*Title Page

Original article

Title: A three-way decomposition of a total effect into direct, indirect, and interactive effects

Author: Tyler J. VanderWeele
Departments of Epidemiology and Biostatistics, Harvard School of Public Health

Corresponding Author:
Tyler J. VanderWeele
Harvard School of Public Health
Departments of Epidemiology and Biostatistics
677 Huntington Avenue
Boston MA 02115
Phone: 617-432-7855
Fax: 617-432-1884
E-mail: tvanderw@hsph.harvard.edu

Abbreviated running head: A three-way decomposition

Word Count (abstract): 158
Word Count (text): 3933

Total number of pages: 25
Pages of text: 23
Pages of tables: 0
Pages of figures: 2

Financial Support: The research was supported by National Institutes of Health grants HD060696 and ES017876.

A three-way decomposition of a total effect into direct, indirect, and interactive effects

Abstract

Recent theory in causal inference has provided concepts for mediation analysis and effect decomposition which allows one to decompose a total effect into a direct and an indirect effect. Here it is shown that what is often taken as an indirect effect can in fact be further decomposed into a "pure" indirect effect and a mediated interactive effect, thus yielding a three-way decomposition of a total effect into a direct, an indirect and an interactive effect. This three-way decomposition applies to difference scales and also to additive ratio scales and additive hazard scales. Assumptions needed for the identification of each of these three effects are discussed and simple formulae are given for each of these three effects when regression models, allowing for interaction, are employed. The three-way decomposition is illustrated by examples from genetic and perinatal epidemiology and discussion is given to what is gained over the traditional two-way decomposition into simply a direct and an indirect effect.

Introduction

There has been considerable interest in methodology for mediation analysis and effect decomposition of a total effect into direct and indirect effects. The recent causal inference literature has allowed for such effect decomposition even in the presence of interactions and in non-linear models.[1–13] The counterfactual quantities used to define these direct and indirect effects accommodated interaction, even at the individual level.[1,2] However the presence of such interaction led to more than one way to decompose the total effect into a direct effect and indirect effect, depending precisely on how the interaction was accounted for.[1,14] In this paper, it is shown that a further decomposition is possible: one can decompose a total effect into a direct effect, an indirect effect and an interactive effect. This further decomposition makes clearer the role of interaction when questions of mediation and pathways are of interest.

The paper is structured as follows. We first review definitions for natural direct and indirect effects and discuss issues concerning accounting for interaction in these decompositions. We then consider a difference scale and give a new three-way decomposition at the individual counterfactual level of a total effect into direct effect, indirect effect, and interactive components. Following this, we show how a similar decomposition can be achieved for ratio scales. We then illustrate how this 3-way decomposition can be carried out using simple regression models and in the following section we revisit two mediation analysis data examples in which direct and indirect effect were estimated and we carry out the 3-way effect decomposition in these settings. We close with discussion of the implications of the results in this paper for our understanding of pathways and mediation.

Natural Direct and Indirect Effects

Let A denote the exposure of interest, Y the outcome, and M a potential mediator, and let C denote a set of baseline covariates. We let Y_a and M_a denote respectively the values of the outcome and mediator that would have been observed had the exposure A been set to level a; let Y_{am} denote the value of the outcome that would have been observed had A been set to level a, and M to m. Suppose we compare two levels of the exposure, a and a^*; for binary exposure we would have $a = 1$ and $a^* = 0$. The controlled direct effect, comparing exposure level $A = a$ to $A = a^*$ and fixing the mediator to level m is defined by $Y_{am} - Y_{a^*m}$ and captures the effect of exposure A on outcome Y, intervening to fix M to m; it may be different for different levels of m.[1,2] It may also be different for different individuals. The natural direct effect[1,2] is defined as $Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$ and differs from controlled direct effects in that the intermediate M is set to the level M_{a^*}, the level that it would have naturally been under some reference condition for the exposure, $A = a^*$. Similarly, the natural indirect effect, can be defined as $Y_{aM_a} - Y_{aM_{a^*}}$, which compares the effect of the mediator at levels M_a and M_{a^*} on the outcome when exposure is set to $A = a$. For the natural indirect effect to be non-zero, the exposure would have to change the mediator and that change in the

mediator would have to change the outcome; natural indirect effects thus capture formally our notion of mediation. Defined thus, for a binary exposure these three effects would be: $Y_{1m} - Y_{0m}$ for the controlled direct effect; $Y_{1M_0} - Y_{0M_0}$ for the natural direct effect; and $Y_{1M_1} - Y_{1M_0}$ for the natural indirect effect. Natural direct and indirect effects have the property that a total effect, $Y_1 - Y_0$, decomposes into a natural direct and indirect effect: $Y_1 - Y_0 = Y_{1M_1} - Y_{0M_0} = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$; the decomposition holds even when there are interactions and non-linearities.

Because the direct and indirect effects above are counterfactual quantities we in general will not be able to compute these for any individual in the population, but under certain assumptions, we might hope to be able to estimate them on average. The expected valued of three effects, conditional on the covariates $C = c$, are defined by: $E[Y_{1m} - Y_{0m}|c]$, $E[Y_{1M_0} - Y_{0M_0}|c]$, and $E[Y_{1M_1} - Y_{1M_0}|c]$ respectively. Under certain no-confounding assumptions, these average controlled direct effect, natural direct effect and natural indirect effect, conditional on the covariates, are identified by the data. For causal diagrams interpreted as non-parametric structural equation models[15], the following four assumptions suffice to identify natural direct and indirect effects from data[2]: (i) the effect the exposure A on the outcome Y is unconfounded conditional on C; (ii) the effect the mediator M on the outcome Y is unconfounded conditional on C; (iii) the effect the exposure A on the mediator M is unconfounded conditional on C; and (iv) there is no effect of the exposure that itself confounds the mediator-outcome relationship. If we let $X \coprod Y|Z$ denote that X is independent of Y conditional on Z then these four assumptions stated formally in terms of counterfactual independence, are: (i) $Y_{am} \coprod A|C$, (ii) $Y_{am} \coprod M|\{A, C\}$, (iii) $M_a \coprod A|C$, (iv) $Y_{am} \coprod M_{a^*}|C$. Average controlled direct effects conditional on C, are identified by assumptions (i) and (ii) alone; natural direct and indirect effects are identified by assumptions (i)-(iv). Some additional technical conditions referred to as consistency and composition are also needed to relate the observed data to counterfactual quantities. The consistency assumption in this context is that when $A = a$, the counterfactual outcomes Y_a and M_a are, respectively, equal

to the observed outcomes Y and M, and that when $A = a$ and $M = m$, the counterfactual outcome Y_{am} is equal to Y. The composition assumption is that $Y_a = Y_{aM_a}$. Further discussion of these assumptions is given elsewhere.[4,16] Note that assumption (iv) requires that there is no effect of the exposure that itself confounds the mediator-outcome relationship. This would hold in Figure 1 but would be violated in Figure 2.

Avin et al.[17] have shown that natural direct and indirect effects are not identified from data in Figure 2 or whenever there is a variable (such as L) that is affected by exposure that in turn confounds the mediator-outcome relationship, irrespective of whether data is available on this exposure-induced confounder or not.

