© International Epidemiological Association 2002 Printed in Great Britain *International Journal of Epidemiology* 2002;**31**:163–165

Fallibility in estimating direct effects

Stephen R Cole[a] **and Miguel A Hernán**[b]

We use causal graphs and a partly hypothetical example from the Physicians' Health Study to explain why a common standard method for quantifying direct effects (i.e. stratifying on the intermediate variable) may be flawed. Estimating direct effects without bias requires that two assumptions hold, namely the absence of unmeasured confounding for (1) exposure and outcome, and (2) the intermediate variable and outcome. Recommendations include collecting and incorporating potential confounders for the causal effect of the mediator on the outcome, as well as the causal effect of the exposure on the outcome, and clearly stating the additional assumption that there is no unmeasured confounding for the causal effect of the mediator on the outcome.

Keywords Causal inference, direct effect, intermediate variable, risk ratio

Accepted 11 July 2001

Epidemiologists often try to identify specific pathways through which an exposure may affect an endpoint. One standard method used to examine a potential pathway from an exposure E through a measured intermediate variable M to an endpoint D is to stratify the measure of effect between E and D by M. The residual association after adjustment for the posited intermediate variable is then interpreted as the direct effect of exposure on the endpoint through pathways other than that pathway marked by the intermediate variable.

Following Robins[1] and Robins and Greenland,[2] we define the direct effect of E on D controlling for M as the causal effect of E on D when the M value of everybody in the population is physically set to a predetermined value. Therefore, there are as many direct effects as levels of M, and hence we generally refer in plural to the direct effects of E on D. Robins[1] and Robins and Greenland[2] demonstrated that the standard method described above does not generally provide an unbiased estimate of the direct effects of exposure. Using a hypothetical example, Poole and Kaufman[3] further demonstrated the inadequacy of this standard method. Below, we revisit this issue using directed acyclic graphs,[4–6] and an example that appends two hypothetical variables to data from the Physicians' Health Study.[7]

Example

The Physicians' Health Study (PHS) was a randomized double blind 2 × 2 factorial placebo-controlled trial of 22 071 US male physicians assigned to aspirin and β-carotene for the primary prevention of cardiovascular disease and cancer.[7] The aspirin component of PHS was stopped in January 1988 after an average of 5 years of follow-up, due primarily to the emergence of a 44% reduction in risk of myocardial infarction (MI) among those randomized to aspirin compared to placebo. Let $E = 1$ represent randomization to aspirin and $E = 0$ randomization to placebo. Likewise, let $D = 1$ represent diagnosis of myocardial infarction (MI) during the defined period of follow-up, with all subjects free of such a diagnosis at the start of follow-up. Let $M = 1$ represent a hypothetical indicator of high platelet aggregation measured without error after allocation to aspirin, but before possible diagnosis of MI. For simplicity, we will assume that M is a dichotomous variable throughout.

Our goal is to estimate the direct effect of aspirin on MI when everybody is treated in such a way that his or her platelet aggregation level is kept low, and similarly when platelet aggregation is kept high. Note that an unambiguous definition of direct effects requires detailed specification of the method by which platelet aggregation levels are physically set. For simplicity, we will assume throughout that such a method exists and is generally accepted, although this is an issue that would require discussion among subject-matter experts.

Because of randomization, the crude risk ratio RR_{ED} of 0.6 (Table 1) has, in principle, a causal interpretation indicating that aspirin exposure decreases the risk of MI. Imagine two investigators, one of who hypothesizes that the effect of exposure to aspirin causes the endpoint solely through reduced platelet

Table 1 Physicians' Health Study (N = 22 071)[7]

	Myocardial infarction	Total
Aspirin		
$E = 1$	139	11 037
$E = 0$	239	11 034

Unadjusted risk ratio = 0.6.

[a] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, MD, USA.

[b] Department of Epidemiology, Harvard School of Public Health, MA, USA.

