Published in final edited form as:

Genet Epidemiol. 2010 May ; 34(4): 327–334. doi:10.1002/gepi.20484.

Case-only gene-environment interaction studies: when does association imply mechanistic interaction?

Tyler J. VanderWeele,

Harvard School of Public Health, Departments of Epidemiology and Biostatistics, 677 Huntington Avenue, Boston, MA 02115, Phone: 617-432-7855

Sonia Hernández-Díaz, and

Harvard School of Public Health, Department of Epidemiology, 677 Huntington Avenue, Boston, MA 02115, Phone: 617-432-3942

Miguel A. Hernán

Harvard School of Public Health, Department of Epidemiology, and Harvard-MIT, Division of Health Sciences and Technology, 677 Huntington Avenue, Boston, MA 02115, Phone: 617-432-0101

Tyler J. VanderWeele: tvanderw@hsph.harvard.edu; Sonia Hernández-Díaz: shernan@hsph.harvard.edu; Miguel A. Hernán: miguel_hernan@post.harvard.edu

Abstract

Case-only studies are often used to identify interactions between a genetic factor and an environmental factor under the assumption both factors are independent in the population. However, interpreting a statistical association between the genetic and the environmental factors among the cases as evidence of a mechanistic gene-environment interaction is not always warranted. Using a mechanistic approach based on the sufficient cause framework, we show association amongst cases can arise between the genetic and environmental factors when there is in fact no mechanistic gene-environment interaction. However, when it can be assumed the genetic and environmental factors themselves can never prevent the outcome, we show a positive association amongst cases implies a mechanistic gene-environment interaction. Without this assumption that the effects of the two factors are never preventive, a multiplicative interaction greater than 2 is needed to conclude the presence of a mechanistic interaction. We furthermore show these tests for mechanistic interaction can be extended to scenarios in which the genetic and environmental factors are negatively associated in the population rather than independent.

Keywords

Case-only; sufficient cause; multiplicative interaction; synergism

INTRODUCTION

The case-only study design is used to identify interactions between the effects of genetic and environmental factors on a disease using data only on diseased individuals (cases). This design estimates statistical interactions on the multiplicative scale under the assumption genetic and environmental factors are independent within the population [Piegorsch et al., 1994; Begg and Zhang, 1994; Khoury and Flanders, 1996; Yang et al., 1999; Schmidt and Schaid, 1999].

However, results from the case-only design can be misleading because of at least two potential problems. First, the design is quite sensitive to the assumption of independence of

genetic and environmental factors. That is, when genetic and environmental factors are associated, even if only mildly, the design may wrongly lead to the conclusion that a multiplicative statistical interaction exists [Albert et al., 2001]. Second, even when a multiplicative statistical interaction does exist, its presence does not imply the existence of biological mechanisms where genetic and the environmental factors interact to cause disease. That is, a statistical interaction is no guarantee of biologic or mechanistic interaction.

Here we describe conditions under which multiplicative statistical interactions can be appropriately interpreted as indications of true mechanistic interactions, and extend the applicability of the case-only design to settings where genetic and environmental factors are not independent but possibly negatively correlated. We first briefly review the case-only design.

METHODS

Consider a follow-up cohort study with dichotomous (i.e., present or absent) genetic exposure G , environmental exposure E , and disease D . Using the notation in Table 1, we define the following risk ratios of disease: $RR_g = [c/(c+d)]/[a/(a+b)]$ is the risk ratio for the genetic factor in the absence of the environmental factor compared with absence of both; $RR_e = [e/(e+f)]/[a/(a+b)]$ is the risk ratio for the environmental exposure in the absence of the genetic factor compared with absence of both; and $RR_{ge} = [g/(g+h)]/[a/(a+b)]$ is the risk ratio for joint presence of the genetic and the environmental exposures compared with absence of both.

We say there is a (statistical) multiplicative interaction for the risk ratio between the effects of the genetic and environmental exposure if $S_{RR} = RR_{ge}/(RR_g \times RR_e)$ is different than 1. A multiplicative interaction is positive if $S_{RR} > 1$ (the coefficient of a product term GE in a log-linear model would be positive), and negative if $S_{RR} < 1$ (the coefficient would be negative). If the genetic and environmental exposures are independent in the population, i.e., $[(g+h) \times (a+b)]/[(c+d) \times (e+f)] = 1$, then a case-only study design can be used to estimate multiplicative interaction because $S_{RR} = ag/ce$ [Yang et al., 1999; Schmidt and Schaid, 1999]. In other words, the magnitude of the multiplicative interaction, if any, can be assessed using an odds ratio relating genetic and environmental exposures amongst cases; and consequently association between genetic and environmental factors amongst cases implies multiplicative interaction.

Prior to these results for risk ratios, Piegorsch et al. [1994] derived a similar result for multiplicative interactions using odds ratios. If the data in Table 1 came from a case-control study with controls sampled from the source population (i.e., replace the label “Non-cases” by “Controls” in the second column of Table 1), one could similarly define the odds ratios: $OR_g = bc/ad$, $OR_e = be/af$ and $OR_{ge} = bg/ah$. We would say there is a positive multiplicative interaction for odds ratios between effects of the genetic and environmental exposure if $S_{OR} = OR_{ge}/(OR_g \times OR_e) > 1$ (the product term coefficient in a logistic model would be positive), and negative if $S_{OR} < 1$ (the coefficient would be negative). For diseases that are rare at all exposure levels, $S_{RR} = ag/ce$ from the case-only will be approximately equal to S_{OR} from the case-control study [Piegorsch et al., 1994]; see Appendix A for further discussion.

