Published in final edited form as:

Eur J Epidemiol. 2011 May ; 26(5): 343–346. doi:10.1007/s10654-011-9588-z.

Subtleties of explanatory language: what is meant by “mediation”?

Tyler J. VanderWeele

Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA, USA tvanderw@hsph.harvard.edu

A distinction is sometimes drawn between what, in the philosophical literature, is called “token causation” as contrasted with “type causation.” Token causation involves statements of the form “X caused Y (in this particular instance)”. Type causation involves statements of the form “X causes Y (in general)”. Epidemiologists are principally concerned with type causation, i.e. with statements of the form “X causes Y.” We ask, what, in general, are the determinants of disease within the population? For better or for worse, issues of token causation—What was the cause in this particular instance? Who was to blame for this specific incident?—are generally left to coroner’s reports and law courts. But token causation and type causation are not unrelated. Once we have established type causation (“X causes Y”) we know there must be instances of token causation (instances in which “X in fact caused Y”) even if we cannot identify the particular instances. And likewise statements of token causation (“X caused Y” in these particular cases) imply statements of type causation (“X causes Y” in general). Epidemiologists, although generally focused on type causation, do have within their methodological toolkit, concepts to help think more carefully about token causation. The sufficient cause framework [1] has for decades been used to help think about and conceptualize the actual causes of particular outcomes.

These distinctions are not simply theoretical. They are important in the interpretation of data analysis. They relate to concepts regularly employed in epidemiologic discourse. Perhaps most importantly for epidemiologic practice, they relate to attributable fractions. Over 20 years ago, Greenland and Robins [2] noted that the use of “attributable fractions” or of expressions of the form “the proportion of disease due to an exposure” were in fact ambiguous. The proportion of the disease “due to an exposure” might refer to the proportion that “would be eliminated if the exposure were eliminated” or it might refer to the proportion of disease that was “actually caused by the exposure.” The two notions are not equivalent because of the possibility of competing causes. Even if an exposure were the actual cause of disease (an instance of “token causation”), it may still be the case that if the exposure were eliminated, the disease would still occur because of some other cause. Greenland and Robins proposed using “excess fraction” for the “proportion that would be eliminated if the exposure were eliminated” and “etiologic fraction” for the “the proportion actually caused by the exposure.” Only the former, not the latter, could in general be identified by data. They argued that these distinctions ought to be kept in mind when estimating and interpreting “attributable fractions.”

Epidemiologists are often not content with stopping at general conclusions about overall effects, about the extent to which, “X causes Y”. Often the pathways by which “X causes Y” are also of interest. For much of modern epidemiology’s history a rather informal regression-based approach has been employed to assess such questions (cf. [3]) but more recently a more formal approach to these questions based on counterfactuals [4, 5] has been advocated and the literature in epidemiology on practical tools to estimate direct and indirect effects has begun to expand rapidly [6–13]. These methods and the counterfactual-based direct and indirect effects they purport to estimate are still, however, based essentially on notions of

“type causation” i.e. the questions they answer take the form, “Are these pathways present in general and to what extent?” rather than, “Did this pathway operate in a given instance?”

In this issue of the *European Journal of Epidemiology*, Suzuki et al. [14] have taken the discussion one level further. They have used the sufficient cause framework to combine notions of token causation with concepts of direct and indirect effects to more precisely clarify what we mean by “mediation.” The use of the sufficient cause framework to shed light on questions of mediation and direct and indirect effects had been considered in prior work [15, 16] but, by appealing to notions of token causation, and by considering what the actual causes of intermediates and outcomes might be, Suzuki et al. are able to draw further distinctions not considered in prior work. At stake is what is meant by “mediation.” And the distinctions they draw, although subtle, are, for mediation, roughly analogous to those drawn by Greenland and Robins for attributable fractions.

Direct and indirect effects have recently been formalized within the counterfactual framework [4, 5] and considering these here will be helpful in following Suzuki et al.'s argument. We will let X denote our exposure, M an intermediate and Y our outcome. We will let Y_x be the potential outcome for each individual if X were set to x and we will likewise let M_x be the potential value of M for each individual if X were set to x . Finally we let Y_{xm} denote the potential outcome Y if, possibly contrary to fact, X were set to x and M were set to m . Robins and Greenland [4] and Pearl [5] defined a controlled direct effect comparing $X = 1$ and $X = 0$ for some level of $M = m$ as $Y_{1m} - Y_{0m}$. They also defined a different sort of direct effect, called a natural direct effect, defined by $Y_{1M_0} - Y_{0M_0}$. This is a comparison of the outcome with versus without treatment in both scenarios setting the intermediate to the level it would have been without treatment. The average natural indirect effect is then defined as $Y_{1M_1} - Y_{1M_0}$ i.e. a comparison of the outcomes under treatment when setting the intermediate to the level it would have been with versus without treatment. These notions of natural direct and indirect effects are desirable in that we can decompose a total effect into a natural indirect effect and a natural direct effect, $Y_1 - Y_0 = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$, and this decomposition holds irrespective of the functional form or interactions which may be present.

