Harvard University
Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 170

A unification of mediation and interaction: a four-way decomposition

Tyler J. VanderWeele*

*Harvard University, tvanderw@hsph.harvard.edu

A unification of mediation and interaction: a four-way decomposition

Tyler J. VanderWeele

Abstract

It is shown that the overall effect of an exposure on an outcome, in the presence of a mediator with which the exposure may interact, can be decomposed into four components: (i) the effect of the exposure in the absence of the mediator, (ii) the interactive effect when the mediator is left to what it would be in the absence of exposure, (iii) a mediated interaction, and (iv) a pure mediated effect. These four components, respectively, correspond to the portion of the effect that is due to neither mediation nor interaction, to just interaction (but not mediation), to both mediation and interaction, and to just mediation (but not interaction). This four-way decomposition unites methods that attribute effects to interactions and methods that assess mediation. Certain combinations of these four components correspond to measures for mediation, while other combinations correspond to measures of interaction previously proposed in the literature. Prior decompositions in the literature are in essence special cases of this four-way decomposition. The four-way decomposition can be carried out using standard statistical models, and software is provided to estimate each of the four components. The four-way decomposition provides maximum insight into how much of an effect is mediated, how much is due to interaction, how much is due to both mediation and interaction together, and how much is due to neither.

Introduction

Methodology for mediation and interaction has developed rapidly over the past decade. Methods for effect decomposition to assess direct and indirect effects have shed light on mechanisms and pathways.[1–19] Other methods and measures have been useful in assessing how much of the effect of one exposure is due to its interaction with another.[20–24] In this paper we provide theory and methods to unite these effect decomposition and attribution methods for mediation and interaction. The paper's central result is that the total effect of an exposure on an outcome, in the presence of a mediator with which the exposure may interact, can be decomposed into four components: components due to just mediation, to just interaction, to both mediation and interaction, and to neither mediation nor interaction.

After presenting this four-way decomposition, we will discuss assumptions for identifying these four components from data and we will relate this four-way decomposition approach to various statistical models. We then discuss the relations between existing measures of mediation and interaction and each of the four components, and show how existing measures of mediation and interaction consist of different combinations of these four components. We show how different effect decomposition and attribution approaches for mediation and interaction can in fact be united within this four-fold framework; when some of the components are combined, the framework presented in this paper essentially collapses to approaches that have been used previously. The greatest insight, however, is arguably gained when the four-fold approach is employed and we illustrate this with an example from genetic epidemiology.

Notation

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 will suppose we want to compare two levels of the exposure, a and a^*; for binary exposure we would have $a = 1$ and $a^* = 0$. For simplicity we will consider the setting of a binary exposure and binary mediator; however more general results that are applicable to arbitrary exposures and mediators are given in the Appendix. We let Y_a and M_a denote respectively the potentially counterfactual values of the outcome and mediator that would have been observed had the exposure A been set to level a. The total effect (TE) of the exposure A on the outcome Y is defined as $Y_1 - Y_0$; the total effect of the exposure A on the mediator M is defined as $M_1 - M_0$. We will not in general ever know what these effects are at the individual level but we might hope to be able to estimate them on average for a population. For the first part of this paper, however, we will be concerned with concepts and only later will we turn to what can be identified with data and under what assumptions.

We will also need counterfactuals of another form. 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. The controlled direct effect, comparing exposure level $A = 1$ to $A = 0$ and fixing the mediator to level m is defined by $Y_{1m} - Y_{0m}$ 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 persons. Finally we will also later consider counterfactuals of the form $Y_{aM_{a^*}}$ which is the outcome Y that would have occurred if we fixed A to a and we fixed M to the level it would have taken if A had been a^*. We will also make some technical assumptions referred to as consistency and composition that 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 equal to the observed outcomes Y and M, respectively, 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.[4,18,25]

A Four-Fold Decomposition

We show in the Appendix that we can decompose the total effect (TE) of A on Y into the following four components:

$$\begin{aligned} Y_1 - Y_0 &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ &\quad + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0). \end{aligned} \tag{1}$$

The first component, $(Y_{10} - Y_{00})$, is the direct effect of the exposure A if the mediator were removed, i.e. fixed to $M = 0$. This effect is sometimes referred to as a 'controlled direct effect' (CDE).[1,2] The second component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, we will call a 'reference interaction' (INT_{ref}). The term $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$ is an additive interaction. It can be rewritten as $(Y_{11} - Y_{00}) - \{(Y_{10} - Y_{00}) + (Y_{01} - Y_{00})\}$ and will be non-zero for a person if the effect on the outcome of setting both the exposure and the mediator to present differs from the sum of the effect of having only the exposure present and the effect of having only the mediator present; additive interaction is generally considered of greatest public health importance[20–22]. The second component in the decomposition in (1) is the product of this additive interaction and M_0. Thus this second component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, is an additive interaction that only operates if the mediator is present in the absence of exposure i.e. when $M_0 = 1$. The third component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$, will be referred to as a 'mediated interaction' (INT_{med}). It is the same additive interaction contrast times $(M_1 - M_0)$. In other words it is an additive interaction that only operates if the exposure has an effect on the mediator so that $M_1 - M_0 \neq 0$. The final component, $(Y_{01} - Y_{00})(M_1 - M_0)$, is the effect of the mediator in the absence of the exposure, $Y_{01} - Y_{00}$, multiplied by the the effect of the exposure on the mediator itself, $M_1 - M_0$. It will be non-zero only if the mediator affects the outcome when the exposure is absent, and the exposure itself affects the mediator. We might refer to this final component as a 'mediated main effect' or, as will be explained below, as a 'pure indirect effect' (PIE).[1,2]

The intuition behind this decomposition is that if the exposure affects the outcome for a particular individual, then at least one of four things must be the case. Either the exposure might affect the outcome through pathways which do not require the mediator (i.e. the exposure affects the outcome even when the mediator is absent); in other words the first component is non-zero. Or alternatively, the exposure effect might operate only in the presence of the mediator (i.e. there is an interaction) and it might also be the case that the exposure itself is not necessary for the mediator to be present (i.e. the mediator itself would be present in the absence of the exposure, though the mediator is itself necessary for the exposure to have an effect on the outcome); in other words, the second component is non-zero. Or alternatively, the exposure effect might operate only in the presence of the mediator (i.e. there is an interaction) and it might also be the case that the exposure itself is in fact needed for the mediator to be present (i.e. the exposure causes the mediator, and the presence of the mediator is itself necessary for the exposure to have an effect on the outcome); in other words, the third component in non-zero. Or finally, it might alternatively be the case that the mediator can cause the outcome in the absence of the exposure, but the exposure is necessary for the mediator itself to be present; in other words, the fourth component is non-zero. The decomposition above, proved in the Appendix, provides a mathematical formalization of this intuition. We could thus rewrite our decomposition as:

$$TE = CDE + INT_{ref} + INT_{med} + PIE.$$

As with the total effect of the exposure on the outcome, $Y_1 - Y_0$, we cannot in general hope

to know the value of each of the four components for a particular individual, but below we will discuss assumptions under which we could estimate measures of these four components on average for a particular population. We will see below that under certain assumptions about confounding the average value of each of four components is given by the following empirical expressions:

$$\begin{aligned} E[CDE] &= (p_{10} - p_{00}) \\ E[INT_{ref}] &= (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\ E[INT_{med}] &= (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\ E[PIE] &= (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}. \end{aligned}$$

where $p_{am} = E(Y|A = a, M = m)$. If we let $p_a = E(Y|A = a)$ we will have following empirical decomposition:

$$\begin{aligned} p_{a=1} - p_{a=0} &= (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\ &\quad + (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\ &\quad + (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \end{aligned} \tag{1b}$$

With such average measures we would be able to assess how much of the total effect is due to (i) neither mediation nor interaction (the first component); how much is due to interaction but not mediation (the second component), how much is due to both mediation and interaction (the third component); and how much of the effect is due to mediation but not interaction (the fourth component). The four components of the total effect are summarized in Table 1.