As the equations above show, case-only designs can be used to estimate statistical interactions when the genetic and environmental factors are unassociated. Unfortunately, even in the absence of association between these two factors, a multiplicative interaction can arise in a case-only study (i.e., $S_{RR} \neq 1$) even when there is no mechanistic interaction between effects of genetic and environmental factors. In this paper we illustrate this problem

through examples and clarify when a multiplicative interaction in a case-only study can be interpreted as indicating a true mechanistic interaction. When gene-environment independence assumption holds, the case-only interaction parameter will have a mechanistic interpretation whenever the original risk ratio interaction parameter has a mechanistic interpretation and we use this observation to clarify the relevance of the interaction parameters in the case-only design for drawing conclusions about mechanism.

RESULTS

We are interested in whether the genetic and environmental factors interact mechanistically in the sense of there being individuals for whom the outcome would occur if both the genetic and environmental factors were present but for whom the outcome would not occur if only one of these two factors were present. We relate this concept of a mechanistic interaction to synergism in the sufficient cause framework [Rothman, 1976; VanderWeele and Robins, 2007a, 2008; VanderWeele, 2009]. We first illustrate the concept of a mechanistic interaction with a couple of examples.

Negative Multiplicative Interaction but No Mechanistic Interaction

In this subsection we show a negative multiplicative interaction can arise in a case-only gene-environment interaction study even when there is no mechanistic interaction between genetic and environmental factors. Consider the following example. Suppose a particular disease D can arise only through one of three mechanisms, one involving the genetic factor, one involving the environmental factor and one involving neither the genetic nor the environmental factor. As is generally required in the case-only interaction study design, suppose the genetic exposure G and the environmental exposure E are independent in the population. The genetic factor alone may not be sufficient for developing disease D but suppose the genetic factor G in combination with some other factors, denoted by A_1 , will lead to disease. Similarly, suppose that the environmental factor E in combination with some other factors, denoted by A_2 , will lead to disease. Finally, let A_0 denote the set of factors necessary for the third mechanism for the disease, which does not require the genetic nor the environmental factor to operate. There are thus three mechanisms for the disease, one involving G and A_1 , one involving E and A_2 and one involving just A_0 . We represent these mechanisms diagrammatically as in Figure 1 (see VanderWeele and Robins [2007b, 2009] for discussion of how such diagrams can be formalized in terms of causal directed acyclic graphs [Pearl, 1995; Friedman et al., 2000; Hernán et al., 2004]). There is no mechanism requiring both G and E to operate; there are no individuals for whom the outcome would occur if both the genetic and environmental factors were present but for whom the outcome would not occur if only one of the two factors were present. For simplicity suppose G, E, A_0 , A_1 and A_2 are all independent in the population. Because A_0 , A_1 and A_2 are all independent of, and do not affect, G or E, the variables A_0 , A_1 and A_2 do not confound effects of G and E on D. Set the probabilities of G and E at 0.2 and 0.5 respectively, the probabilities of A_1 and A_2 both at 0.015 and the probability of A_0 at 0.005. Suppose the target population of interest has 10,000 individuals; the expected cross-classification of cases and non-cases by genetic and environmental exposure status is given in Table 2 (see Appendix B for derivations). We see the outcome is relatively rare under all exposure combinations. If we compute ag/ce , we obtain $20 \times 35 / (20 \times 80) = 0.44$, a negative multiplicative interaction. Note however there is no interaction between G and E in any of the mechanisms; there is no mechanism requiring both G and E to operate. We thus see a negative multiplicative interaction is possible even when there is no mechanistic interaction.

In a gene-environment interaction study of effects of tobacco use and the XRCC1 variant genotype, Albert et al. [2001] found evidence of a multiplicative interaction with $S_{OR} = 0.44$. We see from the example above such a negative multiplicative interaction of this magnitude

could arise even when there is no mechanistic interaction between genetic and environmental exposures. We note the estimate of $S_{OR}=0.44$ in Albert et al. [2001] was obtained from a case-control study; their estimate from the case-only study design is somewhat higher, but still less than 1, owing to violations in the independence assumption. However, the same point applies: it is possible to have a negative multiplicative interaction in the absence of any mechanistic interaction.