The natural direct and indirect effects defined above are referred to by Robins and Greenland[1] as "pure direct effects" and "total indirect effects" respectively. Robins and Greenland use the terminology "pure" and "total" because there are different ways of decomposing an overall effect into direct and indirect effects component. Above, we decomposed the overall or total effect as follows: $Y_1 - Y_0 = Y_{1M_1} - Y_{0M_0} = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$. For the natural direct effect, $Y_{1M_0} - Y_{0M_0}$, we compared average outcomes under exposure versus no exposure, in both cases setting the mediator to what it would have been in the absence of exposure. We might instead compare exposure to no exposure, now in both cases setting the mediator to what it would have been in the presence of exposure. This would be the counterfactual contrast $Y_{1M_1} - Y_{0M_1}$. Likewise in the decomposition above, for the natural indirect effect, $Y_{1M_1} - Y_{1M_0}$, we compared average outcome when exposure is set to present and the mediator is set to the level it would have been with versus without exposure. We might instead compare average outcome when exposure is set to absent and the mediator is set to the level it would have been with versus without exposure. This would be the counterfactual contrast $Y_{0M_1} - Y_{0M_0}$. Robins and Greenland refer to $Y_{1M_1} - Y_{0M_1}$ as the "total direct effect" and $Y_{1M_1} - Y_{1M_0}$ as the "pure indirect effect", in contrast to the "pure direct effect" and "total indirect effect" considered above. We also then have an alternative effect decomposition of an overall effect: $Y_1 - Y_0 = (Y_{1M_1} - Y_{0M_1}) + (Y_{0M_1} - Y_{0M_0})$. We

can thus decompose an overall, $Y_1 - Y_0$, either into a total indirect effect and a pure direct effect, $(Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$, or into a total direct effect and a pure indirect effect, $(Y_{1M_1} - Y_{0M_1}) + (Y_{0M_1} - Y_{0M_0})$.

The "pure" and "total" terminology used by Robins and Greenland essentially arises from different ways of accounting for an interaction. When we decompose an overall or total effect into a pure direct effect and a total indirect effect, the indirect effect "picks up" the interaction; the "pure" in "pure direct effect" effectively indicates that the direct effect does not pick up the interaction. When we decompose an overall effect into a total direct effect and a pure indirect effect, the direct effect picks up the interaction; the "pure" in "pure indirect effect" effectively indicates that the indirect effect does not pick up the interaction. We thus have two different decompositions depending on how we account for the interaction. Traditionally the decomposition has been into the pure direct effect and the total indirect effect. This was arguably in part because of historical reasons as this decomposition was the one initially suggested by Pearl; however, under certain 'monotonicity' assumptions, the total indirect effect, in contrast to the pure indirect effect, would also give more evidence for the actual operation, rather than just the presence, of mediating mechanisms.[18,19] Nonetheless, two decompositions remain and there is some level of arbitrariness or ambiguity in choosing between them. This ambiguity of the choice between the two essentially arises from different ways of accounting for interaction. In the next section, we show that this ambiguity can be eliminated by a three-way decomposition of a total effect into three components: (i) a pure direct effect, (ii) pure indirect effect, and (iii) an interactive effect.

A Three-way Decomposition of a Total Effect into Direct, Indirect and Interactive Effects

For simplicity we will consider the setting of a binary exposure and binary mediator. A more general decomposition for categorical or continuous exposure and mediator is given in the appendix. For binary exposure A, binary mediator M and outcome Y, we show in the

Appendix that we have the following decomposition:

$$Y_1 - Y_0 = (Y_{1M_0} - Y_{0M_0}) + (Y_{0M_1} - Y_{0M_0}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0). \quad (1)$$

The first term in this decomposition is the pure direct effect considered in the previous section. The second term in this decomposition is the pure indirect effect considered in the previous section. The third term in this decomposition, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$, is the product of an interaction between the exposure and the mediator on the outcome, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$, and the effect of the exposure on the mediator, $(M_1 - M_0)$. This interactive effect will be non-zero if and only if it is both the case that the exposure has some effect on the mediator, $(M_1 - M_0) \neq 0$, and if the interaction contrast, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$, is non-zero. We might thus refer to this interactive effect as a "mediated interactive effect". The contrast $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$ is a counterfactual measure of additive interaction. It is considered in more detail elsewhere.[20–23] It can be rewritten as $(Y_{11} - Y_{00}) - \{(Y_{10} - Y_{00}) + (Y_{01} - Y_{00})\}$. It will be non-zero for an individual if the effect on the outcome of setting both the exposure and the mediator to present differs from the sum of the effects of having only one of the exposure or the mediator to present. In the appendix, it is shown that this mediated interactive effect is equal to the difference between the total indirect effect and the pure indirect effect, $(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})$; the mediated interactive effect is also equal to the difference between the total direct effect and the pure direct effect. The three-way decomposition above, and the mediated interactive effect essentially resolves the ambiguity above concerning the choice between decomposition into a pure direct and total indirect effect, or a total direct and pure indirect effect. The ambiguity was created by different ways of accounting for interaction. Instead of specifically assigning such interaction to either the direct effect or the indirect effect we can simply account for it separately.

The decomposition above in (1) applies at the individual counterfactual level. We have considered average direct and indirect effects conditional on the covariates above. The av-

erage interactive mediation effect conditional on covariates $C = c$, could likewise be defined as: $E[(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)|c]$. Under the assumptions (i)-(iv) above (specifically, $Y_{am} \coprod M_{a^*}|C$), we can give a somewhat similar decomposition for the average effect conditional on C:

$$E[Y_1 - Y_0|c] = E[Y_{1M_0} - Y_{0M_0}|c] + E[Y_{0M_1} - Y_{0M_0}|c] + E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]E[M_1 - M_0|c]. \quad (2)$$

The first expression in the decomposition is the average pure direct effect conditional on the covariates C. The second term in this decomposition is the average pure indirect effect considered conditional on the covariates C. The third term in the decomposition is the product of the average causal interaction conditional on covariates C, $E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]$, and the average effect of the exposure on the mediator conditional on covariates C, $E[M_1 - M_0|c]$. As shown in the Appendix, what assumption (iv) essentially gives us here is that the average "mediated interactive effect" is simply equal to the product of the average additive interaction and the average effect of the exposure on the mediator i.e. $E[(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)|c] = E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]E[M_1 - M_0|c]$.

In the Appendix it is also shown that under assumptions (i)-(iv) the average pure direct effect, pure indirect effect, and mediated interactive effect, conditional on covariates $C = c$ are identified from data by the following empirical expressions:

$$\begin{aligned}
E[Y_{1M_0} - Y_{0M_0}|c] &= \sum\nolimits_m \{E[Y|A=1,m,c] - E[Y|A=0,m,c]\}P(m|A=0,c) \\
E[Y_{0M_1} - Y_{0M_0}|c] &= \sum\nolimits_m E[Y|A=0,m,c]\{P(m|A=1,c) - P(m|A=0,c)\}. \\
E[(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)|c] &= \{E[Y|A=1,M=1,c] - E[Y|A=1,M=0,c] \\
&\quad - E[Y|A=0,M=1,c] + E[Y|A=0,M=0,c]\} \\
&\quad \times \{E[M|A=1,c] - E[M|A=0,c]\}. \qquad (3)
\end{aligned}$$

In a subsequent section we will illustrate the estimation of these three effects using regression

models.

