on the strength of relations of one known variable, the collider C , to the study variables E and D . For example, if control of C is suspected of introducing bias (as in Figures 1 and 2), one can estimate the size of that bias by the change in the E - D association upon C adjustment. Nonetheless, whether the change represents collider bias (as opposed to some other phenomenon) may remain quite speculative, because the bias involves variables (A and B in Figures 1 and 2) that may be unmeasured and may not even be known to exist.

Unlike classical confounding, the biases in Figure 1 and 2 depend on formulations of component biases; thus, for them to be large requires all their component biases be even larger, which may strain credibility in some contexts. This requirement seems especially severe for M bias, and is relevant to situations like that in the “bow-tie” diagram in Figure 3a, which has an M pattern embedded within it.¹² Here, C is a classical confounder, but is also a “back-door” collider, as in Figure 2a. If we cannot adjust for A or B , should we stratify on C ? C -stratification removes the classical confounding by A , B , and C , at the cost of introducing M bias (Figure 3b, which shows the confounding within C strata). The above results suggest that the M bias would often be slight, and, if the effects of C are even remotely comparable with those of A and B , the introduced M bias would be outweighed by the removed confounding. Nonetheless, consideration of variance as well as bias

might favor no adjustment for C , or adjustment that minimizes mean-squared error.²⁹ Further study of more general cases would help indicate whether M bias should be of practical concern in such adjustment decisions.

Acknowledgments

I thank Babette Brumback, Stephen Cole, Jay Kaufman, Richard MacLehose and Thomas Richardson for helpful comments.

References

- Pearl J. *Causality*. New York: Oxford, 2000.
- Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. 2nd ed. New York: Springer, 2001.
- Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37–48.
- Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001;12:313–320.
- Kaufman J, Kaufman S. Assessment of structured socioeconomic effects on health. *Epidemiology* 2001;12:157–167.
- Hernán M, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation. *Am J Epidemiol* 2002;155:176–184.
- Cole S, Hernán M. Fallibility in estimating direct effects. *Int J Epidemiol* 2002;31:163–165.
- Greenland S, Brumback BA. An overview of relations among causal modeling methods. *Int J Epidemiol* 2002;31:1030–1037.
- Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biomet Bull* 1946;2:47–53.
- Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott, 1998.
- Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999b;14:29–46.
- Pearl J. Causal diagrams for empirical research (with discussion). *Biometrika* 1995;82:669–710.
- Horwitz RI, Feinstein AR. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *N Engl J Med* 1978;299:1089–1094.
- Hutchison GB, Rothman KJ. Correcting a bias? *N Engl J Med* 1978;299:1129–1130.
- Greenland S, Neutra RR. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. *J Chron Dis* 1981;34:433–438.
- Cornfield J, Haenszel WH, Hammond EC, Lillienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *JNCI* 1959;22:173–203.
- Bross IDJ. Pertinency of an extraneous variable. *J Chron Dis* 1967;20:487–495.
- Schlesselman JJ. Assessing effects of confounding variables. *Am J Epidemiol* 1978;108:3–8.
- Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol I. The Analysis of Case-Control Data*. Lyon, France: IARC, 1980.
- Yanagawa T. Case-control studies: assessing the effect of a confounding factor. *Biometrika* 1984;71:191–194.
- Flanders WD, Khoury MJ. Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. *Epidemiology* 1990;1:239–246.
- Copeland KT, Checkoway H, Holbrook RH, McMichael AJ. Bias due to misclassification in the estimate of relative risk. *Am J Epidemiol* 1977;105:488–495.
- Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;9:361–367.

24. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;3:143–155.
25. Cox DR. *The Planning of Experiments*. New York: Wiley, 1958.
26. Weinberg CR. Toward a clearer definition of confounding. *Am J Epidemiol* 1993;137:1–8.
27. Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*. Gaithersburg, Md: Aspen, 2000;184–187.
28. Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421–429.
29. Greenland S. Reducing mean squared error in the analysis of stratified epidemiologic studies. *Biometrics* 1991;47:773–775.
30. Greenland S, Maldonado G. The interpretation of multiplicative model parameters as standardized parameters. *Stat Med* 1994;13:989–999.
31. Kitagawa EM. Components of a difference between two rates. *J Am Stat Assoc* 1955;50:1168–1194.
32. Gail MH, Wacholder S, Lubin JH. Indirect corrections for confounding under multiplicative and additive risk models. *Am J Ind Med* 1988;13:119–130.
33. Presented at: *WNAR/IMS Meeting*, Los Angeles, CA, June 2002.