Positive Multiplicative Interaction but No Mechanistic Interaction

In this subsection we show, when one of the exposures can be preventive for some individuals and causative for others, even a positive multiplicative interaction can arise when there is in fact no mechanistic interaction between genetic and environmental factors. For example, in a study of stroke comparing high levels of alcohol consumption to low levels, high alcohol consumption might be harmful on average but there may be some individuals for whom high alcohol consumption prevents stroke (i.e. the absence of high alcohol consumption causes stroke). Suppose there are again three causal mechanisms for the outcome D, one involving G and some other factors A_1 , one involving E and some other factors A_2 , and one involving the absence of E, which we denote by E^c , and some other factors A_3 . We can represent these mechanisms diagrammatically as in Figure 2. Note that no mechanism requires both G and E; i.e. there is no mechanistic interaction between G and E. Suppose again the genetic factor G and the environmental factor E are independent in the population above. Let the probability of G and E be 0.4 and 0.5 respectively. Finally, suppose the distributions of A_1 , A_2 and A_3 are independent of G and E and the joint distribution of A_1 , A_2 and A_3 is as given in Table 3. It can then be shown that a cohort of 10,000 can be cross-classified into cases and non-cases by genetic and environmental exposure status as given in Table 4 (see Appendix B for derivations). If a case-only design were utilized for these data, we would obtain a multiplicative interaction of $ag/ce=12 \times 20 / (12 \times 15)=1.33$. We thus see if one of the exposures is causative for some individuals and preventive for others, we can also have a positive multiplicative interaction without a true mechanistic interaction.

Sufficient Cause Framework

The examples given above can be formalized and generalized using the sufficient cause framework [Rothman, 1976; VanderWeele and Robins, 2007a, 2008; VanderWeele, 2009]. If we consider a genetic factor G and an environmental factor E, then we might conceive of several causal mechanisms for the outcome as follows. A particular mechanism might require either G or its absence or neither, and it might require either E or its absence or neither. Each mechanism might also require some other factors which we denote by A_i where A_i variables essentially represent unmeasured or unknown factors necessary for a particular causal mechanism to operate. If we conceive of the mechanisms in this way we could have as many as nine different mechanisms, enumerated as follows:

$$A_0, A_1G, A_2E, A_3G^c, A_4E^c, A_5GE, A_6G^cE, A_7GE^c \text{ and } A_8G^cE^c$$

where G^c and E^c denote the absence of G and E respectively. Whenever all components of any particular mechanism are present, the outcome will occur. The mechanisms are sometimes referred to as sufficient causes because whenever all components of one mechanism are present, these components together suffice for the outcome. In many settings only some subset of these mechanisms will be present. In the two hypothetical examples considered above, there were only three mechanisms: A_0 , A_1G and A_2E in the first example and A_1G , A_2E and A_3E^c in the second example. If G can never be preventive for any individual, only causative, then none of the mechanisms involving G^c will be present and G

is said to have a monotonic effect on D; similarly if E can never be preventive for any individual, only causative, then none of the mechanisms involving E^c will be present and E is said to have a monotonic effect on D. We say this monotonicity assumption holds for G and E when neither the genetic factor G nor the environmental factor E are ever preventive.

A question of interest which may arise when thinking about causal mechanisms in this way is whether one can test for the existence of a mechanism requiring the presence of both G and E to operate. If there is such a mechanism then synergism exists between G and E [Rothman, 1976; VanderWeele and Robins, 2007a]. Another way to think about this question is asking whether there are individuals who would have the outcome D if both G and E were present but who would not if only one of the two were present. If there are such individuals, a sufficient cause interaction is said to be present [VanderWeele and Robins, 2007a, 2008]; in such situations there must be a causal mechanism which requires both G and E to operate and thus synergism must be present.

There are certain situations where one can use data to test for such sufficient cause interactions both where it could be assumed G and E had monotonic effects on D and where such monotonicity could not be assumed [VanderWeele and Robins, 2007a, 2008; VanderWeele, 2009]. In the following sub-sections we discuss how statistical interactions for risk ratios and odds ratios can be used to draw conclusions about mechanistic interactions in case-only studies.

Mechanistic Interactions Using the Case-Only Design

VanderWeele [2009] showed that, under monotonicity, $S_{RR} = RR_{ge}/(RR_g \times RR_e) > 1$ implies a sufficient cause interaction is present, i.e. there is a mechanism requiring both G and E to operate because there are individuals for whom the outcome would occur if both G and E were present but for whom the outcome would not occur if only one were present. Thus, under monotonicity and the standard assumption of independence between G and E in the population, if $ag/ce > 1$ so the odds ratio relating G and E among cases is greater than 1, then a mechanistic interaction must be present. These results require no confounding of effects of G and E on D, an issue which we return to in the discussion section.

In many situations the monotonicity assumption will be unreasonable. However, progress in testing for mechanistic interactions is still sometimes possible. In Appendix C, we show $S_{RR} > 2$ implies a mechanistic interaction if both G and E are causative on average, even if not necessarily for every individual. Thus in a case-only study if the genetic and environmental factors are on average causative over the population then a sufficiently large positive multiplicative interaction, $S_{RR} > 2$, suffices to infer the presence of a mechanistic interaction even if the individual-level monotonicity assumption does not hold.

In summary, under the standard assumption of independent genetic and environmental factors in a population, one can use case-only designs to test for mechanistic interaction. If the monotonicity assumption that the effects of G and E on D are never preventive for any individual holds, then a multiplicative interaction greater than 1 implies a true mechanistic interaction. If the individual-level monotonicity assumption does not hold but the factors are causative on average, then a multiplicative interaction greater than 2 implies a true mechanistic interaction.