It was noted above that when using a two-way decomposition of a total effect into a direct and an indirect effect there was ambiguity in how this was done and in the manner in which interaction was accounted for. The total effect could be decomposed into the sum of a total indirect effect and a pure direct effect or into a pure indirect effect and a total direct effect. The three-way decomposition arguably lends support to the approach of using the total indirect effect and the pure direct effect. This is because the total indirect effect is itself composed of the pure indirect effect and a mediated interaction. If the indirect effect that we use in a two-way decomposition of a total effect into direct and indirect effects is to capture the entirety of the effect that is in some sense mediated then it arguably ought to include the mediated interaction as well. Fortunately, it is the decomposition of a total effect into a total indirect effect and a pure direct effect that has most often been employed in practice and in software and, as noted above, there are other theoretical arguments for sometimes preferring this particular decomposition.[18,19] However, once again, with the 3-way decomposition, one need not decide between alternative two-way decompositions and alternative approaches to accounting for interaction. The mediated interactive effect can be left as its own component in the decomposition.

A Three-way Decomposition on the Ratio Scale

Thus far we have been considering the definition of these direct, indirect and interaction effects on a difference scale. Often in epidemiology risk ratios or odds ratios are used for convenience, or ease of interpretation, or to account for study design. Direct and indirect effect have also been considered on risk ratio and odds ratio scales.[5,12] For example we could define the conditional total effect risk ratio by $RR_c^{TE} = E[Y_1|c]/E[Y_0|c]$. We could likewise define the pure direct effect risk ratio by $RR_c^{DE} = E[Y_{1M_0}|c]/E[Y_{0M_0}|c]$ and the pure indirect effect risk ratio by $RR_c^{IE} = E[Y_{0M_1}|c]/E[Y_{0M_0}|c]$. As shown in the eAppendix we then have

the following decomposition for the excess relative risks:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + \left(\frac{E[Y_{1M_1}|c]}{E[Y_{0M_0}|c]} - \frac{E[Y_{1M_0}|c]}{E[Y_{0M_0}|c]} - \frac{E[Y_{0M_1}|c]}{E[Y_{0M_0}|c]} + 1 \right) \quad (4)$$

On the left hand side of this equation, the term $(RR_c^{TE} - 1)$ is the excess relative risk for the total effect. On the right hand side of the equation, we have a 3-way decomposition. The first term in this decomposition is the excess relative risk for the pure direct effect, the second term is the excess relative risk for the pure indirect effect, and the final term could be interpreted as a measure of mediated excess relative risk due to interaction. We will refer to this quantity as $RERI_{mediated}$. When using a ratio scale, epidemiologists will sometimes use a quantity called the "relative excess risk due to interaction"[24] or the "interaction contrast ratio"[20]. The causal relative excess risk due to interaction if M were binary would be defined as:

$$RERI_{causal} = \frac{E[Y_{11}|c]}{E[Y_{00}|c]} - \frac{E[Y_{10}|c]}{E[Y_{00}|c]} - \frac{E[Y_{01}|c]}{E[Y_{00}|c]} + 1. \quad (5)$$

It assesses whether there is additive interaction but does so using ratios. The mediated relative excess risk due to interaction in (4) is analogous to the regular causal relative excess risk due to interaction in (5) but with replacing $m = 1$ and $m = 0$ in (5) with M_1 and M_0, respectively, in (4).

In fact under assumption (iv), $RERI_{mediated}$ is equal to $RERI_{causal}$ times a scaling factor:

$$RERI_{mediated} = \phi \times RERI_{causal}$$

where ϕ is given by $\phi = E[M_1 - M_0|c]E[Y_{00}|c]/E[Y_{0M_0}|c]$.

In any case, analogous to the decomposition for the total effect defined on a difference scale, we can decompose the excess relative risk for a total effect into the sum of the excess relative risk for the pure direct effect, the excess relative risk for the pure indirect effect, and

the mediated relative excess risk due to interaction:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}. \quad (6)$$

These quantities are likewise all identified under assumptions (i)-(iv); estimation of direct and indirect effect risk ratios is described elsewhere.[5,12] Similar decompositions would hold also for an odds ratio scale. In the eAppendix, we describe a simple estimation approach for the ratio scale using regressions that allow for interaction. Likewise, in the eAppendix we discuss how similar three-way decompositions hold for direct and indirect effects for hazard ratios[9–11], allowing one to decompose the excess hazard ratio for a total effect into the sum of an excess hazard ratio for the direct effect, an excess hazard ratio for the indirect effect, and the hazard ratio equivalent of the mediated relative excess risk due to interaction.

One final point is perhaps worth noting. Using odds ratios, which will approximate risk ratios when the outcome is rare, VanderWeele and Vansteelandt[5] used a decomposition of a total effect risk ratio (odds ratio) into a product of a pure direct effect risk ratio and a total indirect effect risk ratio where the total indirect risk ratio would be defined as $RR_c^{TIE} = E[Y_{1M_1}|c]/E[Y_{1M_0}|c]$ so that $RR_c^{TE} = RR_c^{TIE} \times RR_c^{DE}$. VanderWeele and Vansteelandt proposed as a measure of the proportion mediated on the risk difference scale the measure $\frac{RR_c^{DE}(RR_c^{TIE}-1)}{(RR_c^{TE}-1)}$. It is shown in the eAppendix that the numerator in this quantity, $RR_c^{DE}(RR_c^{TIE} - 1)$, is in fact equal to $(RR_c^{IE} - 1) + RERI_{mediated}$ i.e. to the sum of the excess relative risk for the pure indirect effect plus the mediated relative excess risk due to interaction. These are the latter two terms in the decomposition in (6).

Direct, Indirect and Interactive Effects with Regression

Suppose that assumptions (i)-(iv) hold, that Y and M are continuous and the following

regression models for Y and M are correctly specified:

$$\begin{aligned} E[Y|a,m,c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ E[M|a,c] &= \beta_0 + \beta_1 a + \beta_2' c. \end{aligned}$$

VanderWeele and Vansteelandt[4] derived expressions for natural direct and indirect effects from these two regressions. However, as discussed above we can further decompose such effects into a pure direct effect, a pure indirect effect and a mediated interactive effect. It is shown in the eAppendix that for exposure levels a and a^* the pure direct effect and pure indirect effect are given by:

$$\begin{aligned} E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] &= \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*) \\ E[Y_{aM_a} - Y_{aM_{a^*}}|c] &= (\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*) \end{aligned}$$

and the mediated interactive effect is given by

$$E[Y_{aM_a} - Y_{aM_{a^*}} - Y_{a^*M_a} + Y_{a^*M_{a^*}}|c] = \theta_3\beta_1(a - a^*)(a - a^*).$$

The sum of the pure indirect effect and the mediated interactive effect is equal to $(\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$, which is the total indirect effect derived by VanderWeele and Vansteelandt. If the exposure were binary the pure direct, pure indirect and mediated interactive effects would respectively simply be: $\{\theta_1 + \theta_3(\beta_0 + \beta_2' E[C])\}$, $\theta_2\beta_1$, and $\theta_3\beta_1$. Standard errors for estimators of these quantities could be derived using the delta method along the lines of VanderWeele and Vansteelandt[4] or by using bootstrapping. In the eAppendix we also derive similar expressions for a binary outcome for the pure direct effect risk ratio, the pure indirect effect ratio and the mediated relative excess risk due to interaction.