Table 1. The Four Basic Components of the Total Effect (the following four components sum to the total effect $TE = Y_1 - Y_0$)

Effect	Counterfactual Definition	Empirical Analogue
Controlled Direct Effect (CDE)	$(Y_{10} - Y_{00})$	$(p_{10} - p_{00})$
Reference Interaction (INT_{ref})	$(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$	$(p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0)$
Mediated Interaction (INT_{med})	$(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$	$(p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}$
Pure Indirect Effect (PIE)	$(Y_{01} - Y_{00})(M_1 - M_0) = (Y_{0M_1} - Y_{0M_0})$	$(p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}$.

If we let $E[TE]$ denote the average total effect for the population (equal to $p_{a=1} - p_{a=0} = E(Y|A = 1) - E(Y|A = 0)$ in the absence of confounding), then we could also consider the proportion of the total effect that is due to each of these four components using the ratios $\frac{E[CDE]}{E[TE]}$, $\frac{E[INT_{ref}]}{E[TE]}$, $\frac{E[INT_{med}]}{E[TE]}$, and $\frac{E[PIE]}{E[TE]}$. We could also assess the overall proportion due to mediation by summing the proportions due to the mediated interaction and to the pure indirect effect, i.e. $\frac{E[INT_{med}]+E[PIE]}{E[TE]}$. We could likewise assess the overall proportion due to interaction by summing the proportions due to the reference interaction and to the mediated interaction, i.e. $\frac{E[INT_{ref}]+E[INT_{med}]}{E[TE]}$. Such proportion measures, however, generally only make sense reporting if all of the components are in the same direction (e.g. all positive or all negative). The statistical properties of such proportion measures can also be highly variable, and hence problematic, if the total effect is close to zero as might be the case if some of the components were positive and others negative. Similar comments pertain to other proportion measures described below.

We will first consider the no-confounding assumptions that allow us to estimate these four components on average, and statistical methods to carry out such estimation. We will later consider the relationships between this four-fold decomposition and other concepts from the literatures on mediation and interaction that involve effect decomposition and attribution.

Identification of the Effects

Our discussion thus far has been primarily conceptual. As we have noted, the individual level effects in the four-way decomposition cannot be identified from the data, but under certain no-confounding assumptions the four components can be identified from the data on average for a population. As discussed further in the Appendix, for a causal diagram interpreted as non-parametric structural equation models of Pearl,[18] the following four assumptions suffice to identify each of the four components from the data: (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, A); (iii) the effect the exposure A on the mediator M is unconfounded conditional on C; and (iv) none of the mediator-outcome confounders are themselves affected by the exposure. These are the same four assumptions that are often used in the literature on mediation.[2,4,5] If we let $X \perp\!\!\!\perp Y|Z$ denote that X is independent of Y conditional on Z, then these four assumptions stated formally in terms of counterfactual independence are: (i) $Y_{am} \perp\!\!\!\perp A|C$, (ii) $Y_{am} \perp\!\!\!\perp M|\{A, C\}$, (iii) $M_a \perp\!\!\!\perp A|C$, and (iv) $Y_{am} \perp\!\!\!\perp M_{a^*}|C$. Note that assumption (iv) requires that none of the mediator-outcome confounders are themselves affected by the exposure. This assumption would hold in Figure 1 but would be violated in Figure 2.

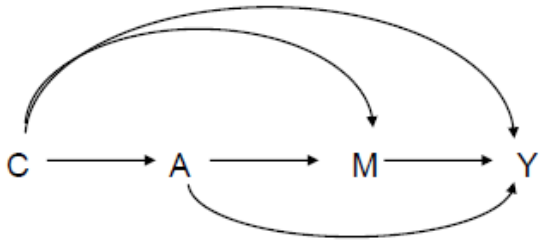

Figure 1: Mediation with exposure A, outcome Y, mediator M, and confounders C.

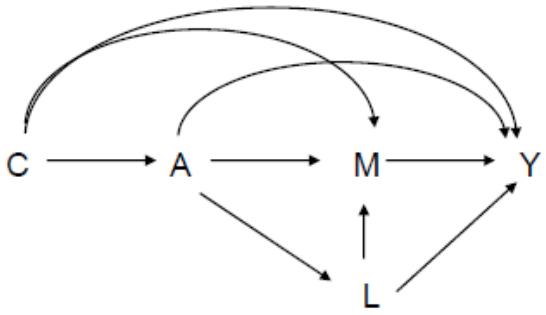

Figure 2: Mediation with a mediator-outcome confounder L that is affected by the exposure.

If these four assumptions held without covariates then we would have the empirical formulae given

above:

$$\begin{aligned} E[CDE] &= (p_{10} - p_{00}) \\ E[INT_{ref}] &= (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\ E[INT_{med}] &= (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\ E[PIE] &= (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}. \end{aligned}$$

More general formulae involving covariates and with arbitrary exposures and mediator (rather than binary) are given in the Appendix.

The counterfactual statement of assumption (iv), $Y_{am} \perp\!\!\!\perp M_{a^*}|C$, is somewhat controversial as it involves what are sometimes called 'cross-world' independencies. It would hold in Figure 1 interpreted as a non-parametric structural equation model,[18] but may not hold under other interpretations of causal diagrams.[19] We noted above that the empirical equivalent of our four-way decomposition was (1b). As shown in the Appendix, this decomposition holds without any assumptions at all about confounding. However, to interpret each of the components causally does require assumptions about confounding. Assumptions (i)-(iv) above allow for interpreting each of the components as population average causal effects of each of the four components in the four-way individual level counterfactual decomposition: CDE, INT_{ref}, INT_{med}, and PIE. In the Appendix we also discuss how a slightly weaker interpretation is also valid essentially under just assumptions (i)-(iii) alone, without requiring the more controversial assumptions (iv).

Also of interest is the fact that the controlled direct effect, CDE, only requires assumption (i) and (ii) to be identified.[1,2] This does not require the more controversial cross-world independence assumptions. The average controlled direct effect is sometimes subtracted from the average total effect to get a proportion eliminated measure $E[PE] := E[TE] - E[CDE]$. Whenever we can identify the total effect and the controlled direct effect we can calculate this portion eliminated measure. Interestingly, as described further below, the four-way decomposition gives a more mechanistic interpretation of this portion eliminated measure: the portion eliminated is the sum of the reference interaction, the mediated interaction, and the pure indirect effect ($PE = INT_{ref} + INT_{med} + PIE$) i.e. it is the portion due to either mediation or interaction or both. We cannot empirically separate apart these three components without using stronger assumptions such as (i)-(iv) above. However, whenever we can identify the total effect and the controlled direct effect (which we can do under much weaker assumptions) we can obtain also the sum of the three other components since they are simply the difference between the total effect and the controlled direct effect.

Relation to Statistical Models

Suppose that assumptions (i)-(iv) hold, that Y and M are continuous and that the following regression models for Y and M are correctly specified:

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

It is shown in the eAppendix that for exposure levels a and a^*, and for setting the mediator to 0 in the controlled direct effect (see Online Appendix for other settings of mediator for the CDE),

the four components are given by:

$$\begin{aligned} E[CDE|c] &= \theta_1(a-a^*) \\ E[INT_{ref}|c] &= \theta_3(\beta_0+\beta_1 a^*+\beta_2' c)(a-a^*) \\ E[INT_{med}|c] &= \theta_3\beta_1(a-a^*)(a-a^*) \\ E[PIE|c] &= (\theta_2\beta_1+\theta_3\beta_1 a^*)(a-a^*) \end{aligned}$$

If the exposure were binary, the pure direct, reference interaction, mediated interaction, and pure indirect effects would, respectively, simply be: θ_1, $\theta_3(\beta_0+\beta_2' c)\}$, $\theta_3\beta_1$, and $\theta_2\beta_1$. 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. SAS code to implement this approach to obtain estimates and confidence intervals is provided in the eAppendix. The eAppendix likewise provides a straightforward modeling approach, and SAS code, when the mediator is binary rather continuous.