Relaxing the Population Independence Assumption

The case-only study design is quite sensitive to violations of the assumption of independence between genetic and environmental factors [Albert et al., 2001; Gatto et al., 2004]. In the absence of independence, the case-only estimate ag/ce no longer estimates multiplicative interaction. However, tests for mechanistic interactions described in the

previous section can be applied in the absence of independence when the genetic and environmental factors are negatively associated in the population (see Appendix C for proof). In particular, if G and E are negatively associated and monotonicity holds, then $ag/ce > 1$ implies the presence of a mechanistic interaction. If individual monotonicity does not hold but effects of G and E are causative on average, then $ag/ce > 2$ implies the presence of a mechanistic interaction.

Khoury and Flanders [1996] discuss an example of genetic variation in alcohol and aldehyde dehydrogenases which are suspected risk factors for alcoholism and alcohol-related liver damage. They note individuals with genetic variants leading to delayed alcohol metabolism may have an increased flushing response after alcohol ingestion and thus may be less likely to seek alcohol. This would create a negative association between the genetic factor and the environmental factor. The standard assumptions of the case-only design would thus be violated. It would, however, still be possible to use tests described above to test for a mechanistic interaction since the association between the genetic and environmental factors is negative.

Examples from the Literature

Negative interaction $RR < 1$ —Wu et al. [2007] studied the effects of TNF-308 polymorphisms and smoking on the risk of childhood asthma. From data provided in their supplemental material, one can obtain a case-only multiplicative interaction estimate of $S_{RR} = 0.56$ (95% CI: 0.33–0.95). The authors hypothesize exposure to second hand smoking overwhelms the smaller effect of TNF-308 polymorphism on TNF production, and in the absence of the smoking exposure, the effect of polymorphism may be more apparent. It follows from the discussion above, this negative multiplicative interaction does not imply the presence of mechanistic interaction.

Positive interaction $1 < RR < 2$ —Milne et al. [2006] studied interaction between maternal folate supplementation and child's methylenetetrahydrofolate reductase (MTHFR) gene polymorphisms among infants with acute lymphoblastic leukaemia (ALL). Both folate supplementation and the MTHFR polymorphisms have been associated with a protective effect for ALL. The case-only OR for MTHFR C677T genotype and folate supplementation was 1.25 (95% CI: 0.31–5.10). In addition to statistical instability due to small sample sizes, this OR could be explained by the possibility the effect of this gene is causative for some individuals and preventive for others. As was made clear above, even if the case-only estimate of 1.25 had been statistically significant, conclusions could not be drawn about mechanistic interaction unless the monotonicity assumption held and effects of MTHFR gene and folate supplementation were in the same direction for all individuals.

Positive interaction $RR > 2$ —Using a case-only design, Bennett et al. [1999] studied the interaction between passive smoking and glutathione S-transferase M1 (GSTM1) on lung cancer risk among non-smokers. Because GST enzymes detoxify some carcinogenic components of tobacco smoke to make them excretable, it was hypothesized certain polymorphisms could increase susceptibility to tobacco smoke. The authors genotyped 106 lung cancer cases and estimated a case-only OR estimate of 2.6 (95% CI: 1.1–6.1) comparing GSTM1 with passive smoking exposure. Even the lower bound of the confidence interval suggests a mechanistic interaction assuming effects of passive smoking and glutathione S-transferase M1 (GSTM1) on lung cancer risk are monotonic for all individuals. The point estimate suggests a mechanistic interaction is present even without this monotonicity assumption.

DISCUSSION

In this paper we have discussed a number of results relating to tests for mechanistic interactions in case-only designs. We have shown the multiplicative interactions often estimated in case-only studies may not imply an interaction in a mechanistic sense. Negative multiplicative interactions can arise without any mechanistic interaction. Furthermore, unless it can be assumed neither the genetic nor the environmental factor are ever preventive for any individual, even positive multiplicative interactions can arise without a mechanistic interaction being present. Thus without further assumptions, estimates of multiplicative interactions in case-only studies may be of limited interest in drawing conclusions about interactions which are biologically meaningful.

We showed, however, if it can be assumed neither the genetic nor the environmental factor are ever preventive for any individual, then a positive multiplicative interaction does imply a mechanistic interaction. Without this monotonicity assumption, a multiplicative interaction greater than 2 would be needed to conclude the presence of a mechanistic interaction; but without monotonicity, one must further know that the main effects of the two exposures are both non-negative. These main effects cannot be estimated from a case-only design; so prior knowledge concerning the sign of the main effects is required and this is a limitation of our results. The results may still be of use for the purposes of hypothesis generation but eventually data from, and analogous method [VanderWeele and Robins, 2007a, 2008; VanderWeele, 2009] for, case-control and cohort studies will be needed to validate the hypotheses.

We further showed assuming independence between genetic and the environmental factors is not strictly necessary: one can draw conclusions about mechanistic interactions even when genetic and environmental factors are negatively associated in population. However, the settings in which it is known a priori these factors are negatively correlated in the population are quite limited. It will thus generally be important to collect data on controls and then perform more robust analysis before firm conclusions about mechanistic interpretation are drawn.