Illustrations

We will consider two data examples using methods from causal mediation analysis to decompose a total effect into natural direct and indirect effects. Here we will revisit these examples and give the 3-way decompositions. VanderWeele et al.[25] used lung cancer case-control data to examine the extent to which the effect of chromosome 15q25.1 rs8034191 C alleles on lung cancer risk was mediated by cigarettes smoked per day. rs8034191 C alleles had been found to be associated with both smoking[26,27] and lung cancer[28–30] but there had been debate as to whether the effects on lung cancer were direct or mediated by smoking. As the outcome, lung cancer, is rare, odds ratios approximate risk ratios. Using lung case-control data and logistic regression models, controlling for sex, age, education, restricting to Caucasians, and allowing for gene-by-smoking interaction, it was found that comparing 2 to 0 C alleles gave a pure direct effect odds ratio of 1.72 (95% CI: 1.34, 2.21), a total indirect effect odds ratio of 1.028 (95% CI: 0.99, 1.07), and a total effect odds ratio of $1.72 \times 1.028 = 1.77$ (95% CI: 1.38, 2.26), with proportion mediated $RR_c^{DE}(RR_c^{TIE}-1)/(RR_c^{TE}-1) = 1.72(1.028-1)/(1.77-1) = 6.3\%$. Most of the effect was found to be not through increasing cigarettes per day i.e. direct. If we now use the 3-way decomposition for risk ratios:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}$$

we find, $RR_c^{DE} = 1.72$, $RR_c^{IE} = 1.014$, $RERI_{mediated} = 0.036$. Thus of the excess relative risk, $(1.77 - 1) = 0.77$, for the total effect, $(1.72 - 1)/0.77 = 93.7\%$ is attributable to the pure direct effect, $(1.014 - 1)/0.77 = 1.7\%$ is attributable to the pure indirect effect, and $0.036/0.77 = 4.6\%$ is attributable to the mediated interaction and once again the overall proportion mediated is $1.7\% + 4.6\% = 6.3\%$. Of the mediated effect, which is itself small proportion, most of this mediated effect is due to the mediated interaction, rather than a pure indirect effect.

In another example, Ananth and VanderWeele[31] examined the extent to which the effect of placental abruption on perinatal mortality was mediated by medically induced preterm

birth using NCHS birth certificate files from 1995-2002. Allowing for potential interaction between abruption and preterm birth, and controlling for various socio-demographic variables, Ananth and VanderWeele found that the pure direct effect risk ratio was 10.18 (95% CI: 9.80, 10.58), the total indirect effect risk ratio was 1.35 (95% CI: 1.33, 1.38), and the total effect risk ratio was $10.18 \times 1.35 = 13.76$ (95% CI: 13.45, 14.08), with proportion mediated: $RR_c^{DE}(RR_c^{TIE} - 1)/(RR_c^{TE} - 1) = 10.18(1.35 - 1)/(13.76 - 1) = 28.1\%$. If we now use the 3-way decomposition for ratios:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}$$

we find $RR_c^{DE} = 10.18$, $RR_c^{IE} = 2.47$, $RERI_{mediated} = 2.11$. Thus of the excess relative risk, $(13.76 - 1) = 12.76$, for the total effect, $(10.18 - 1)/12.76 = 71.9\%$ is attributable to the pure direct effect, $(2.47 - 1)/12.76 = 11.5\%$ is attributable to the pure indirect effect, and $2.11/12.76 = 16.6\%$ is attributable to the mediated interaction and once again the overall proportion mediated is $11.5\% + 16.6\% = 28.1\%$.

Both these examples would require assumptions (i)-(iv) above held conditional on the covariates, a point to which we turn below. Discussion of these assumptions in their respective substantive contexts can be found in VanderWeele et al.[25] and Ananth and VanderWeele[31].

Discussion

The principle behind the results in this paper was utilizing the difference between the total indirect effect and the pure indirect effect (or, equivalently, the total direct effect and pure direct effect) as a measure of interaction, a mediated interactive effect. The interpretation of this difference between two indirect effects as a measure of interaction required justification. In the case of these effects defined as the difference of counterfactuals, we saw that the difference between the total indirect effect and the pure indirect effect was in fact the product of a causal interaction defined in terms of counterfactuals and the effect of the exposure on the mediator. We thus referred to this effect as a mediated interactive effect. For this

effect to be non-zero for an individual, an interaction had to be present and the exposure had to have an effect on the mediator. We saw also that for conditional effects on the difference scale that the conditional average of this mediated interactive effect could, under the assumption of no exposure-induced mediator-outcome confounder, be expressed as the product of the standard additive interaction contrast and the average conditional effect of the exposure on the mediator. In the case of the ratio scale, we saw that our interactive effect, again ultimately arising from taking the difference between a total indirect effect and a pure indirect effect, could be interpreted as a mediated analogue of the relative excess risk due to interaction. Further discussion of the 3-way decomposition for hazard ratios[9–11,32,33] or for direct and indirect effects in the presence of a mediator-outcome confounder affected by exposure[34] are given in the eAppendix. In all of these cases, we were thus able to decompose the total effect into a direct effect, and indirect effect and an interactive effect.

The chief difficulty in estimating the components of this 3-way decomposition are the strong assumptions required for their identification. These assumptions were no confounding of the exposure-outcome, mediator-outcome and exposure-mediator relationships, conditional on the covariates, and further that there is no mediator-outcome confounder affected by the exposure. These are strong assumptions; however the assumptions for the 3-way decomposition are no stronger than those that are required to estimate direct and indirect effects generally. Moreover, the extent to which violations of these assumptions would affect inference can be assessed through sensitivity analysis for the pure direct and indirect effects.[6,8] Future research could perhaps also adapt sensitivity analysis for interactions[35] to extend such techniques to the mediated interaction considered in this paper.

As noted above, and as has been done in the past, we could of course simply decompose a total or overall effect into two components: into the pure direct effect and the total indirect effect. This raises the question of which of these decompositions is to be preferred - the 2-way or the 3-way - and what it is that is ultimately of interest when we carry out effect decomposition. The two-way decomposition is simpler, but the three-way decomposition has

the potential to give additional insight. It allows us to assess how much of the total indirect effect is due to a mediated interaction versus a pure indirect effect. It makes clearer the role of interaction in mediation analysis. The utility of a method should arguably in the end be judged by the insight it gives into actual applications. Time and use over numerous data examples will thus ultimately make clearer the extent to which the three-way decomposition proposed in this paper is helpful in practice.

Acknowledgements. The author thanks Fan Mu and Nina Paynter for insightful questions that prompted this research and Sander Greenland and Eric Tchetgen Tchetgen for helpful comments on a draft of the paper.