Correspondence: Dr Stephen Cole, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St E-7139 Baltimore, MD 21205, USA. E-mail: scole@jhsph.edu

aggregation (i.e. no direct effects of E on D). We may depict this hypothesis using a directed acyclic graph,

$$E \rightarrow M \rightarrow D \quad (1).$$

If graph (1) represented the true state of nature, standard adjustment (stratification or regression) for M would leave no pathway from E to D (i.e. E would not be a cause of D within strata of M) and therefore the stratum-specific risk ratios $RR_{ED|M=m}$ (m = 0, 1) and the Mantel-Haenszel summary risk ratio $RR_{ED|M}$ would be 1. The second investigator agrees that aspirin use directly effects platelet aggregation but hypothesizes that the effect of aspirin on risk of MI goes through some other undetermined pathway not mediated by platelet aggregation (M), thereby proposing an alternative graph,

$$E \rightarrow M \quad D \quad (2).$$

Under graph (2), adjustment for M would not alter the crude risk ratio RR_{ED}. Of course, one could hypothesize more complex alternate worlds where the effect of aspirin is both mediated through platelet aggregation and has a residual direct effect (possibly mediated through its anti-inflammatory properties). The partly hypothetical data from Table 2, with Mantel-Haenszel $RR_{ED|M} = 0.6$, apparently supports the second investigator's hypothesis. A naïve interpretation of this RR could lead to the conclusion that E exerts a direct protection on D after adjusting for M. We explain below why this interpretation is incorrect for this example.

The true causal graph, unknown to both investigators, is

$$U \quad E \rightarrow M \rightarrow D \quad (3),$$

where U = 1 represents a hypothetical but unknown genotype which simultaneously predisposes one to increased platelet aggregation and decreased risk of MI through another pathway. Note that graph (3) is consistent with the first investigator's hypothesis, but she did not consider the existence of U. Recall that the genotype U, or any other pre-randomization variable, is not associated with aspirin assignment due to randomization. A cursory glance at causal graph (3) may leave one with the impression that the $RR_{ED|M}$ should equal 1, not 0.6. However, it can be shown using the theory of directed acyclic graphs that stratifying on M will induce a spurious (non-causal) association between E and D because (1) E predicts M, and (2) the causal effect of M on D is confounded by U (refs[5,8,9] for general explanations of causal graphs geared towards epidemiologists).

An intuitive understanding may be gleaned from the following. First, among people with high platelet aggregation (M = 1), those randomized to aspirin will be more likely to harbour the predisposing genotype U (since the cause of their high platelet aggregation cannot be due to the absence of aspirin). Similarly, among people with low platelet aggregation (M = 0), those randomized to placebo will be more likely to lack the genotype U. Thus within both strata of M, there is a positive association between U and E ($RR_{UEU|M=m} > 1$ for m = 0, 1). Second, U is inversely associated with D ($RR_{UD|M=m} = 0.1$ for m = 0, 1). Given these two conditions, and an understanding of the directionality of confounding, an inverse association between E and D is generally expected after conditioning on M. The magnitude of this spurious association will depend upon the associations between the unmeasured genotype and platelet aggregation, and between the genotype and MI, which were relatively strong in this example ($RR_{UM} = 2.7$ and $RR_{UD} = 0.2$). As our example shows, this spurious association may appear even if E has no direct effect on D. We consider confounding to be present for the causal effect of an exposure A on an outcome B when there exists a common cause of A and B.[9] In summary, we would expect a spurious association between E and D whenever (a) there is confounding for the effect of M on D (e.g. due to their common cause U), and (b) E predicts M. Note that M does not have to be on a causal pathway between E and D for the spurious association to occur.

One way to avoid this spurious association is to stratify on U in addition to M. Since this is a partly hypothetical example, we can stratify on both U and M and obtain a Mantel-Haenszel $RR_{ED|UM} = 1.0$ (Table 3). Under our causal graph (3), $RR_{ED|UM}$ has a simple causal interpretation as the direct effect of E on D within levels of both U and M. If the direct effects varied across the four strata defined by U and M, one would likely present the four direct effects separately. Alternatively, one may wish to calculate an average causal direct effect (by pooling over levels of U and M).[6] In real world applications U would remain unobserved and therefore estimation of $RR_{ED|MU}$ would be impossible.