All of the remarks above would apply also to studies of gene-gene interaction. However, a further limitation of the current results is they apply only in settings in which genetic and environmental exposures are considered binary. Extensions to exposures that are categorical, ordinal or continuous should be considered in future work. Some progress has been made on mechanistic interactions for categorical and ordinal exposures [VanderWeele, 2010] and future work could extend this to case-only designs.

We have focused here on genetic and environmental factors which are on average causative in a population. If one of the factors is protective on average, similar results can be derived by recoding the relevant risk factor. If, for example, the environmental factor E is preventive (on average), one could define a new factor F denoting the absence of E; then the factor F will be causative on average and one could use the results described here to test for a mechanistic interaction between G and F. To apply the results described above to test for mechanistic interactions between G and F, we must assume G and F are either independent or negatively associated (i.e. G and E must be either independent or positively associated to test for mechanistic interaction between G and F).

Our results require the associations of G and E with D reflect causal effects of G and E on D i.e. the effects of G and E on D are not confounded. For example, this requirement would not be met if the association between G and D were due to linkage disequilibrium or population stratification. Thus, it may be necessary to adjust for confounding variables. If these covariates are all binary or categorical, it is possible to use the results above to test for

mechanistic interactions in each stratum of the covariates, as long as G and E are independent (or negatively associated) in all strata of the covariates where tests for interactions are conducted. If there are many covariates or if some of the covariates are continuous, statistical models, such as logistic regression, must be employed¹ and modeling assumptions concerning functional form must be made. When such modeling assumptions are made, additional caution must be exercised when trying to interpret interactions biologically or mechanistically because these conclusions are valid only if the functional form assumptions hold at least to a reasonable approximation [VanderWeele, 2009]. Reliance on modeling assumptions can be alleviated somewhat by using multiply robust semiparametric modeling approaches [Chatterjee and Carroll, 2005; Chen, 2007; Vansteelandt et al., 2008; Tchetgen Tchetgen and Robins, 2009].

In summary, the case-only design can be a useful exploratory method to evaluate the presence of gene-environment interactions and possibly identify interactions in a mechanistic sense. Our results should clarify the interpretation and extend the applicability of the case-only designs for identifying gene-environment interaction.

References

- Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol*. 2001; 154:687–693. [PubMed: 11590080]
- Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomark Prev*. 1994; 3:173–175.
- Bennett WP, Alavanja MCR, Blomeke B, Vähäkangas KH, Castrén K, Welsh JA, Bowman ED, Khan MA, Flieder DB, Harris CC. Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. *J Natl Cancer Inst*. 1999; 91:2009–2014. [PubMed: 10580025]
- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92:399–418.
- Chen YH. A semi-parametric odds ratio model for measuring association. *Biometrics*. 2007; 63:413–421. [PubMed: 17688494]
- Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comp Biol*. 2000; 7:601–620.
- Gatto NM, Campbell UB, Rundle AG, Ahsan H. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. *Int J Epidemiol*. 2004; 33:1014–1024. [PubMed: 15358745]
- Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiol*. 2004; 15:615–625.
- Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Epidemiol*. 1996; 144:207–213.
- Milne E, de Klerk NH, van Bockxmeer F, Kees UR, Thompson JR, Baker D, Armstrong BK. Is there a folate-related gene-environment interaction in the etiology of childhood acute lymphoblastic leukemia? *Int J Cancer*. 2006; 119:229–232. [PubMed: 16450391]
- Pearl J. Casual diagrams for empirical research (with discussion). *Biometrika*. 1995; 82:669–710.
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statist Med*. 1994; 13:153–162.
- Rothman KJ. Causes. *Am J Epidemiol*. 1976; 104:587–592. [PubMed: 998606]
- Schmidt S, Schaid DJ. Potential misinterpretation of the case-only study to assess gene-environment interaction. *Am J Epidemiol*. 1999; 150:878–885. [PubMed: 10522659]
- Tchetgen Tchetgen, EJ.; Robins, JM. Harvard School of Public Health Technical Report. 2009. The semi-parametric case-only estimator.
- VanderWeele TJ. Sufficient cause interactions and statistical interactions. *Epidemiol*. 2009; 20:6–13.
- VanderWeele TJ. Sufficient cause interactions for categorical and ordinal exposures. Revised for *Biometrika*. 2010

- VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component cause framework. *Epidemiol.* 2007a; 18:329–339.
- VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *Am J Epidemiol.* 2007b; 166:1096–1104. [PubMed: 17702973]
- VanderWeele TJ, Robins JM. Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika.* 2008; 95:49–61.
- VanderWeele TJ, Robins JM. Minimal sufficient causation and directed acyclic graphs. *Ann Statist.* 2009; 37:1437–1465.
- Vansteelandt S, VanderWeele TJ, Tchetgen Tchetgen EJ, Robins JM. Multiply robust inference for statistical interactions. *J Am Statist Assoc.* 2008; 103:1693–1704.
- Wu H, Romieu I, Sienra-Monge J-J, del Rio-Navarro BE, Anderson DM, Dunn EW, Steiner LL, Lara-Sanchez IC, London SJ. Parental smoking modifies the relation between genetic variation tumor necrosis factor-alpha (TNF) and childhood asthma. *Environ Health Persp.* 2007; 115:616–622.
- Yang Q, Khoury MJ, Fengzhu S, Flanders WD. Case-only design to measure gene-gene interaction. *Epidemiol.* 1999; 10:167–170.