For the natural indirect effect, $Y_{1M_1} - Y_{1M_0}$, to be non-zero it would have to be the case that M_1 and M_0 were different, i.e. that the exposure changed the mediator, and that this difference in the mediator changed the outcome when the exposure was present so that Y_{1M_1} would be different from Y_{1M_0} . This comparison, $Y_{1M_1} - Y_{1M_0}$, is sometimes referred to as a “total natural indirect effect” [4]; the use of “total” denotes the fact that when we examine whether the potential change in the mediator from M_0 to M_1 changes the outcome, we do so whilst fixing the exposure to be present i.e. we consider the contrast, $Y_{1M_1} - Y_{1M_0}$. We might alternatively consider whether a potential change in the mediator from M_0 to M_1 changes the outcome while fixing the exposure to be absent; this would be the contrast, $Y_{0M_1} - Y_{0M_0}$, and this contrast is sometimes referred to as the “pure natural indirect effect,” the “pure” indicating that we consider whether the potential change in the mediator changes the outcome when exposure is absent. How might these “indirect effects” relate to actual mediation?

Ordinarily, we would think that mediation would be present if one of the natural indirect effects is non-zero. However, Suzuki et al. [14] draw a further distinction between the presence of a mediating pathway and the operation of a mediating pathway that qualifies the sense in which the presence of an “indirect effect” constitutes mediation.

Consider the sufficient cause diagram of Suzuki et al. [14–16], given here again in Fig. 1. Let us suppose that it were the case that the average value of “pure natural indirect effect”,

$E[Y_{0M1} - Y_{0M0}]$, was non-zero. Suzuki et al. [14] note that within the sufficient cause framework this would imply that there must be a pathway from X to M to Y , namely the B_3M sufficient cause must be present. Suzuki et al. thus say in this instance that a mediating pathway is present. But does this mediating pathway actually operate in the population? Suzuki et al. argue that it may not. Suppose that the background causes A_2 , B_2 and B_3 are present and that all others are absent (i.e. $A_1 = B_1 = B_6 = 0$). Although the mediating pathway through the sufficient B_3M is present and will, for some individuals, be capable of causing the outcome when the exposure is present, it may be the case that whenever the exposure is present, a different sufficient cause, B_2X , operates and is the actual cause of the outcome, even though the mediating pathway would also be sufficient for the outcome if the pathway involving B_2X were blocked. It may thus be the case that mediating path is present and would be sufficient for the outcome but that it in fact never normally operates; that whenever it is set in motion, it is pre-empted by some other sufficient cause for the outcome, B_2X , so that the mediating pathway is not in fact the actual cause. Suzuki et al. thus distinguish the “presence of mediation” from the “operation of mediation.”

Can we ever draw conclusions about the operation of mediation then? Under the setting considered by Suzuki et al. in which there are no sufficient causes requiring the absence of X or the absence of M (sometimes referred to as a monotonicity assumption, i.e. as in Fig. 1), the answer is yes. Suzuki et al. show that if the average “total natural indirect effect”, $E[Y_{1M1} - Y_{1M0}]$, is non-zero then a mediating pathway must also actually operate. This is essentially because for individuals with the sufficient cause B_2X present, the “total natural indirect effect”, $Y_{1M1} - Y_{1M0}$ will be zero. If the average total natural indirect effect is non-zero then for some individuals B_2X must be absent and the sufficient cause that is actually responsible for the change in Y must either be B_3M or B_6XM , both of which are mediating paths. Thus if the “total natural indirect effect” is non-zero then the mediating path must also, at least in some contexts, operate; it must be the actual cause. It is the “total natural direct effect” not the “pure natural direct effect” that gives this conclusion.

In a sense Suzuki's et al.'s argument gives priority to the “total natural direct effect” over the “pure natural direct effect” when assessing mediation. It is the “total natural direct effect” that indicates the operation, not simply the presence, of mediating pathways. And presumably it is the actual, rather than potential, operation of pathways that is in view when questions of mediation are posed. Fortunately, this “total natural direct effect” was the one first considered by Pearl, and, as a result, many of the more recent methods [8–10, 13] for mediation analysis (although they can in principle be used for both total and pure effects), take the total natural indirect effect as the default. Non-zero “total natural indirect effects” using these methods would, when valid, thus imply the operation and not simply the presence of mediation.