Appendix

We first show that the decomposition in (1) holds. As noted in the text, we can decompose the total effect into a total indirect effect and a pure direct effect:

$$Y_1 - Y_0 = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$$

By adding and subtracting the pure indirect effect, $(Y_{0M_1} - Y_{0M_0})$, we obtain

$$Y_1 - Y_0 = (Y_{1M_0} - Y_{0M_0}) + (Y_{0M_1} - Y_{0M_0}) + \{(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})\}$$

The third quantity in this decomposition is the difference between the total indirect effect and the pure indirect effect. This quantity is also equal to the difference between the total direct effect and the pure direct effect, $(Y_{1M_1} - Y_{0M_1}) - (Y_{1M_0} - Y_{0M_0})\}$. We will consider the value that this difference between the total indirect and the pure indirect effect, $(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})$, might take under several different scenarios. If $M_0 = M_1$, then both indirect effects are 0 and so the difference is 0. If $M_1 = 1$ and $M_0 = 0$ then $(M_1 - M_0) = 1$ and the difference will be $(Y_{11} - Y_{10} - Y_{01} + Y_{00}) = (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$. If $M_1 = 0$ and $M_0 = 1$ then $(M_1 - M_0) = -1$ and the difference will be $(-Y_{11} + Y_{10} + Y_{01} - Y_{00}) =$

$(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$. Thus, the difference $(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})$ is always equal to $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$ and we have:

$$Y_1 - Y_0 = (Y_{1M_0} - Y_{0M_0}) + (Y_{0M_1} - Y_{0M_0}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0). \quad (1)$$

We will now establish the decomposition in (2) for conditional effects. We will in fact establish a more general result for an arbitrary exposure and mediator (not restricting to binary exposure and mediator). We have that $E[Y_a - Y_{a^*}|c] =$

$$\begin{aligned} & E[Y_{aM_a} - Y_{a^*M_a}|c] + E[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] \\ = \; & E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] + E[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] + \{E[Y_{aM_a} - Y_{a^*M_a}|c] - E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]\} \end{aligned}$$

where the first quantity is the conditional pure direct effect, the second is the conditional pure indirect effect and the third is the difference between the conditional total direct effect and the conditional pure direct effect. Under assumption (iv) that $Y_{am} \coprod M_{a^*}|C$ we have that this difference is:

$$\begin{aligned} & \{E[Y_{aM_a} - Y_{a^*M_a}|c] - E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]\} \\ = \; & \sum\nolimits_m E[Y_{am} - Y_{a^*m}|M_a = m, c]P(M_a = m|c) - \sum\nolimits_m E[Y_{am} - Y_{a^*m}|M_{a^*} = m, c]P(M_{a^*} = m|c) \\ = \; & \sum\nolimits_m E[Y_{am} - Y_{a^*m}|c]P(M_a = m|c) - \sum\nolimits_m E[Y_{am} - Y_{a^*m}|c]P(M_{a^*} = m|c) \\ = \; & \sum\nolimits_m E[Y_{am} - Y_{a^*m}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\ = \; & \sum\nolimits_m E[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \end{aligned}$$

where m^* is an arbitrary value of M, and where the first equality follows by iterated expectations, the second by assumption (iv), and the fourth because for some fixed level of m^*, $\sum_m E[Y_{a^*m}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} = 0$ and $\sum_m E[Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} = 0$. Thus, for arbitrary exposure and mediator, under (iv) we have

the decomposition of the conditional effect:

$$\begin{aligned} E[Y_a - Y_{a^*}|c] &= E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] + E[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] \\ &\quad + \sum\nolimits_m E[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \end{aligned}$$

where the first term is the pure direct effect, the second is the pure indirect effect, and the third is a mediated interactive effect. If we consider binary exposure and mediator with $a = 1, a^* = 0, m^* = 0$ we have

$$\begin{aligned} & \sum\nolimits_m E[Y_{1m} - Y_{0m} - Y_{10} + Y_{00}|c]\{P(M_1 = m|c) - P(M_0 = m|c)\} \\ = & \sum\nolimits_m E[Y_{1m} - Y_{0m}|c]\{P(M_1 = m|c) - P(M_0 = m|c)\} \\ = & E[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\ & + E[Y_{10} - Y_{00}|c]\{P(M_1 = 0|c) - P(M_0 = 0|c)\} \\ = & E[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\ & + E[Y_{10} - Y_{00}|c][1 - P(M_1 = 1|c) - \{1 + P(M_0 = 1|c)\}] \\ = & E[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\ & - E[Y_{10} - Y_{00}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\ = & E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]\{E[M_1|c] - E[M_0|c]\} \end{aligned}$$

and so we have

$$E[Y_1 - Y_0|c] = E[Y_{1M_0} - Y_{0M_0}|c] + E[Y_{0M_1} - Y_{0M_0}|c] + E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]E[M_1 - M_0|c], \quad (2)$$

thus establishing the decomposition in (2).

For the identification formulae in (3), under assumptions (i)-(iv) that

$$\begin{aligned} E[Y_{1M_0} - Y_{0M_0}|c] &= \sum\nolimits_m \{E[Y|A=1,m,c] - E[Y|A=0,m,c]\}P(m|A=0,c) \\ E[Y_{0M_1} - Y_{0M_0}|c] &= \sum\nolimits_m E[Y|A=0,m,c]\{P(m|A=1,c) - P(m|A=0,c)\} \end{aligned}$$

has been established elsewhere.[2] We have shown above that under (iv),

$$E[(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)|c] = E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]E[M_1 - M_0|c]$$

and under (i) and (ii) the first term in this product is equal to $\{E[Y|A=1,M=1,c] - E[Y|A=1,M=0,c] - E[Y|A=0,M=1,c] + E[Y|A=0,M=0,c]\}$ and under (iii) the second term in this product is equal to $E[M|A=1,c] - E[M|A=0,c]$, thus establishing the identification formulae in (3).

References

1. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology.* 1992; 3:143-155.

2. Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence.* San Francisco: Morgan Kaufmann; 2001:411-420.

3. van der Laan MJ, Petersen ML. Direct effect models. *International Journal of Biostatistics*, 2008, Article 23.

4. 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: 457-468.

5. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis with a dichotomous outcome. *American Journal of Epidemiology*, 2010;172:1339-1348.

6. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychological Methods, 2010;15:309-334.

7. Imai K, Keele L, Tingley D, Yamamoto T. Causal mediation analysis using R. In: H.D. Vinod (ed.), Advances in Social Science Research Using R. New York: Springer (Lecture Notes in Statistics), p.129-154, 2010.

8. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 2010;21:540-551.

9. Lange T, Hansen JV Direct and indirect effects in a survival context. Epidemiology, 2011;22:575-581.

10. VanderWeele TJ. Causal mediation analysis with survival data. Epidemiology, 2011;22:582-585.

11. Tchetgen Tchetgen EJ. On causal mediation analysis with a survival outcome. *International Journal of Biostatistics*, 2011;7:Article 33, 1-38.

12. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Technical Report.

13. Ten Have TR, Joffe MM. A review of causal estimation of effects in mediation analyses. *Stat Methods Med Res* 2012;21: 77-107

14. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In Highly Structured Stochastic Systems, Eds. P. Green, N.L. Hjort, and S. Richardson, 70-81. Oxford University Press, New York, 2003.