Binary Outcomes and the Ratio Scale

Thus far we have been considering the definition of these four components 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. By dividing the decomposition in (1b) by $p_{a=0}$ we can rewrite this decomposition on the ratio scale as

$$\begin{aligned} RR_{a=1}-1 &= \kappa(RR_{10}-1)+\kappa(RR_{11}-RR_{10}-RR_{01}+1)P(M=1|A=0) \\ &\quad +\kappa(RR_{11}-RR_{10}-RR_{01}+1)\{P(M=1|A=1)-P(M=1|A=0)\} \\ &\quad +\kappa(RR_{01}-1)\{P(M=1|A=1)-P(M=1|A=0)\} \end{aligned} \tag{2}$$

where $RR_{a=1}=\frac{p_{a=1}}{p_{a=0}}$ is the relative risk for exposure A comparing $A=1$ to the reference category $A=0$, and $RR_{am}=\frac{p_{am}}{p_{00}}$ is the relative risk for comparing categories $A=a, M=m$ to the reference category $A=0, M=0$, and where κ is a scaling factor which is given by $\kappa=\frac{p_{00}}{p_{a=0}}$. Note also here that the term, $(RR_{11}-RR_{10}-RR_{01}+1)$, is Rothman's excess relative risk due to interaction ($RERI$), and is a measure of additive interaction using ratios.[20]

The decomposition in (2) involves decomposing the excess relative risk for the exposure A, $RR_{a=1}-1$, into four components on the excess relative risk scale involving, as before, (i) the controlled direct effect of A when $M=0$, (ii) a reference interaction, (iii) a mediated interaction, and (iv) a mediated main effect. Note that although the right hand side of the decomposition involves a scaling factor κ, if what we are interested in is the proportion of the effect attributable to each of the components, then if we take any particular component and divide it by the sum of all the components, then the scaling drops out. The proportion of the effect attributable to each of the four components is thus given by the expressions in Table 2.

Table 2. Proportion attributable to the controlled direct effect (PA_{CDE}), the reference interaction (PA_{INTref}), the mediated interaction (PA_{INTmed}) and the pure indirect effect (PA_{PIE}) when using a ratio scale.

$$
\begin{aligned}
PA_{CDE} &= \frac{(RR_{10}-1)}{(RR_{10}-1)+(RERI)P(M=1|A=1)+(RR_{01}-1)\{P(M=1|A=1)-P(M=1|A=0)\}} \\
PA_{INTref} &= \frac{(RERI)P(M=1|A=0)}{(RR_{10}-1)+(RERI)P(M=1|A=1)+(RR_{01}-1)\{P(M=1|A=1)-P(M=1|A=0)\}} \\
PA_{INTmed} &= \frac{(RERI)\{P(M=1|A=1)-P(M=1|A=0)\}}{(RR_{10}-1)+(RERI)P(M=1|A=1)+(RR_{01}-1)\{P(M=1|A=1)-P(M=1|A=0)\}} \\
PA_{PIE} &= \frac{(RR_{01}-1)\{P(M=1|A=1)-P(M=1|A=0)\}}{(RR_{10}-1)+(RERI)P(M=1|A=1)+(RR_{01}-1)\{P(M=1|A=1)-P(M=1|A=0)\}}.
\end{aligned}
$$

The four-fold proportion attributable measures given in Table 2 allow us to estimate the proportion of the total effect attributable just to mediation (PA_{PIE}), just due to interaction (PA_{INTref}), due to both mediation and interaction (PA_{INTmed}), or due to neither mediation nor interaction (PA_{CDE}). Further technical details concerning the four-way decomposition on the ratio scale and for obtaining estimates and confidence intervals using logistic regression for the outcome along with linear regression for a continuous mediator or a second logistic regression for a binary mediator is given in the eAppendix. SAS code to implement this approach is also given in the eAppendix.

Illustration

We will consider a data example from genetic epidemiology to illustrate the four-way decomposition. Specifically, we consider the extent to which the effect of chromosome 15q25.1 rs8034191 C alleles on lung cancer risk is mediated by cigarettes smoked per day and/or due to interaction with this smoking measure. 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. VanderWeele et al.[31] used methods from the causal mediation analysis literature to assess whether the effect was direct or indirect and found that most of the effect was not mediated by cigarettes per day (the total indirect effect was very small and the pure direct effect was large). In large meta-analyses, Truong et al.[32] found no association between the genetic variants amongst never smokers suggesting strong interaction between the variants and smoking behavior; VanderWeele et al.[31] likewise reported statistical evidence of interaction. Here we will use the four-way decomposition to assess how much of the effect is due to each of the components.

We use data on 1836 cases and 1452 controls from a lung cancer case-control study at Massachusetts General Hospital; see Miller et al.[33] or VanderWeele et al.[31] for further details on the study. As the exposure we compare 2 versus 0 C alleles, and use cigarettes per day as the mediator (the square root of this measure is used so that the measure is more normally distributed). Covariates adjusted for in the analysis include sex, age, education, and smoking duration. Analyses are restricted to Caucasians. Because the outcome, lung cancer, is rare, odds ratios approximate risk ratios. We fit a logistic regression model for lung cancer on the variants, smoking, their interaction, and the covariates; and a linear regression model for smoking on the variants and covariates. Confidence intervals are obtained using the delta method. Details of this modeling approach in the context of the four-way decomposition are given in the eAppendix; SAS code is also provided. Results are summarized in Table 3.

Table 3. Proportions of the effect of genetic variants on lung cancer due to mediation and interaction with smoking (cigarettes per day)

Component	Excess Relative Risk	Proportion Attributable	Other Proportions
CDE	0.30 (95% CI: $-0.19, 0.79$)	39% (95% CI: $-11\%, 89\%$)	Overall Proportion to Interaction:
INT_{ref}	0.42 (95% CI: $0.11, 0.73$)	55% (95% CI: $8\%, 101\%$)	59% (95% CI: $9\%, 109\%$)
INT_{med}	0.034 (95% CI: $-0.02, 0.09$)	4% (95% CI: $-3\%, 11\%$)	Overall Proportion to Mediation:
PIE	0.014 (95% CI: $-0.01, 0.04$)	2% (95% CI: $-1\%, 5\%$)	6% (95% CI: $-3\%, 15\%$)
Total	0.77 (95% CI: $0.33, 1.21$)	100%	

The overall risk ratio comparing 2 versus 0 C alleles was 1.77 (95% CI: $1.33, 2.21$) for an excess relative risk of $1.77 - 1 = 0.77$ (95% confidence interval = $0.33, 1.21$). We decompose this excess relative risk into the four components. The component due to the pure indirect effect is 0.014 (95% CI: $-0.01, 0.04$); the component due to the mediated interaction is 0.034 (95% CI: $-0.02, 0.09$); the component due to the reference interaction is 0.42 (95% CI: $0.11, 0.73$); and the component due to the controlled direct effect (if smoking were fixed to 0) is 0.30 (95% CI: $-0.19, 0.79$). The four components sum to the excess relative risk: $0.014 + 0.034 + 0.42 + 0.30 \approx 0.77$. Of the four components, the reference interaction is most substantial, highlighting the important role of interaction in this context. The overall proportion mediated (the sum of the pure indirect effect and the mediated interaction, divided by the excess relative risk) is quite small 6.2% (95% CI: $-2.7\%, 15.1\%$), as had been indicated in the analyses of VanderWeele et al.[31] The overall proportion attributable to interaction (the reference interaction plus the mediated interaction, divided by the excess relative risk) is relatively substantial 59.2% (95% CI: $9.2\%, 109.3\%$). Mediation may play a role here (and probably does as the variants do affect smoking and smoking affects lung cancer) but interaction, between the variants and smoking, is clearly much more important in this context.