APPENDIX A

Independence Assumption of the Case-Only Study

We note that the independence assumption for the odds ratio multiplicative interaction in a case-only study is sometimes articulated as an assumption of the independence of the genetic and environmental factor amongst the non-cases. This is in fact the assumption that is required mathematically for the odds ratio interaction derivation of Piegorsch et al. [1994] that $S_{OR}=ag/ce$. Because the genetic factor and the environmental factor are both causes of the outcome this assumption will almost never hold exactly [Hernán et al., 2004; Gatto et al., 2004; VanderWeele and Robins, 2007b]. Although the assumption will almost never hold exactly, the rare disease assumption in conjunction with the assumption of independent genetic and environmental factors in the population together imply that the genetic and environmental exposures are *approximately* independent amongst the non-cases. It is for this reason that Piegorsch et al. [1994] originally articulated the case-only study design assumptions as requiring both that the genetic and environmental factors are independent in the population and that the disease is rare in the population. The result relating the risk ratio multiplicative interaction to the odds ratio between the genetic and environmental exposures amongst the cases as derived in Yang et al. [1999] and Schmidt and Schaid [1999] does not require a rare disease assumption since the assumption of independence of the genetic and environmental factors in the population overall, rather than amongst the non-cases, is required in their derivation.

As noted above, if the genetic factor and the environmental factor are both causes of the outcome, then these two factors will almost never be exactly independent amongst the non-cases [Hernán et al., 2004; Gatto et al., 2004; VanderWeele and Robins, 2007b]. This is because conditioning on a consequence of two factors will often induce conditional correlation between the two factors even if the two factors are unconditionally independent [Pearl, 1995; Hernán et al., 2004]; this phenomenon is sometimes referred to as collider stratification [Hernán et al., 2004]. In some instances, the conditional correlation may only hold in one of two strata of the conditioning variable (e.g. only amongst the cases or only amongst the non-cases). The results of Yang et al. [1999] and Schmidt and Schaid [1999] show that if the genetic and environmental factors are independent in the population and if there is no interaction on the multiplicative scale in the effects of the two factors on the outcome then there will be no conditional correlation between the two factors amongst the cases; in such situations there will still be correlation between the two factors amongst the non-cases but again, if the rare disease assumption holds, then this correlation may be small.

APPENDIX B

Derivation of the Numbers of Cases and Non-Cases in Table 2

In Figure 1 and Table 2, we assumed three mechanisms for the outcome: one involving the genetic factor G in combination with some other factors denoted by A_1 , one involving the environmental factor E in combination with some other factors denoted by A_2 , and a final mechanism requiring some factors A_0 (but not requiring either the genetic or the environmental factor). It was assumed that G, E, A_0 , A_1 and A_2 were all independent in the population. It was furthermore assumed that the probabilities of G and E were 0.2 and 0.5 respectively, that the probabilities of A_1 and A_2 were both 0.015 and that the probability of A_0 was 0.005 and that there was a hypothetical target population of 10,000 individuals. The numbers in Table 2 are generated as follows.

$$P(D=1|G=0,E=0) = P(A_0=1) = .005$$

$$P(G=0,E=0) = P(G=0)*P(E=0) = (.8)(.5) = .40$$

$$\text{Total Number with } G=0 \text{ and } E=0 \text{ is given by } (10,000)*(.40) = 4,000$$

$$\text{The cases with } G=0 \text{ and } E=0 \text{ is given by } (4,000)*(.005) = 20$$

$$\text{Non-cases with } G=0 \text{ and } E=0 \text{ is given by } (4,000)*(.995) = 3980$$

$$P(D=1|G=1,E=0) = 1 - P(D=0|G=1,E=0) = 1 - P(A_0=0)*P(A_1=0) = 1 - (.995)*(.985) = .020$$

$$P(G=1,E=0) = P(G=1)*P(E=0) = (.2)(.5) = .10$$

$$\text{Total Number with } G=1 \text{ and } E=0 \text{ is given by } (10,000)*(.10) = 1,000$$

$$\text{The cases with } G=1 \text{ and } E=0 \text{ is given by } (1,000)*(.02) = 20$$

$$\text{Non-cases with } G=1 \text{ and } E=0 \text{ is given by } (1,000)*(.98) = 980$$

$$P(D=1|G=0,E=1) = 1 - P(D=0|G=0,E=1) = 1 - P(A_0=0)*P(A_2=0) = 1 - (.995)*(.985) = .020$$

$$P(G=0,E=1) = P(G=0)*P(E=1) = (.8)(.5) = .40$$

$$\text{Total Number with } G=0 \text{ and } E=1 \text{ is given by } (10,000)*(.40) = 4,000$$

$$\text{The cases with } G=0 \text{ and } E=1 \text{ is given by } (4,000)*(.02) = 80$$

$$\text{Non-cases with } G=0 \text{ and } E=1 \text{ is given by } (4,000)*(.98) = 3920$$

$$P(D=1|G=1,E=1) = 1 - P(D=0|G=1,E=1) = 1 - P(A_0=0)*P(A_1=0)*P(A_2=0) = 1 - (.995)*(.985)*(.985) = .035$$

$$P(G=1,E=1) = P(G=1)*P(E=1) = (.2)(.5) = .10$$

$$\text{Total Number with } G=1 \text{ and } E=1 \text{ is given by } (10,000)*(.10) = 1,000$$

$$\text{The cases with } G=1 \text{ and } E=1 \text{ is given by } (1,000)*(.035) = 35$$

$$\text{Non-cases with } G=1 \text{ and } E=1 \text{ is given by } (1,000)*(.965) = 965$$