15. Pearl J. Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press, 2nd edition, 2009.

16. VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology 2009, 20(1):880-883.

17. Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2005;357-363.

18. VanderWeele TJ. Subtleties of explanatory language: what is meant by "mediation"? *European Journal of Epidemiology*, 26:343-346.

19. Suzuki E, Yamamoto E, Tsuda,T. Identification of operating mediation and mechanism in the sufficient-component cause framework. *European Journal of Epidemiology*, 2011;26:347-57.

20. Rothman KJ, Greenland S, Lash TL. "Concepts of interaction," chapter 5, in Modern Epidemiology, 3rd edition. Philadelphia: Lippincott Williams and Wilkins, 2008.

21. VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component cause framework. Epidemiology, 2007;18:329-339.

22. VanderWeele, T.J. and Robins, J.M. (2008). Empirical and counterfactual conditions for sufficient cause interactions. Biometrika, 95:49-61.

23. VanderWeele TJ. Sufficient cause interactions and statistical interactions. Epidemiology, 2009;20:6-13.

24. Rothman KJ. Modern Epidemiology. 1st ed. Little, Brown and Company, Boston, MA, 1986.

25. VanderWeele TJ, Asomaning K, Tchetgen Tchetgen EJ, Han Y, Spitz MR, Shete S, Wu X, Gaborieau V, Wang Y, McLaughlin J, Hung RJ, Brennan P, Amos CI, Christiani DC, Lin X. Genetic variants on 15q25.1, smoking and lung cancer: an assessment of mediation and interaction. American Journal of Epidemiology, 2012 doi: 10.1093/aje/kwr467.

26. Saccone SF, Hinrichs AL, Saccone NL, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet. 2007;16(1):36-49.

27. Spitz MR, Amos CI, Dong Q, et al. The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. J Natl Cancer Inst. 2008;100(21):1552-1556.

28. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008;452(7187):633-637.

29. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008;40(5):616-622.

30. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008;452(7187):638-642.

31. Ananth CV, VanderWeele TJ. Placental abruption and perinatal mortality with preterm delivery as a mediator: disentangling direct and indirect effects. American Journal of Epidemiology, 2011;174:99-108.

32. Li R, Chambless L. Test for additive interaction in proportional hazards models. Annals of Epidemiology, 2007;17:227-236.

33. VanderWeele TJ. Causal interactions in the proportional hazards model. Epidemiology, 2011;22:713-717.

34. Robins JM. A new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect. Mathematical Modelling, 1986;7:1393-1512.

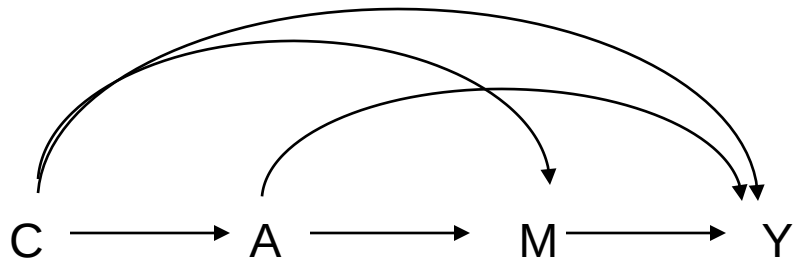
35. VanderWeele TJ, Mukherjee B, Chen J. Sensitivity analysis for interactions under unmeasured confounding. Statistics in Medicine, Article first published online: 4 OCT 2011, DOI: 10.1002/sim.4354.

Figure Legends

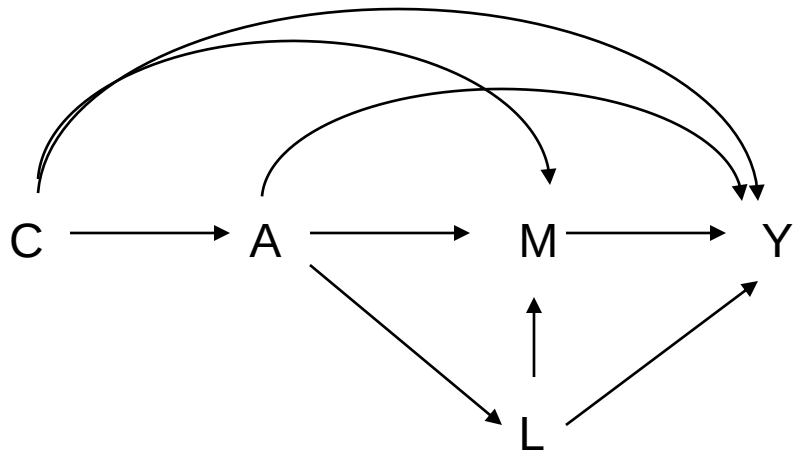
Figure 1: Mediation with exposure A, outcome Y, mediator M, and confounders C.

Figure 2: Mediation with a mediator-outcome confounder L that is affected by the exposure.

Figure



Figure



eAppendix for "A three-way decomposition of a total effect into direct, indirect, and interactive effects"

Proofs for 3-way Decomposition on a Ratio Scale

On a risk ratio scale, the conditional total effect risk ratio is defined by $RR_c^{TE} = E[Y_a|c]/E[Y_{a^*}|c]$; the pure direct effect risk ratio is defined by $RR_c^{DE} = E[Y_{aM_{a^*}}|c]/E[Y_{a^*M_{a^*}}|c]$ and the pure indirect effect risk ratio is defined by $RR_c^{IE} = E[Y_{a^*M_a}|c]/E[Y_{a^*M_{a^*}}|c]$. We have that:

$$\begin{aligned} E[Y_a - Y_{a^*}|c] &= E[Y_{aM_a} - Y_{a^*M_a}|c] + E[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] \\ E[Y_a - Y_{a^*}|c] &= E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] + E[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] \\ &\quad + \{E[Y_{aM_a} - Y_{a^*M_a}|c] - E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]\}. \end{aligned}$$

Dividing both sides of the equation by$E[Y_{a^*M_{a^*}}|c]$ gives:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + \left(\frac{E[Y_{aM_a}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{aM_{a^*}}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{a^*M_a}|c]}{E[Y_{a^*M_{a^*}}|c]} + 1 \right)$$

If we define $RERI_{mediated}$ by $RERI_{mediated} = \left(\frac{E[Y_{aM_a}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{aM_{a^*}}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{a^*M_a}|c]}{E[Y_{a^*M_{a^*}}|c]} + 1 \right)$ then we have

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}.$$

If we define the total indirect effect risk ratio as $RR_c^{TIE} = E[Y_{aM_a}|c]/E[Y_{aM_{a^*}}|c]$ then the total effect risk ratio decomposes as $RR_c^{TE} = RR_c^{TIE} \times RR_c^{DE}$. VanderWeele and Vansteelandt[5] proposed as a measure of the proportion mediated on the risk difference scale

the measure $\frac{RR_c^{DE}(RR_c^{TIE}-1)}{(RR_c^{TE}-1)}$. The numerator in this quantity is in fact equal to

$$\begin{aligned} RR_c^{DE}(RR_c^{TIE}-1) &= RR_c^{TE} - RR_c^{DE} \\ &= \frac{E[Y_{aM_a}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{aM_{a^*}}|c]}{E[Y_{a^*M_{a^*}}|c]} \\ &= \left(\frac{E[Y_{a^*M_a}|c]}{E[Y_{a^*M_{a^*}}|c]} - 1\right) + \left(\frac{E[Y_{aM_a}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{aM_{a^*}}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{a^*M_a}|c]}{E[Y_{a^*M_{a^*}}|c]} + 1\right) \\ &= (RR_c^{IE}-1) + RERI_{mediated}. \end{aligned}$$

i.e. the sum of the excess relative risk for the pure indirect effect plus the mediated relative excess risk due to interaction.