Relation to Mediation Decompositions

We will first discuss the relations between the four components above and concepts from the mediation analysis literature, and we will then discuss relations with the interaction analysis literature. As above, our four-fold decomposition is:

$$\begin{aligned} Y_1 - Y_0 &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ &\quad + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0). \end{aligned}$$

The first component, $(Y_{10} - Y_{00})$, is referred to in the mediation analysis literature as a 'controlled direct effect' (CDE) of the exposure when fixing the mediator to level $M = 0$. We can also consider controlled direct effects which set M to a level other than 0 and in the Discussion section and further in the Appendix we consider four-way decompositions involving these alternative controlled direct effects. The fourth component in the four-way decomposition, $(Y_{01} - Y_{00})(M_1 - M_0)$, what we referred to above as a 'mediated main effect' is in fact equivalent to what in the mediation analysis literature is sometimes referred to as a 'pure indirect effect' (PIE). It is shown in the Appendix that:

$$PIE := Y_{0M_1} - Y_{0M_0} = (Y_{01} - Y_{00})(M_1 - M_0).$$

The counterfactual contrast, $Y_{0M_1} - Y_{0M_0}$, in the mediation analysis literature is referred to as a 'pure indirect effect'[1] or as a type of 'natural direct effect'.[2] This contrast $Y_{0M_1} - Y_{0M_0}$ compares what would happen to the outcome if the mediator were changed from the level M_0 (the level it would be in the absence of the exposure) to M_1 (the level it would be in the presence of exposure) while in both counterfactual scenarios fixing the exposure itself to be absent. It will be non-zero if and only if the exposure changes the mediator (so that M_0 and M_1 are different) and the mediator

itself has an effect on the outcome even in the absence of the exposure. However, this is, in fact, the same quantity as what we had in our decomposition above, namely $(Y_{01}-Y_{00})(M_1-M_0)$. Note that writing the pure indirect effect as $(Y_{01}-Y_{00})(M_1-M_0)$ gives a representation of the pure indirect effect that does not require nested counterfactuals of the form Y_{0M_1}. This may be of interest as sometimes objections are made to the pure indirect effect on the grounds that nested counterfactuals of the form Y_{0M_1} are difficult to interpret. The third component, $(Y_{11}-Y_{10}-Y_{01}+Y_{00})(M_1-M_0)$, was recently considered in the mediation analysis literature and called a 'mediated interaction' (INT_{med}).[34] As discussed in the Appendix and elsewhere[34], this mediated interaction can also be written as $(Y_{1M_1}-Y_{0M_1}-Y_{1M_0}+Y_{0M_0})$. The component we have not yet considered, the second component, $(Y_{11}-Y_{10}-Y_{01}+Y_{00})(M_0)$, what we referred to above as a 'reference interaction' (INT_{ref}) has no analogue in the current literature. However, it is shown in the Appendix that the sum of the first and second component does have an analogue in the mediation analysis literature and it is equal to what is sometimes called in the mediation analysis literature the 'pure direct effect' (PDE) defined as $Y_{1M_0}-Y_{0M_0}$ which compares what would happen to the outcome in the presence versus the absence of the exposure if, in both cases, the mediator were set to whatever it would be for that individual in the absence of exposure. In other words we have that

$$\begin{aligned} PDE := Y_{1M_0}-Y_{0M_0} &= (Y_{10}-Y_{00})+(Y_{11}-Y_{10}-Y_{01}+Y_{00})(M_0) \\ &= CDE+INT_{ref}. \end{aligned}$$

The pure direct effect is the sum of a controlled direct effect and, our second component, the reference interaction. If in our four-way decomposition above we replace the first two components with the pure direct effect and write the fourth component as the pure indirect effect we obtain:

$$Y_1-Y_0 = PDE+INT_{med}+PIE. \tag{3}$$

In other words, we can decompose the total effect into a pure direct effect, a pure indirect effect, and a mediated interaction. This decomposition in (3) was the three-way decomposition provided by VanderWeele[34] in 2013. However, even this three-way decomposition is relatively new and prior to this, a two-way decomposition was the norm in the mediation analysis literature. As discussed in the Appendix and in VanderWeele,[34] the sum of the mediated interaction and the pure indirect effect is equal to what in the mediation analysis literature is sometimes called a 'total indirect effect' (TIE), defined as $Y_{1M_1}-Y_{1M_0}$. Whereas, the pure indirect effect, $Y_{0M_1}-Y_{0M_0}$, compares changing the mediator from M_0 to M_1 while fixing the exposure itself to be absent, the total indirect effect, $Y_{1M_1}-Y_{1M_0}$, compares changing the mediator from M_0 to M_1 fixing the exposure to present. With the total indirect so defined we have $TIE = PIE+INT_{med}$ i.e. $(Y_{1M_1}-Y_{1M_0}) = (Y_{0M_1}-Y_{0M_0})+(Y_{11}-Y_{10}-Y_{01}+Y_{00})(M_1-M_0)$. We can then combine the mediated interaction and the pure indirect effect in the decomposition in (3), into a total indirect effect to obtain the more standard 2-way decomposition in the mediation analysis literature:

$$Y_1-Y_0 = PDE+TIE. \tag{4}$$

This is the decomposition that has been used most often in the causal inference literature when assessing direct and indirect effects; this two-way decomposition was first proposed in 1992 by Robins and Greenland[1]; and it is the decomposition that most of the existing software packages for causal mediation analysis have focused on.[8,16] This two-way decomposition also provides the counterfactual formalization for the decompositions typically employed in the social science literature on mediation.[35] However, as we have seen above, the pure direct effect is itself a combination of

two components: a controlled direct effect and the reference interaction ($PDE = CDE + INT_{ref}$). And the total indirect effect is a combination of two components, the pure indirect effect and the mediated interaction ($TIE = PIE + INT_{med}$). When these effects are estimated on average, a proportion mediated measure, $\frac{E[TIE]}{E[TE]}$, is sometimes used which can also be re-written as $\frac{E[TIE]}{E[TE]} = \frac{E[INT_{med}]+E[PIE]}{E[TE]}$.

Yet another decomposition is worth noting in the mediation analysis literature. Sometimes the mediated interaction in the decomposition in (3) is combined with pure direct effect, rather than with the pure indirect effect, for an alternative two-way decomposition. As discussed in the Appendix and in VanderWeele[34], the sum of the mediated interaction and the pure direct effect is equal to what in the mediation analysis literature is sometimes called a 'total direct effect' (TDE),[1] defined as $Y_{1M_1} - Y_{0M_1}$. The total and the pure direct effects are sometimes also called 'natural direct effects'[2] and the total and the pure indirect effects are sometimes called 'natural indirect effects'[2]. A summary of the various composite effects is given in Table 4.

Table 4. Composite Effects

Effect	Counterfactual Definition	Composite Relationship
Total Indirect Effect (TIE)	$(Y_{1M_1} - Y_{1M_0})$	$TIE = PIE + INT_{med}$
Pure Direct Effect (PDE)	$(Y_{1M_0} - Y_{0M_0})$	$PDE = CDE + INT_{ref}$
Total Direct Effect (TDE)	$(Y_{1M_1} - Y_{0M_1})$	$TDE = CDE + INT_{ref} + INT_{med}$
Portion Eliminated (PE)	$(Y_1 - Y_0) - (Y_{10} - Y_{00})$	$PE = PIE + INT_{ref} + INT_{med}$
Portion Attributable to Interaction (PAI)	$(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1)$	$PAI = INT_{ref} + INT_{med}$

Of interest here is that the total direct effect contains three components: the controlled direct effect, the reference interaction, and the mediated interaction. As we move from the first to third of these components, we see they increasing involve the mediator in more substantial ways. The controlled direct effect, $(Y_{10} - Y_{00})$, operates completely independent of the mediator; for this to be non-zero the direct effect must be present even when the mediator is absent. The reference interaction, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, requires the mediator to operate but the effect does not come about by the exposure changing the mediator - it simply requires that the mediator itself is present even when the exposure is absent; the effect is 'unmediated' in the sense that it does not operate by the exposure changing the mediator, but it requires the presence of the mediator nonetheless. The third component, the mediated interaction, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$, is a type of mediated effect; it requires that the exposure change the mediator; but it is also a direct effect insofar as an interaction must also be present (the effect of the exposure is different for different levels of the mediator); the third component thus not only involves the mediator but it is a mediated effect, and a direct effect as well. It is for this reason that it is sometimes combined with the pure indirect effect to obtain the total indirect effect, and sometimes combined with the pure direct effect to obtain the total direct effect.