Appendix

Bias Expressions

For binary C , D , E , let R_{ED} , R_{CD} and R_{CE} be the E - D , C - D , and C - E odds ratios given $C = 0$, $E = 0$, and $D = 0$, respectively, and let R_{CED} be the ratio of E - D odds ratios at $C = 1$ and $C = 0$. Then the ratio of the marginal E - D odds ratio to R_{ED} is

$$(R_{CED}R_{CD}R_{CE}r + 1)(r + 1)/([R_{CD}r + 1][R_{CE}r + 1]) \quad (A1)$$

where $r = p/(1 - p)$ (see Yanagawa,²⁰ formula. 2.2). formula 1 then follows when $R_{CED} = 1$ (the constant odds-ratio model) upon substituting $p/(1 - p)$ for r . Differentiation of formula 1 with respect to r shows that its maximum is formula 2, with maximum at $r = 1/G$; differentiation of formula 2 with respect to $Q = (R_{CD}/R_{CE})^{1/2}$ with G fixed shows that it is maximized when $Q = 1$ (ie, when $R_{CD} = R_{CE}$). The resulting maximum $(R + 1)^2/4R$ is a sharp bound on the bias given G .

Extensions to Figures 1 and 2

The bounds for the C -specific E - D odds ratios in Figures 1 and 2 follow by two and three applications of formula 2, respectively, under the given constant-effect models. As formula 1 shows, the marginal E - D odds ratio will equal the C -specific E - D odds ratio if C is independent of E given D or of D given E . The R_{CE} in formulas 1 and 2 is D -specific, but in the second application of

formula 2 to Figure 1, R_{BE} is substituted for R_{CE} , where R_{BE} is the C -specific B - E odds ratio (and is constant across C). To see that R_{BE} is also C - D -specific, as needed, denote the conditioning variables explicitly in the subscripts, eg, R_{BE} is now $R_{BE|C}$. From standard graphical rules (eg, Pearl¹²), Figures 1 and 2 imply that E and D are independent given B and C ($R_{ED|BC} = 1$), which in turn implies that the C - D -specific odds ratio $R_{BE|CD}$ equals $R_{BE|C}$ (substitute $R_{ED|BC}$ and $R_{BD|CE}$ into formula 1; the result equals 1 because $R_{ED|BC} = 1$). A parallel argument implies that $R_{AE|CD}$ equals $R_{AE|C}$ in Figure 2. These arguments are unnecessary if $D = 1$ is rare at all B - E - C levels in the expected sample, for then $R_{BE|C}$ will approximate $R_{BE|CD}$ regardless of the E - D association. The resulting bounds are not sharp, however, and become more conservative as R increases. In addition, the M -bias bound is obtained by substituting the unconditional maximum C - E and C - D odds ratio $(R + 1)^2/4R$ in place of the conditional odds ratios $R_{CE|D}$ and $R_{CD|E}$ in formula 2; nonetheless, in the examples the error from this approximation is slight (even when $R = 16$) because $R_{CE|D}$, $R_{CD|E}$ and $R_{ED|C}$ are not large.

Extensions to Other Models

To ease sequential application of formula 2 in Figures 1 and 2, I assumed first-order logistic models for effects, which imply absence of three-way log-linear interactions under those diagrams (eg, under a logistic dependence of C on D and E , the C -specific E - D odds ratio R_{ED} is constant across C). When moderate interactions are present (eg, if $R_{CED} \neq 1$ but is closer to 1 than each lower-order relative R_{ED} , R_{CD} and R_{CE}), the above results may be viewed as approximations by reinterpreting the conditional odds ratios (such as R_{ED} , R_{CD} and R_{CE} in formulas 1 and 2) as odds ratios standardized to the total-sample distribution. See the case-control results in Greenland and Maldonado³⁰ for examples of the approximation error under classical confounding. Yanagawa²⁰(theorem 1) generalized formula 1 directly to allow $R_{CED} \neq 1$, as in formula A1). Kitagawa³¹ gave a formula for confounding of risk differences, whereas Schlesselman¹⁸ and Flanders and Khoury²¹ gave formulas for risk ratios and polytomous C ; see also Gail *et al.*³² In these extensions it remains the case that the confounding is smaller than the smallest constituent effect, and much smaller than the average constituent effect.