As shown in the Appendix to the text, under assumption (iv), we have for a binary exposure and mediator that

$$E[Y_{1M_1} - Y_{0M_1}|c] - E[Y_{1M_0} - Y_{0M_0}|c] = E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]E[M_1 - M_0|c].$$

Dividing both sides of the equation by $E[Y_{0M_0}|c]$ gives:

$$\left(\frac{E[Y_{1M_1}|c]}{E[Y_{0M_0}|c]} - \frac{E[Y_{1M_0}|c]}{E[Y_{0M_0}|c]} - \frac{E[Y_{0M_1}|c]}{E[Y_{0M_0}|c]} + 1\right) = \frac{1}{E[Y_{0M_0}|c]}E[Y_{11}-Y_{10}-Y_{01}+Y_{00}|c]E[M_1-M_0|c]$$

and thus

$$\begin{aligned} &\left(\frac{E[Y_{aM_a}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{aM_{a^*}}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{a^*M_a}|c]}{E[Y_{a^*M_{a^*}}|c]} + 1\right) \\ = &\frac{E[Y_{00}|c]}{E[Y_{0M_0}|c]}\left(\frac{E[Y_{11}|c]}{E[Y_{00}|c]} - \frac{E[Y_{10}|c]}{E[Y_{00}|c]} - \frac{E[Y_{01}|c]}{E[Y_{00}|c]} + 1\right)E[M_1 - M_0|c], \end{aligned}$$

from which we have

$$RERI_{mediated} = \frac{E[M_1 - M_0|c]E[Y_{00}|c]}{E[Y_{0M_0}|c]}RERI_{causal}.$$

A Three-way Decomposition on the Hazards Scale

When time-to-event data is under consideration a hazard scale is often employed. Natural direct and indirect effects have likewise been discussed for the hazards scale.[9–11] Lange and Hansen[9] considered the analysis of natural direct and indirect effects using an additive hazards model; VanderWeele considered the analysis of natural direct and indirect effects using a proportional hazards model. Let T denote a time-to-event outcome and let T_a denote the counterfactual event time if A had been set to a; likewise we let T_{am} denote the counterfactual event time if A had been set to a and M had been set to m. With these definitions we can also consider nested counterfactual event times. For example, $T_{aM_{a^*}}$ is an individual's event time if the exposure had been set to a and the mediator had been set to the level it would have been had exposure been a^*. We will use $\lambda_V(t)$ and $\lambda_V(t|c)$ for the hazard or conditional hazard at time t, that is the instantaneous rate of the event conditional on $V \geq t$.

We can decompose the overall difference in hazards for a total effect as the sum of natural indirect and direct effects on the hazard difference scale:

$$\lambda_{T_a}(t) - \lambda_{T_{a^*}}(t) = \left[\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t)\right] + \left[\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)\right].$$

We could further rewrite this as follows:

$$\begin{aligned}\lambda_{T_a}(t) - \lambda_{T_{a^*}}(t) &= \left[\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)\right] + \left[\lambda_{T_{a^*M_a}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)\right] \\ &\quad + \left(\left[\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t)\right] - \left[\lambda_{T_{a^*M_a}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)\right]\right)\end{aligned}$$

where the first term is the pure direct effect on the hazard difference scale, the second is the pure indirect effect on the hazard difference scale and the final term is the difference between the total indirect effect and the pure indirect effect on the hazard difference scale. If we then

divide both sides of this equation by $\lambda_{T_{a^*}}(t)$ we obtain:

$$\left(\frac{\lambda_{T_a}(t)}{\lambda_{T_{a^*}}(t)} - 1\right) = \left(\frac{\lambda_{T_{aM_{a^*}}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - 1\right) + \left(\frac{\lambda_{T_{a^*M_a}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - 1\right) + \left(\frac{\lambda_{T_{aM_a}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - \frac{\lambda_{T_{aM_{a^*}}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - \frac{\lambda_{T_{a^*M_a}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} + 1\right)$$

where the first term in this decomposition is the excess hazard ratio for the pure direct effect, the second is the excess hazard ratio for the pure indirect effect, and the third term is the hazard ratio equivalent of the mediated relative excess risk due to interaction[32,33].

A Three-way Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder

Consider a setting in which there is a variable L that is affected by exposure A and in turn affects both M and Y as in Figure 2. Although natural direct and indirect effects with M as the mediator are not identified in this setting[17], alternative effects which randomly set M to a value chosen from the distribution of a particular exposure level can be identified. Let $G_{a|c}$ denote a random draw from the distribution of the mediator amongst those with exposure status a conditional on $C = c$. Let a and a^* be two values of the exposure e.g. for binary exposure we may have $a = 1$ and $a^* = 0$. The effect $E(Y_{aG_{a|c}}|c) - E(Y_{aG_{a^*|c}}|c)$ is then the effect on the outcome of randomly assigning an individual who is given the exposure to a value of the mediator from the distribution of the mediator amongst those given exposure versus no exposure, conditional on covariates; this is an effect through the mediator. Next consider the effect $E(Y_{aG_{a^*|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)$; this is a direct effect comparing exposure versus no exposure with the mediator in both cases randomly drawn from the distribution of the population when given the absence of exposure, conditional on covariates. Finally, the effect $E(Y_{aG_{a|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)$ compares the expected outcome when having the exposure with the mediator randomly drawn from the distribution of the population when given the exposure, conditional on covariates to the expected outcome when not having the exposure with the mediator randomly drawn from the distribution of the population when not exposed, conditional on covariates. With effects thus defined we have the decomposition: $E(Y_{aG_{a|c}}|c) -$

$E(Y_{a^*G_{a^*|c}}|c) = \{E(Y_{aG_{a|c}}|c) - E(Y_{aG_{a^*|c}}|c)\} + \{E(Y_{aG_{a^*|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)\}$ so that the total effect decomposes into the sum of the effect through the mediator and the direct effect. These effects arise from randomly choosing for each individual a value of the mediator from the distribution of the mediator amongst all of those with a particular exposure.