When we combine the pure direct effect and mediated interaction to get the total direct effect, $TDE := Y_{1M_1} - Y_{0M_1} = PDE + INT_{med}$, we have the alternative 2-way decomposition of the total effect into the sum of the total direct effect and the pure indirect effect:

$$Y_1 - Y_0 = TDE + PIE. \tag{5}$$

This decomposition was likewise proposed by Robins and Greenland[1] in 1992. Relatively easy-to-use software is currently available to estimate the components of the two-way decompositions in (4) and (5) on average for a population, under the assumptions described later in the paper. Note that in the decomposition in (5), the total direct effect consists of three of the four basic components (the controlled direct effect, the reference interaction, and the mediated interaction), whereas the

pure indirect effect constitutes a single component. The mediated interaction is, however, arguably part of the effect that is mediated and thus, when questions of mediation are of interest, it is arguably (4), rather than (5), that is to be preferred when assessing the extent of mediation.[9,10,34] However, whether the pure indirect effect or the total indirect effect is of interest may depend upon the context.[6]

A final measure that is used in the mediation analysis literature is sometimes referred to as the "portion eliminated" (PE).[1,12] As noted above, this is generally defined as the difference between the total effect and the controlled direct effect: $PE := (Y_1 - Y_0) - CDE$. It is the portion of the effect of the exposure that would remain if the mediator were fixed to 0. The portion eliminated may be of interest insofar as it allows one to assess how much of the effect of the exposure can be eliminated or prevented by intervening on the mediator; for this reason it is sometimes argued to be of policy interest.[1,6,12] The four-way decomposition above in fact shows that this portion eliminated measure is equal to the sum of the other three components: the reference interaction, the mediated interaction, and the pure indirect effect i.e. $PE = INT_{ref} + INT_{med} + PIE$ and we can write the total effect as $TE = CDE + PE$. The four-way decomposition provides a causal interpretation for the difference between the total effect and the controlled direct effect: it is the portion of the effect attributable to mediation, or interaction, or both. When the portion eliminated is estimated at the population level, sometimes a proportion eliminated measure is also calculated as $\frac{E[TE]-E[CDE]}{E[TE]}$ which we could also rewrite as $\frac{E[INT_{ref}]+E[INT_{med}]+E[PIE]}{E[TE]}$; note that this is different from the proportion mediated measure considered earlier which was $\frac{E[NIE]}{E[TE]} = \frac{E[INT_{med}]+E[PIE]}{E[TE]}$. The proportion eliminated includes in the numerator the reference interaction (since this part of the effect is eliminated if the mediator is removed); the proportion mediated does not include the reference interaction in the numerator (since this is not part of the mediated effect).[12]

We have seen then a number of different decompositions. However, when we are interested in questions of mediation, we need not choose between the two-way decompositions, or even the three-way decomposition, but can in fact use the decomposition into four components above so as to assess the portion of the total effect that is attributable just to mediation, just to interaction, to both mediation and interaction, or to neither mediation nor interaction. The four-way decomposition allows us to accomplish this. The various decompositions within the context of mediation are summarized in Table 5, but the four-way decomposition here essentially provides a framework which encompasses them all.

Table 5. Mediation Decompositions

Number of Components	Decomposition
2-Way Decomposition[a]	$TE = TIE + PDE$
2-Way Decomposition[b]	$TE = TDE + PIE$
2-Way Decomposition[c]	$TE = CDE + PE$
3-Way Decomposition[d]	$TE = PDE + PIE + INT_{med}$
4-Way Decomposition[e]	$TE = CDE + INT_{ref} + INT_{med} + PIE$

a $(Y_1 - Y_0) = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$
b $(Y_1 - Y_0) = (Y_{1M_1} - Y_{0M_1}) + (Y_{0M_1} - Y_{0M_0})$
c $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + [(Y_1 - Y_0) - (Y_{10} - Y_{00})]$
d $(Y_1 - Y_0) = (Y_{0M_1} - Y_{0M_0}) + (Y_{1M_0} - Y_{0M_0}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$
e $Y_1 - Y_0 = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0)$

Relation to Interaction Decompositions

VanderWeele and Tchetgen Tchetgen[24] recently considered attributing a portion of the total effect of one exposure on an outcome that is due to an interaction with a second exposure. Here we will relate this to the four-way decomposition above. Our four-way decomposition above was expressed as:

$$\begin{aligned} Y_1 - Y_0 &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ &\quad + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0). \end{aligned} \tag{1}$$

which we also wrote as: $TE = CDE + INT_{ref} + INT_{med} + PIE$. Suppose now that instead of considering how much of the total effect is mediated versus direct, as in the previous section, we were interested in the portion due to interaction. In our four-way decomposition, two of the four components (the second and the third involve) an interaction. We could thus define the portion due to interaction as their sum: $PAI := INT_{ref} + INT_{med} = (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) = (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1)$ and we would then have the 3-way decomposition:

$$\begin{aligned} TE &= CDE + PAI + PIE \\ &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1) + (Y_{01} - Y_{00})(M_1 - M_0). \end{aligned} \tag{6}$$

The total effect can be decomposed into the effect of A with M absent (CDE), a pure indirect effect (PIE), and a portion due to interaction (PAI). Consider now the empirical analogue of this decomposition using the expressions in (1b). We let $p_{am} = E[Y|A = a, M = m]$ and, $p_a = E[Y|A = a]$, and $p_m = E[Y|M = m]$ and we have from (1b): $(p_{a=1} - p_{a=0}) =$

$$(p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 1) + (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}. \tag{7}$$

We again have the decomposition of the average total effect of A, into what is essentially the average controlled direct effect, the average portion attributable to interaction, and the average pure indirect effect. The middle component is the component due to interaction and the proportion of the effect due to interaction could then be assessed by: $(p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 1)/(p_{a=1} - p_{a=0})$.

In fact, the decomposition given above in (7) is that which VanderWeele and Tchetgen[24] used when attributing effects to interactions. Several points are worth noting. First, the decomposition in (6) and (7) for the portion attributable to interaction follows quite clearly from the four-way decomposition. The decomposition in (6) is the decomposition at the individual counterfactual level analogous to the empirical decomposition in (7) given by VanderWeele and Tchetgen Tchetgen.[24] Second, VanderWeele and Tchetgen Tchetgen considered two cases, one in which A and M were independent and one in which they are not. The decomposition in (7) was that which was proposed when A affected M. When A and M are independent, the decomposition in (7) reduces to $(p_{a=1} - p_{a=0}) = (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1)$ and we likewise have a similar decomposition for the total effect of M on Y: $(p_{m=1} - p_{m=0}) = (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(A = 1)$. Likewise, on a ratio scale, when A does not affect M, the third and fourth components in Table 2 become 0 and we are left with $PA_{CDE} = \frac{(RR_{10}-1)}{(RR_{10}-1)+(RERI)P(M=1)}$ and $PA_{INTref} = \frac{(RERI)P(M=1)}{(RR_{10}-1)+(RERI)P(M=1)}$ which are also the expressions given by VanderWeele and Tchetgen Tchetgen[24] for attributing effects to interactions on a ratio scale. When A affects M, the decomposition for the total effect of A on Y is altered and we must use the decomposition in (7). Finally, when A does not affect Y, we have an analogous individual counterfactual level decomposition as (6) then reduces to: $TE =$

$(Y_{10}-Y_{00})+(Y_{11}-Y_{10}-Y_{01}+Y_{00})(M)$ since when A does not affect M, $M_1 = M_0 = M$. All of this also follows from our four-way decomposition which encompasses all of the prior decompositions. These decompositions are all summarized in Table 6.