Derivation of the Numbers of Cases and Non-Cases in Table 4

In Figure 2 and Table 4, we assumed three mechanisms for the outcome: one involving the genetic factor G in combination with some other factors denoted by A_1 , one involving the environmental factor E in combination with some other factors denoted by A_2 , and a final mechanism requiring some factors A_3 along with the absence of the genetic factor. It was

assumed that G and E were all independent in the population with probabilities 0.4 and 0.5. The distributions of A_1 , A_2 and A_3 were assumed to be independent of G and E but not of one another; it was assumed A_1 , A_2 and A_3 had the following joint distribution:

$$P(A_1=1, A_2=0, A_3=1) = .004$$

$$P(A_1=0, A_2=1, A_3=0) = .004$$

$$P(A_1=1, A_2=0, A_3=0) = .001$$

$$P(A_1=1, A_2=1, A_3=0) = .001$$

$$P(A_1=0, A_2=0, A_3=0) = .99$$

It was assumed that there was a hypothetical target population of 10,000 individuals. The numbers in Table 4 are generated as follows. From the joint distribution of A_1 , A_2 and A_3 we can calculate the following probabilities which we will need below: $P(A_3=1) = .004$, $P(A_2=1) = .005$, $P(A_1=1 \text{ or } A_3=1) = .006$, $P(A_1=1 \text{ or } A_2=1) = .010$. We then have the following:

$$P(D=1|G=0, E=0) = P(A_3=1) = .004$$

$$P(G=0, E=0) = P(G=0) \cdot P(E=0) = (.6)(.5) = .30$$

$$\text{Total Number with } G=0 \text{ and } E=0 \text{ is given by } (10,000) \cdot (.30) = 3,000$$

$$\text{The cases with } G=0 \text{ and } E=0 \text{ is given by } (3,000) \cdot (.004) = 12$$

$$\text{Non-cases with } G=0 \text{ and } E=0 \text{ is given by } (3,000) \cdot (.996) = 2988$$

$$P(D=1|G=1, E=0) = P(A_1=1 \text{ or } A_3=1) = .006$$

$$P(G=1, E=0) = P(G=1) \cdot P(E=0) = (.4)(.5) = .20$$

$$\text{Total Number with } G=1 \text{ and } E=0 \text{ is given by } (10,000) \cdot (.20) = 2,000$$

$$\text{The cases with } G=1 \text{ and } E=0 \text{ is given by } (2,000) \cdot (.006) = 12$$

$$\text{Non-cases with } G=1 \text{ and } E=0 \text{ is given by } (2,000) \cdot (.994) = 1988$$

$$P(D=1|G=0, E=1) = P(A_2=1) = .005$$

$$P(G=0, E=1) = P(G=0) \cdot P(E=1) = (.6)(.5) = .30$$

$$\text{Total Number with } G=0 \text{ and } E=1 \text{ is given by } (10,000) \cdot (.30) = 3,000$$

$$\text{The cases with } G=0 \text{ and } E=1 \text{ is given by } (3,000) \cdot (.005) = 15$$

$$\text{Non-cases with } G=0 \text{ and } E=1 \text{ is given by } (3,000) \cdot (.995) = 2985$$

$$P(D=1|G=1, E=1) = P(A_1=1 \text{ or } A_2=1) = .010$$

$$P(G=1, E=1) = P(G=1) \cdot P(E=1) = (.4)(.5) = .20$$

$$\text{Total Number with } G=1 \text{ and } E=1 \text{ is given by } (10,000) \cdot (.20) = 2,000$$

$$\text{The cases with } G=1 \text{ and } E=1 \text{ is given by } (2,000) \cdot (.010) = 20$$

$$\text{Non-cases with } G=1 \text{ and } E=1 \text{ is given by } (2,000) \cdot (.990) = 1980$$

APPENDIX C

Tests for Mechanistic Interactions Without the Monotonicity Assumption

Let $S_{RR} = RR_{ge} / (RR_g \times RR_e)$. VanderWeele [2009] showed that if $\log(S_{RR}) > \log(2) - \log(RR_g)$ and if $\log(S_{RR}) > \log(2) - \log(RR_e)$ then a sufficient cause