We might further decompose this as follows:

$$\begin{aligned} E(Y_{aG_{a|c}}|c) - E(Y_{a^*G_{a^*|c}}|c) &= \{E(Y_{aG_{a^*|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)\} + \{E(Y_{a^*G_{a|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)\} \\ &\quad + [\{E(Y_{aG_{a|c}}|c) - E(Y_{a^*G_{a|c}}|c)\} - \{E(Y_{aG_{a^*|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)\}] \end{aligned}$$

where the first term in the decomposition is the randomized intervention analogue of the pure direct effect, the second is the randomized intervention analogue of the pure indirect effect, and the third is the difference between the randomized intervention analogue of the total direct effect and the pure direct effect. We now show that this third term in fact has the interpretation of an interaction. We have that

$$\begin{aligned} & \{E(Y_{aG_{a|c}}|c) - E(Y_{a^*G_{a|c}}|c)\} - \{E(Y_{aG_{a^*|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)\} \\ = & \sum\nolimits_m E[Y_{am} - Y_{a^*m}|G_{a|c} = m, c]P(G_{a|c} = m|c) - \sum\nolimits_m E[Y_{am} - Y_{a^*m}|G_{a^*|c} = m, c]P(G_{a^*|c} = m|c) \\ = & \sum\nolimits_m E[Y_{am} - Y_{a^*m}|c]P(M_a = m|c) - \sum\nolimits_m E[Y_{am} - Y_{a^*m}|c]P(M_{a^*} = m|c) \\ = & \sum\nolimits_m E[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \end{aligned}$$

where m^* is an arbitrary value of M. This final expression can be interpreted as a measure of interaction. For binary exposure and mediator, if we set $a = 1, a^* = 0, m^* = 0$ then as in

the Appendix of the text for we also have here:

$$
\begin{aligned}
& \sum\nolimits_m E[Y_{1m} - Y_{0m} - Y_{10} + Y_{00}|c]\{P(M_1 = m|c) - P(M_0 = m|c)\} \\
= & \sum\nolimits_m E[Y_{1m} - Y_{0m}|c]\{P(M_1 = m|c) - P(M_0 = m|c)\} \\
= & E[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
& + E[Y_{10} - Y_{00}|c]\{P(M_1 = 0|c) - P(M_0 = 0|c)\} \\
= & E[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
& + E[Y_{10} - Y_{00}|c][1 - P(M_1 = 1|c) - \{1 - P(M_0 = 1|c)\}] \\
= & E[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
& - E[Y_{10} - Y_{00}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
= & E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]\{E[M_1|c] - E[M_0|c]\}.
\end{aligned}
$$

Thus for binary exposure and mediator, even in the presence of an exposure-induced mediator-outcome confounder, we would have the three way effect decomposition:

$$
\begin{aligned}
E(Y_{1G_{1|c}}|c) - E(Y_{0G_{0|c}}|c) = & \{E(Y_{1G_{0|c}}|c) - E(Y_{0G_{0|c}}|c)\} + \{E(Y_{0G_{1|c}}|c) - E(Y_{0G_{0|c}}|c)\} \\
& + E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]\{E[M_1|c] - E[M_0|c]\}.
\end{aligned}
$$

To identify these effects the following conditions suffice: Assumptions (i) $Y_{am} \coprod A|C$ and (iii) $M_a \coprod A|C$ above, that conditional on C there is no unmeasured exposure-outcome or exposure-mediator confounding, along with an assumption (ii*) that $Y_{am} \coprod M|\{A, C, L\}$, i.e. that conditional on (A, C, L), there is no unmeasured confounding of the mediator-outcome relationship. These three assumptions would hold in the causal diagram in Figure 2. Under the three assumptions, each of these component are identified from data and it follows from

the g-formula[34] that:

$$\begin{aligned} E(Y_{aG_{a^*|c}}|c) - E(Y_{a^*G_{a^*|c}}|c) &= \sum\nolimits_{l,m} \{E[Y|a,l,m,c]P(l|a,c) - E[Y|a^*,l,m,c]P(l|a^*,c)\}P(m|a^*,c) \\ E(Y_{a^*G_{a|c}}|c) - E(Y_{a^*G_{a^*|c}}|c) &= \sum\nolimits_{l,m} E[Y|a^*,l,m,c]P(l|a^*,c)\{P(m|a,c) - P(m|a^*,c)\} \end{aligned}$$

and $E[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]\{E[M_1|c] - E[M_0|c]\}$

$$\begin{aligned} = \; & \{\sum\nolimits_l E[Y|A=1,l,m=1,c]P(l|A=1,c) - \sum\nolimits_l E[Y|A=1,l,m=0,c]P(l|A=1,c) \\ & - \sum\nolimits_l E[Y|A=0,l,m=1,c]P(l|A=0,c) + \sum\nolimits_l E[Y|A=0,l,m=0,c]P(l|A=0,c)\} \\ & \times \{E[M|A=1,c] - E[M|A=0,c]\} \end{aligned}$$

Proofs for Direct, Indirect and Interactive Effects from Regression Models

For Y and M continuous, under assumptions (i)-(iv) and correct specification of the regression models for Y and M:

$$\begin{aligned} E[Y|a,m,c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ E[M|a,c] &= \beta_0 + \beta_1 a + \beta_2' c, \end{aligned}$$

VanderWeele and Vansteelandt[4] showed that the pure direct effect was given by:

$$E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$$

and that the total indirect effect was given by

$$E[Y_{aM_a} - Y_{aM_{a^*}}|c] = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

and likewise the pure indirect effect by

$$E[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] = (\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*).$$

The mediated interactive effect is given by the difference between the total indirect effect and the pure indirect effect and is thus equal to:

$$\begin{aligned} & E[Y_{aM_a} - Y_{aM_{a^*}}|c] - E[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] \\ = \; & (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*) - (\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*) \\ = \; & \theta_3\beta_1(a - a^*)(a - a^*). \end{aligned}$$

Suppose now instead that Y were binary and M continuous, that assumptions (i)-(iv) held, that the outcome was rare and that the following regressions were correctly specified:

$$\begin{aligned} logit(P(Y = 1|a, m, c)) &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ E[M|a, c] &= \beta_0 + \beta_1 a + \beta_2' c. \end{aligned}$$

VanderWeele and Vansteelandt[5] derived expressions for natural direct and indirect effect odds ratio (risk ratios) from these two regressions that hold approximately provided the outcome is rare. The expressions hold exactly for risk ratios, even for common outcomes, if the logistic regression model is replaced by a log-linear model. As noted above, we can decompose the excess relative risk for a total effect into the sum of the excess relative risk for the pure direct effect, the excess relative risk for the pure indirect effect, and the mediated relative excess risk due to interaction:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}.$$

VanderWeele and Vansteelandt[5] showed that the pure direct effect risk ratio and the pure

indirect effect risk ratio were given by:

$$\begin{aligned} RR_c^{DE} &= \exp[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})] \\ RR_c^{IE} &= \exp[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)] \end{aligned}$$

where σ^2 is the variance of the error term in the linear regression model for M.

From the derivations of VanderWeele and Vansteelandt[5], we have that the total effect is given by:

$$RR_c^{TE} = \exp[\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})]$$

and from this it follows that $RERI_{mediated}$ is equal to:

$$\begin{aligned} & \left(\frac{E[Y_{1M_1}|c]}{E[Y_{0M_0}|c]} - \frac{E[Y_{1M_0}|c]}{E[Y_{0M_0}|c]} - \frac{E[Y_{0M_1}|c]}{E[Y_{0M_0}|c]} + 1\right) \\ = \; & \exp[\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})] \\ & - \exp[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})] \\ & - \exp[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)] + 1. \end{aligned}$$

Again standard errors for this expression could be derived using the delta method along the lines of the derivations in the Online Appendix of VanderWeele and Vansteelandt[5] or by using bootstrapping.