Table 6. Interaction Decompositions

Number of Components	Decomposition
2-Way Decomposition (No Mediation)[a]	$TE = CDE + PAI$
3-Way Decomposition[b]	$TE = CDE + PAI + PIE$
4-Way Decomposition[c]	$TE = CDE + INT_{ref} + INT_{med} + PIE$

a $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M)$

b $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1) + (Y_{01} - Y_{00})(M_1 - M_0)$

c $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0)$

Although in the more general setting when A affects M, we can estimate the portion due to interaction on the average level using the three-way decomposition in (7), there is no need to use only a three-way decomposition; we can instead use the four-way decomposition in (1) and the empirical expressions in (1b) to further divide the portion due to interaction into that which is due to interaction but not mediation (the reference interaction, $E[INT_{ref}]$) and the portion due to interaction and mediation (the mediated interaction $E[INT_{med}]$). Such a four-way decomposition, in which the portion attributed to interaction is itself further divided may shed additional insight.

Perhaps most importantly, this four-way decomposition, which helps better understand the portions of a total effect due to interaction, is exactly the same decomposition that was used above to shed insight into what portions of the total effect were mediated and which portions were direct. The same four-way decomposition was useful in assessing both mediation and interaction. The same four components are used in assessing mediation and interaction, but the components are combined in different ways to assess these different phenomena. However, the four-way decomposition itself essentially provides a unification of these phenomena of mediation and interaction. The four-fold decomposition underlies the various more specific decompositions in assessing both mediation and interaction.

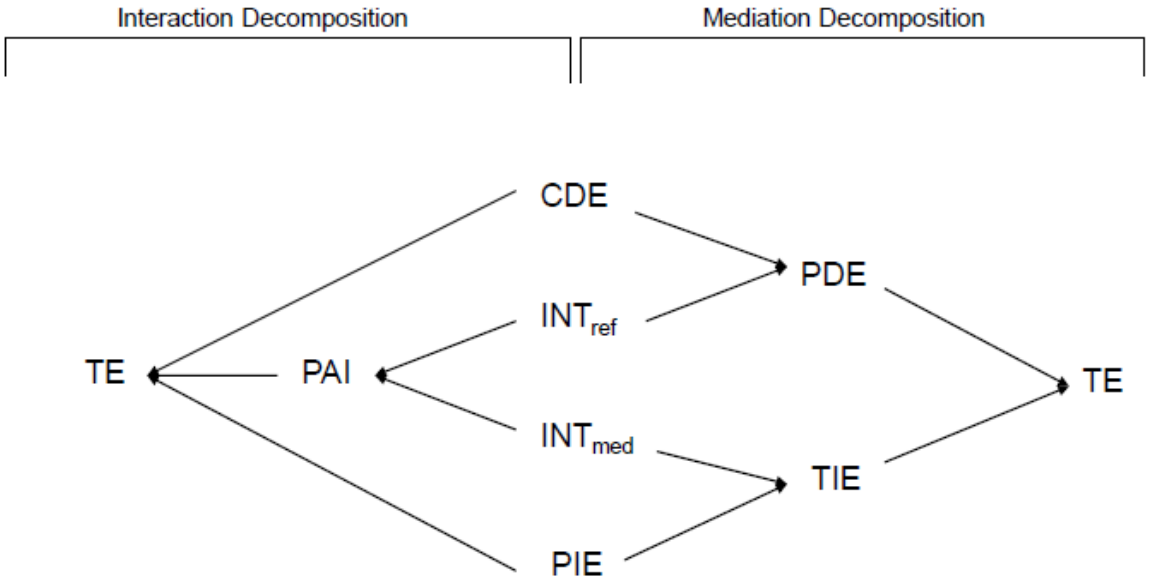

Figure 3: The four-fold decomposition encompasses both decompositions for mediation and interaction. For interaction, the reference interaction (INT_{ref}) and the mediated interaction (INT_{med}) combine to the portion attributable to interaction (PAI). The portion attributable to interaction (PAI) combine with the controlled direct effect (CDE) and the pure indirect effect (PIE) to give the total effect (TE). For mediation, the controlled direct effect and the reference interaction (INT_{ref}) combine to give the pure direct effect (PDE); the pure indirect effect (PIE) combines with the mediated interaction (INT_{med}) to give the total indirect effect (TIE); and the pure direct effect (PDE) combines with total indirect effect (TIE) to give the total effect (TE).

As illustrated in Figure 3, the four components form the backbone of both the various mediation decompositions (Figures 3-5) and the interaction decomposition (Figure 3).

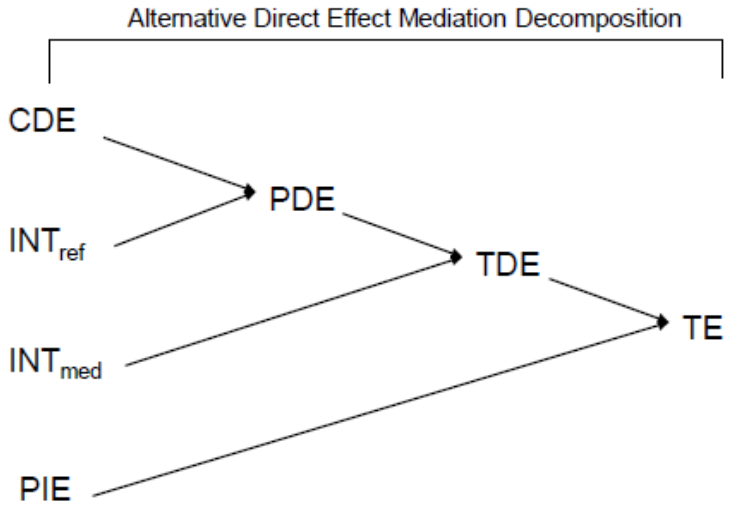

Figure 4: As an alternative mediation decomposition, the controlled direct effect and the reference interaction (INT_{ref}) combine to give the pure direct effect (PDE); the pure direct effect (PDE) and the mediated interaction (INT_{med}) combine to give the total direct effect (TDE); and the total direct effect (TDE) and the pure indirect effect (PIE) combine to give the total effect (TE).

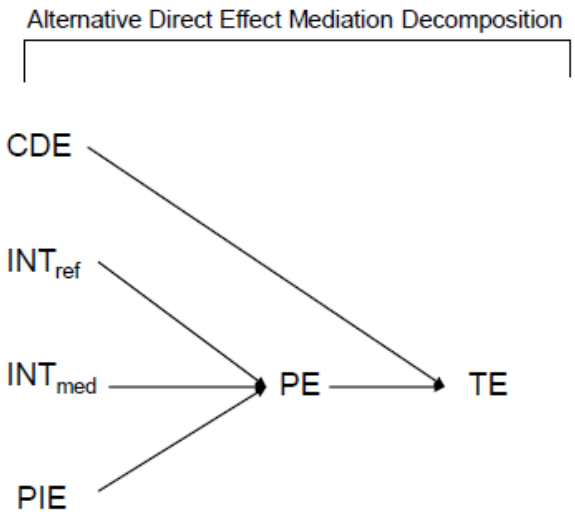

Figure 5: As an alternative mediation decomposition, the difference between the total effect (TE) and the controlled direct effect (PE) is sometimes called the portion elimianted (PE) and it is equal to the sum of the reference interaction (INT_{ref}), the mediated interaction (INT_{med}), and the pure indirect effect (PIE).

Once, again, however, the greatest insight is arguably gained when the four-fold approach is used to assess simultaneously the portions of the total effect that are due just to mediation, just to interaction, to both mediation and interaction, and to neither mediation nor interaction.

Discussion

The four-way decomposition here encompasses and unites previous decompositions in the literature, both concerning mediation and concerning interaction. The results here have also provided a mechanistic interpretation to the difference between a total effect and a controlled direct effect; this contrast has been used to assess policy implications and it is more easily identified than many other causal quantities concerning mediation; the results here show that it also has a mechanistic interpretation as well. We have also shown how the four-way decomposition in this paper can be carried out on a difference scale and on a ratio scale, we have related the various components to standard statistical models, and in the eAppendix we have provided software code to carry out the estimation of the various components of the decomposition using such regression models. We have seen that in addition to reporting the four components, an investigator can also easily report, along with these, the overall proportion attributable to interaction, the overall proportion mediated, and the proportion of the effect that would be eliminated if the mediator were removed. As seen in the empirical example in genetic epidemiology, the approach described here can shed considerable insight into the relationships between an exposure and a mediator with an outcome, and into the role of both mediation and interaction in these relationships.