The distinctions here are admittedly subtle. Epidemiology appears, with these distinctions, to be approaching the level of nuance sometimes manifest in philosophical disputes about language. One might legitimately question whether this level of nuance is important in the interpretation of epidemiologic data and in policy decision making. The distinctions, I believe, are not irrelevant. They shape the language we use. They draw attention to the limits of what we can conclude from analysis. And they will influence development of methods and tools.

There is feedback between the questions that are of scientific or policy interest and the methods that are used to answer them. Questions of pathways are of interest; methods for mediation emerge. There is reflection within the methodological community on the nature of these questions concerning pathways and mediation, and the methodology evolves. These methods are supplied to the general epidemiologic community and are used to answer

questions of scientific and policy interest. The user of the methodology is, by his or her very use of the methodology, influenced by the reflections and refinements that have taken place. Subtle distinctions may shape methodological development and the tools that are made available. But these methods and tools eventually become those used in practice. If methodological reflection on what is meant by mediation results in priority of the “total natural direct effect” over the “pure natural direct effect” and methods take the “total natural direct effect” as a default this will influence practicing epidemiologists whether or not they are aware of the underlying distinctions. The distinctions may only be relevant to the practicing epidemiologist indirectly—but they will exert influence on practice by shaping the methods that are available and routinely employed.

Part of the task of the methodological community is settling on the concepts and methods appropriate for answering particular questions. Different questions will require different concepts and different methodology. I have argued elsewhere [17], for example, that counterfactual-based notions of what is sometimes called “principal stratification” [18], although of some use in assessing certain questions of causal inference, are essentially useless for assessing mediation per se (i.e. an effect through the intermediate itself) and that natural indirect effects as described above need to be used for this purpose. The work of Suzuki et al. qualifies further what can be concluded, even from natural indirect effects, about mediation as so conceived.

Importantly, the paper of Suzuki et al. [14] makes the monotonicity assumption of no sufficient causes involving the absence of the exposure or the absence of the mediator. Further work could be done in considering how we might understand the actual operation of mediation, and how this relates to direct and indirect effects, when these monotonicity assumptions do not hold.

References

1. Rothman KJ. Causes. *Am J Epidemiol.* 1976; 104:587–92. [PubMed: 998606]
2. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol.* 1988; 128:1185–97. [PubMed: 3057878]
3. Kaufman JS. Decomposing that takes a lot of supposing. *Am J Epidemiol.* 2010; 172:1349–51. [PubMed: 21036956]
4. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology.* 1992; 3:143–55. [PubMed: 1576220]
5. Pearl, J. Proceedings of the seventeenth conference on uncertainty and artificial intelligence. Morgan Kaufmann; San Francisco: 2001. Direct and indirect effects.; p. 411-20.
6. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology.* 2009; 20:18–26. [PubMed: 19234398]
7. Vansteelandt S. Estimating direct effects in cohort and case-control studies. *Epidemiology.* 2009; 20:851–60. [PubMed: 19806060]
8. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and its interface (Special Issue on Mental Health and Social Behavioral Science).* 2009; 2:457–68.
9. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis with a dichotomous outcome. *Am J Epidemiol.* 2010; 172:1339–48. [PubMed: 21036955]
10. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology.* 2010; 21:540–51. [PubMed: 20479643]
11. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods.* 2010; 15:309–34. [PubMed: 20954780]
12. Lange T, Hansen JV. Direct and indirect effects in a survival context. *Epidemiology.* in press.
13. VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology.* in press.

14. Suzuki E, Yamamoto E, Tsuda T. Identification of operating mediation and mechanism in the sufficient-component cause framework. *Eur J Epidemiol.* in press.
15. Hafeman D. A sufficient cause based approach to the assessment of mediation. *Eur J Epidemiol.* 2008; 23:711–21. [PubMed: 18798000]
16. VanderWeele TJ. Mediation and mechanism. *Eur J Epidemiol.* 2009; 24:217–24. [PubMed: 19330454]
17. VanderWeele TJ. Principal stratification—uses and limitations. *Int J Biostat.* in press.
18. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics.* 2002; 58:21–9. [PubMed: 11890317]

Fig. 1. Causal diagram for exposure X, mediator M and outcome Y representing sufficient causes for the mediator (i.e. A_1, A_2X) and for the outcome (i.e. B_1, B_2X, B_3M, B_6XM)