When U is difficult or impossible to measure, one could remove the spurious association by stratifying on any measured variable C that is in the causal pathway between U and M (or between U and D), for example

$$U \rightarrow C \quad E \rightarrow M \rightarrow D \quad (4).$$

Table 2 Physicians' Health Study, stratified by hypothetical level of post-randomization platelet aggregation (N = 22 071)

	Myocardial infarction	Total
High platelet aggregation (M = 1)		
E = 1	38	3015
E = 0	187	8027
Low platelet aggregation (M = 0)		
E = 1	101	8022
E = 0	52	3007

Mantel-Haenszel adjusted risk ratio = 0.6.

Table 3 Physicians' Health Study, stratified by hypothetical level of post-randomization platelet aggregation and a pre-randomization unobserved genotype (N = 22 071)

	Myocardial infarction	Total
U = 0, high platelet aggregation (M = 1)		
E = 1	15	259
E = 0	148	2746
U = 0, low platelet aggregation (M = 0)		
E = 1	94	5264
E = 0	51	2745
U = 1, high platelet aggregation (M = 1)		
E = 1	23	2756
E = 0	39	5281
U = 1, low platelet aggregation (M = 0)		
E = 1	7	2758
E = 0	1	262

Mantel-Haenszel adjusted risk ratio = 1.0.

Note that stratifying on either *U* or *C* in data generated from graph (4) eliminates confounding of the effect of *M* on *D*.[5] But if *C* itself is affected by exposure, for example,

(5),

then no standard method (stratification or regression) can estimate the direct effects of interest without bias. Therefore, in cases where either *U* or *C* are affected by exposure *E*, a more general analytical framework for direct effects estimation is needed, such as Robins' non-parametric (g-formula[1,10]) or semi-parametric (direct effect nested structural models,[11] marginal structural models[12]) causal methods. As long as either *U* or *C* is measured and there is no model misspecification, these methods provide estimates that can be causally interpreted as the direct effect of *E* on *D* controlling for *M*.

Conclusion

Estimating direct effects requires the absence of unmeasured confounding for the effect of both the exposure and the intermediate variable on the outcome. If both of these conditions do not hold then no method is able to provide unbiased estimates the direct effects of exposure. Unfortunately, these assumptions of no unmeasured confounding are not testable given observed data. The directed acyclic graphs we presented are overly simplistic depictions of an often-complex world. However, understanding the assumptions necessary for correct estimation of direct effects in such simple worlds is prerequisite to any extension towards more realistic scenarios.

A recommendation to practising epidemiologists who wish to estimate direct effects of exposure is threefold. First, at the study planning stage collect potential confounders for the mediator-outcome association, as well as the exposure-outcome association. Second, at the study analysis stage, incorporate these additional variables (using standard or causal methods, as needed) in an attempt to correctly control for both paths of confounding. In particular, causal methods are generally required when one needs to adjust for variables affected by the exposure. Third, when communicating research results, clearly state and examine the additional assumption that there is no unmeasured confounding for the causal effect of the mediator on the outcome.

Acknowledgements

We thank Drs James Robins, Sander Greenland and Nancy Cook, as well as three anonymous reviewers for helpful suggestions and comments. Dr Cole was partly supported by the US National Institute of Allergy and Infectious Diseases by means of the data coordinating centers for the Multicenter AIDS Cohort (U01-Al-35043) and Women's Interagency HIV (U01-Al-42590) studies.

References

[1] Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect. *Mathematical Modeling* 1986;**7:**1393–512.

[2] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;**3:**143–55.

[3] Poole C, Kaufman JS. What does standard adjustment for downstream mediators tell us about social effect pathways? *Am J Epidemiol* 2000;**151:**S52.

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

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

[6] Pearl J. *Causality.* New York: Cambridge, 2000.

[7] Final report on the aspirin component of the ongoing Physicians' Health Study. Steering Committee of the Physicians' Health Study Research Group. *N Engl J Med* 1989;**321:**129–35.

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

[9] Hernán MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2001;In press.

[10] Robins J. The control of confounding by intermediate variables. *Stat Med* 1989;**8:**679–701.

[11] Robins JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: Glymour C, Cooper G (eds). *Computation, Causation, and Discovery*. Menlo Park, CA: AAAI Press/The MIT Press, 1999, pp.349–405.

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