interaction is present even when the monotonicity assumption does not hold. If it is known that the genetic and environmental factors are on average causative over the population (i.e. if the main effect risk ratios are not less than 1 so that $\log(RR_g) \geq 0$ and $\log(RR_e) \geq 0$) then the conditions needed to conclude the presence of a sufficient cause interaction, namely $\log(S_{RR}) > \log(2) - \log(RR_g)$ and $\log(S_{RR}) > \log(2) - \log(RR_e)$, will be satisfied if $\log(S_{RR}) > \log(2)$ i.e. if $S_{RR} > 2$. Thus in a case-only study if the genetic and environmental factors are on average causative over the population, then a sufficiently large positive multiplicative interaction, $S_{RR} > 2$, suffices to conclude the presence of a mechanistic interaction even if the individual-level assumption of no preventive effects does not hold. Alternatively, if both $RR_g > 2$ and $RR_e > 2$ then we would have that $\log(2) - \log(RR_g) < 0$ and $\log(2) - \log(RR_e) < 0$ and thus the conditions needed to conclude the presence of a sufficient cause or mechanistic interaction, namely $\log(S_{RR}) > \log(2) - \log(RR_g)$ and $\log(S_{RR}) > \log(2) - \log(RR_e)$, would be satisfied if $\log(S_{RR}) > 0$ i.e. if $S_{RR} > 1$. In a case-only study neither RR_g nor RR_e can be estimated [Piergorsch et al., 1994; Khoury and Flanders, 1996]. If however it were known a priori that both of the main effects risk ratios were greater than 2 then a positive multiplicative interaction suffices to conclude the presence of a mechanistic interaction. In summary, even without the monotonicity assumption of no preventive effects for either of the factors, one can conclude the presence of a mechanistic interaction if either it is known that both main effects are non-negative and $S_{RR} > 2$ or if it is known a priori that both of the main effects risk ratios are greater than 2 and $S_{RR} > 1$.

Relaxing the Independence Assumption

Here we show that tests for mechanistic interactions apply not only when the genetic and environmental exposures are independent in the population but also when the genetic and environmental exposures are negatively associated in the population. Let Ψ_{case} denote the odds ratio relating the genetic and environmental exposures amongst the cases and let Ψ_{total} denote the odds ratio relating the genetic and environmental exposures in the population. Schmidt and Schaid [1999] showed that $S_{RR} = \Psi_{case} / \Psi_{total}$. If the genetic and environmental exposures are independent in the population then Ψ_{total} will be 1 and thus the magnitude of the multiplicative interaction S_{RR} will be equal to Ψ_{case} . See Albert et al. [2001] for similar discussion concerning the multiplicative interaction for the odds ratio.

Suppose now that the genetic and environmental factors are negatively associated in the population then Ψ_{total} will be less than 1. This implies that $S_{RR} > \Psi_{case}$. From this it follows that we could apply the tests for mechanistic interactions described in the previous section for case-only studies even when the association between the genetic and environmental factors is negative (i.e. the factors do not have to be independent). If the association between the genetic and environmental factors is negative (or zero) and if the effects of G and E on D are monotonic then if it were found that $\Psi_{case} > 1$ then one could conclude that $S_{RR} > 1$ and thus one could conclude the presence of a mechanistic interaction between G and E. Similarly, if the monotonicity assumption is not reasonable one could still conclude the presence of a mechanistic interaction if $\Psi_{case} > 2$ and if both factors were causative on average (non-negative main effects), even if the association between the genetic and environmental factors is negative.

An alternative derivation for the result where G and E are negatively associated but when G and E have positive monotonic effects on D are monotonic can be obtained from the results of VanderWeele and Robins [2007b, 2009]. VanderWeele and Robins showed that if G and E had monotonic effects on D and if G and E were negatively associated (or independent) then if there were no sufficient cause which required both G and E to operate then the association between G and E amongst the cases must be non-positive i.e. $\Psi_{case} \leq 1$. From this it follows that if G and E have monotonic effects on D, and if G and E are negatively

associated (or independent) in the population and it is found that $\Psi_{\text{case}} > 1$ then one could conclude the presence of a mechanistic interaction.

Figure 1.
Mechanisms for the outcome D in which there is no mechanistic interaction between G and E but in which there may be a negative multiplicative interaction.

Figure 2.
Mechanisms for the outcome D in which there is no mechanistic interaction between G and E but in which there may be a positive multiplicative interaction.

Table 1

Gene-environment interaction in a cohort study

Genetic Exposure	Environmental Exposure	Cases D=1	Non-cases D=0
G=0	E=0	a	b
G=1	E=0	c	d
G=0	E=1	e	f
G=1	E=1	g	h

Table 2

Example of data when there is no mechanistic interaction but a negative multiplicative interaction

Genetic Exposure	Environmental Exposure	Cases	Non-cases	Total
G=0	E=0	20	3980	4000
G=1	E=0	20	980	1000
G=0	E=1	80	3920	4000
G=1	E=1	35	965	1000

Table 3

Joint distribution of background factors for example in which there is no mechanistic interaction but a positive multiplicative interaction

$P(A_1=1, A_2=0, A_3=1) = .004$
$P(A_1=0, A_2=1, A_3=0) = .004$
$P(A_1=1, A_2=0, A_3=0) = .001$
$P(A_1=1, A_2=1, A_3=0) = .001$
$P(A_1=0, A_2=0, A_3=0) = .99$

Table 4

Example of data when there is no mechanistic interaction but a positive multiplicative interaction

Genetic Exposure	Environmental Exposure	Cases	Non-cases	Total
G=0	E=0	12	2988	3000
G=1	E=0	12	1988	2000
G=0	E=1	15	2985	3000
G=1	E=1	20	1980	2000