In the text here we have focused on a binary exposure and binary mediator, and the controlled direct effect we have been considering is that in which the mediator is fixed to being absent. Much more general results are given in the Appendix and the approach in fact applies to arbitrary exposures and mediators. Moreover, instead of focusing on a controlled direct effect that fixes the

mediator to be absent, one can consider controlled direct effects that fix the mediator to some other level, m^*. Similar four-way decompositions can be carried out wherein the first component is the controlled direct effect with the mediator fixed to level m^*. When this is done, the reference interaction term changes because, with the mediator fixed to m^* (rather than 0), the controlled direct effect then picks up some of the effect of the interaction between the exposure and the mediator. With the controlled direct effect in which the mediator is fixed to m^*, the interpretation of the reference interaction is then the portion of the effect due to the interaction between the exposure and the mediator that is not mediated, and also not captured by the controlled direct effect. Again, the results in the Appendix cover very general settings and will thus likely be of use in a variety of contexts. The code in the eAppendix likewise provides practical and relatively easy-to-use software tools to implement the approaches here in a wide range of settings. The central limitations of the approach developed here is the strong assumptions being made about confounding; these are, however, similar assumptions to those made in the literature on mediation that only focuses on simpler decompositions. Future research could examine the robustness of each of the four components to confounding and measurement error. For example, recent work indicates that interaction terms may be more robust to confounding,[36] but that interaction terms when the two exposures are correlated may be particularly sensitive to measurement error;[37,38] different components may be robust to different forms of bias. Future work could also extend existing sensitivity analysis techniques for mediation and interaction[7,8,36–38] to each of the four components.

Prior work on mediation within the counterfactual framework has accommodated potential interaction. The approach here makes the role of interaction, and its separate contribution beyond mediation, clearer, and unites, within a single framework, the phenomena of mediation and interaction.

Appendix

In the Appendix we will no longer restrict attention to binary exposure and mediator and will consider an arbitrary exposure and mediator. We will assume we are comparing two exposure levels a and a^*. We give the general four-way decomposition result in Proposition 1.

Proposition 1. For any level m^* of M we have $Y_a - Y_{a^*}$

$$\begin{aligned} &= (Y_{am^*} - Y_{a^*m^*}) + \sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}) 1(M_{a^*} = m) \\ &\quad + \sum\nolimits_m (Y_{am} - Y_{a^*m}) \{1(M_a = m) - 1(M_{a^*} = m)\} + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \end{aligned}$$

Proof. We have that $Y_a - Y_{a^*}$

$$\begin{aligned} &= Y_{aM_a} - Y_{a^*M_{a^*}} \\ &= (Y_{aM_a} - Y_{a^*M_a}) + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \\ &= (Y_{aM_{a^*}} - Y_{a^*M_{a^*}}) + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) + (Y_{aM_a} - Y_{a^*M_a} - Y_{aM_{a^*}} + Y_{a^*M_{a^*}}) \\ &= (Y_{am^*} - Y_{a^*m^*}) + \{(Y_{aM_{a^*}} - Y_{a^*M_{a^*}}) - (Y_{am^*} - Y_{a^*m^*})\} \\ &\quad + (Y_{aM_a} - Y_{a^*M_a} - Y_{aM_{a^*}} + Y_{a^*M_{a^*}}) + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \\ &= (Y_{am^*} - Y_{a^*m^*}) + \sum\nolimits_m \{(Y_{am} - Y_{a^*m}) - (Y_{am^*} - Y_{a^*m^*})\} 1(M_{a^*} = m) \\ &\quad + \sum\nolimits_m \{(Y_{am} - Y_{a^*m}) 1(M_a = m) - (Y_{am} - Y_{a^*m}) 1(M_{a^*} = m)\} + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \\ &= (Y_{am^*} - Y_{a^*m^*}) + \sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}) 1(M_{a^*} = m) \\ &\quad + \sum\nolimits_m (Y_{am} - Y_{a^*m}) \{1(M_a = m) - 1(M_{a^*} = m)\} + (Y_{a^*M_a} - Y_{a^*M_{a^*}}). \blacksquare \end{aligned}$$

The four components of the decomposition in general form are thus

$$\begin{aligned} CDE(m^*) &: = (Y_{am^*} - Y_{a^*m^*}) \\ INT_{ref}(m^*) &: = \sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}) 1(M_{a^*} = m) \\ INT_{med} &: = \sum\nolimits_m (Y_{am} - Y_{a^*m}) \{1(M_a = m) - 1(M_{a^*} = m)\} \\ PIE &: = (Y_{a^*M_a} - Y_{a^*M_{a^*}}). \end{aligned}$$

Note we can also rewrite $INT_{med} = \sum_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}) \{1(M_a = m) - 1(M_{a^*} = m)\}$ and we can rewrite $PIE = \sum_m (Y_{a^*m} - Y_{a^*m^*}) \{1(M_a = m) - 1(M_{a^*} = m)\}$. Doing so with binary A and M and setting $a = 1, a^* = 0, m^* = 0$ gives us the decomposition in (1) in the text:

$$\begin{aligned} Y_1 - Y_0 &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ &\quad + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0). \end{aligned} \tag{1}$$

The decomposition also has an empirical analogue given in the next Proposition.

Proposition 2. For any level m^* of M we have $E[Y|a,c] - E[Y|a^*,c] =$

$$
\begin{aligned}
= \quad & \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\} \\
& + \int \{E[Y|a,m,c] - E[Y|a^*,m,c] - E[Y|a,m^*,c] + E[Y|a^*,m^*,c]\} dP(m|a^*,c) \\
& + \int \{E[Y|a,m,c] - E[Y|a^*,m,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
& + \int E[Y|a^*,m,c]\{dP(m|a,c) - dP(m|a^*,c)\}.
\end{aligned}
$$

Proof. We have that $E[Y|a,c] - E[Y|a^*,c]$

$$
\begin{aligned}
= \quad & E[Y|a,m^*,c] - E[Y|a^*,m^*,c] + \{E[Y|a,c] - E[Y|a,m^*,c]\} - \{E[Y|a^*,c] - E[Y|a^*,m^*,c]\} \\
= \quad & E[Y|a,m^*,c] - E[Y|a^*,m^*,c] + \int \{E[Y|a,m,c] - E[Y|a,m^*,c]\} dP(m|a,c) \\
& - \int \{E[Y|a^*,m,c] - E[Y|a^*,m^*,c]\} dP(m|a^*,c) \\
= \quad & \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\} \\
& + \int \{E[Y|a,m,c] - E[Y|a^*,m,c]\} - \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\} dP(m|a,c) \\
& + \int \{E[Y|a^*,m,c] - E[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
= \quad & \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\} \\
& + \int \{E[Y|a,m,c] - E[Y|a^*,m,c]\} - \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\} dP(m|a^*,c) \\
& + \int \{E[Y|a,m,c] - E[Y|a^*,m,c]\} - \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
& + \int \{E[Y|a^*,m,c] - E[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\}. \\
= \quad & \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\} \\
& + \int \{E[Y|a,m,c] - E[Y|a^*,m,c] - E[Y|a,m^*,c] + E[Y|a^*,m^*,c]\} dP(m|a^*,c) \\
& + \int \{E[Y|a,m,c] - E[Y|a^*,m,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
& + \int E[Y|a^*,m,c]\{dP(m|a,c) - dP(m|a^*,c)\}. \blacksquare
\end{aligned}
$$

Note we can also rewrite the third term as $\int \{E[Y|a,m,c] - E[Y|a^*,m,c]\} - \{E[Y|a,m^*,c] - E[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\}$ and the fourth term as $\int \{E[Y|a^*,m,c] - E[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\}$. Doing so with binary A and M,

and setting $a = 1, a^* = 0, m^* = 0$ gives decomposition (1b) in the text:

$$\begin{aligned} p_{a=1} - p_{a=0} &= (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\ &\quad + (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\ &\quad + (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \end{aligned} \tag{1b}$$

Note the decomposition in Proposition 2 is a property of the expectations and probabilities. It does not require confounding assumptions. However, to interpret the components as causal effects, confounding assumptions are required. We will begin our discussion of confounding by first considering non-parametric structural equations.[18] Consider the following four confounding assumptions: (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, A); (iii) the effect the exposure A on the mediator M is unconfounded conditional on C; and (iv) none of the mediator-outcome confounders are themselves affected by the exposure. If we let $X \perp\!\!\!\perp Y|Z$ denote that X is independent of Y conditional on Z, then these four assumptions stated formally in terms of counterfactual independence are: (i) $Y_{am} \perp\!\!\!\perp A|C$, (ii) $Y_{am} \perp\!\!\!\perp M|\{A, C\}$, (iii) $M_a \perp\!\!\!\perp A|C$, and (iv) $Y_{am} \perp\!\!\!\perp M_{a^*}|C$.

Proposition 3. Under assumptions (i)-(iv) we have:

$$\begin{aligned} E[CDE(m^*)|c] &= \{E[Y|a, m^*, c] - E[Y|a^*, m^*, c]\} \\ E[INT_{ref}(m^*)|c] &= \int \{E[Y|a, m, c] - E[Y|a^*, m, c] - E[Y|a, m^*, c] + E[Y|a^*, m^*, c]dP(m|a^*, c) \\ E[INT_{med}|c] &= \int \{E[Y|a, m, c] - E[Y|a^*, m, c]\}\{dP(m|a, c) - dP(m|a^*, c)\} \\ E[PIE|c] &= \int E[Y|a^*, m, c]\{dP(m|a, c) - dP(m|a^*, c)\}. \end{aligned}$$

Proof. The first equality is established by Robins[39], the fourth by Pearl[2], the third by VanderWeele[34]. For the second equality we have $E[INT_{ref}(m^*)|c]$

$$\begin{aligned} &= E\left[\sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*})1(M_{a^*} = m)|c\right] \\ &= \int_m E[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]dP(M_{a^*} = m|c) \\ &= \int_m \{E[Y_{am}|c] - E[Y_{a^*m}|c] - E[Y_{am^*}|c] + E[Y_{a^*m^*}|c]\}dP(M_{a^*} = m|c) \\ &= \int_m \{E[Y_{am}|a, m, c] - E[Y_{a^*m}|a^*, m, c] - E[Y_{am^*}|a, m^*, c] + E[Y_{a^*m^*}|a^*, m^*, c]\}dP(M_{a^*} = m|a^*, c) \\ &= \int_m \{E[Y|a, m, c] - E[Y|a^*, m, c] - E[Y|a, m^*, c] + E[Y|a^*, m^*, c]\}dP(M = m|a^*, c) \end{aligned}$$

where the second equality follows by assumption (iv) and the fourth by assumptions (i)-(iii). In fact, the other three equalities in Proposition 3 can be established in much the same way.■

We can also interpret the terms in the decomposition in Proposition 2 causally under assumptions (i)-(iii) alone, though the causal interpretation is slightly weaker.

Proposition 4. Under assumptions (i)-(iii) we have:

$$\{E[Y|a,m^*,c]-E[Y|a^*,m^*,c]\}=E[Y_{am^*}-Y_{a^*m^*}|c]$$

$$\int\{E[Y|a,m,c]-E[Y|a^*,m,c]-E[Y|a,m^*,c]+E[Y|a^*,m^*,c]\}dP(m|a^*,c)$$

$$=\int E[Y_{am}-Y_{a^*m}-Y_{am^*}+Y_{a^*m^*}|c]dP(M_{a^*}|c)$$

$$\int\{E[Y|a,m,c]-E[Y|a^*,m,c]\}\{dP(m|a,c)-dP(m|a^*,c)\}$$

$$=\int E[Y_{am}-Y_{a^*m}|c]\{dP(M_a|c)-dP(M_{a^*}|c)$$

$$\int E[Y|a^*,m,c]\{dP(m|a,c)-dP(m|a^*,c)\}=\int E[Y_{a^*m}|c]\{dP(M_a|c)-dP(M_{a^*}|c)\}.$$

.

Proof. The first equality is established by Robins[39], the second in the final four lines of the proof of Proportion 3 above, the third in VanderWeele[34] and the fourth, using slightly different notation by Didelez et al.[40]■

Note we can also rewrite the right hand side of the third equality as $\int\{E[Y_{am}-Y_{a^*m}|c]-E[Y_{am^*}-Y_{a^*m^*}|c]\}\{dP(M_a|c)-dP(M_{a^*}|c)$ and the right hand side of the fourth equality as $\int\{E[Y_{a^*m}-Y_{a^*m^*}|c]\}\{dP(M_a|c)-dP(M_{a^*}|c)\}$. The right hand side of the equalities in Proposition 4 are causal quantities but rather than directly taking population averages of the four components of the decomposition, the effects of A and M on Y are integrated over the distribution of M under different exposure settings. As discussed further in the eAppendix, these effects can be interpreted as randomized interventional analogues of the four components of the decomposition. They only require assumptions (i)-(iii) for identification (i.e. they do not require the more controversial cross-world independence assumption (iv)) but the causal interpretation of these randomized interventional analogues is somewhat weaker.

Finally, we note that some earlier literature on interaction with binary exposures made use of different response types where individuals were classified according to their joint counterfactual outcomes $(Y_{00},Y_{01},Y_{10},Y_{11})$. Under the response type classification for mediation given by Hafeman and VanderWeele[41] in which it is assumed that the monotonicity assumption that Y_{am} is non-decreasing in a and in m holds, the four components of the four-decomposition could be written as $CDE(0)=Y_{(2)}+Y_{(4)}$, $INT_{ref}(0)=M_{(1)}(Y_{(8)}-Y_{(2)})$, $INT_{med}=M_{(2)}(Y_{(8)}-Y_{(2)})$, and $PIE=M_{(2)}(Y_{(2)}+Y_{(6)})$ where $M_{(1)}$ is a binary indicator such that $M_{(1)}=1$ if $M_0=M_1=1$ and $M_{(1)}=0$ otherwise; $M_{(2)}$ is a binary indicator such that $M_{(2)}=1$ if $M_0=0,M_1=1$ and $M_{(2)}=0$ otherwise; $Y_{(2)}$ is a binary indicator such that $Y_{(2)}=1$ if $(Y_{00}=0,Y_{01}=1,Y_{10}=1,Y_{11}=1)$ and $Y_{(2)}=0$ otherwise; $Y_{(4)}$ is a binary indicator such that $Y_{(4)}=1$ if $(Y_{00}=0,Y_{01}=0,Y_{10}=1,Y_{11}=1)$ and $Y_{(4)}=0$ otherwise; $Y_{(6)}$ is a binary indicator such that $Y_{(6)}=1$ if $(Y_{00}=0,Y_{01}=1,Y_{10}=0,Y_{11}=1)$ and $Y_{(6)}=0$ otherwise; and $Y_{(8)}$ is a binary indicator such that $Y_{(8)}=1$ if $(Y_{00}=0,Y_{01}=1,Y_{10}=1,Y_{11}=1)$ and $Y_{(8)}=0$ otherwise.

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

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. Hafeman D, Schwartz S. Opening the black box: a motivation for the assessment of mediation. International Journal of Epidemiology, 2009;38:838-845.

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

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

9. VanderWeele TJ. Subtleties of explanatory language: what is meant by "mediation"? *European Journal of Epidemiology*, 2011;26:343-346.

10. 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.

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

12. VanderWeele TJ. Policy-relevant proportions for direct effects. Epidemiology, 2013;24:175-176.

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

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

15. Tchetgen Tchetgen, E.J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. Annals of Statistics, 40(3):1816-1845.

16. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological Methods, 2013; 18:137-150.

17. 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.

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

19. Robins JM, Richardson TS. Alternative graphical causal models and the identification of direct effects. In P. Shrout (Ed.): Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures. Oxford University Press, 2010.

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

21. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. Epidemiology, 1992; 3:452-56.

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

23. VanderWeele TJ. Reconsidering the denominator of the attributable proportion for interaction. European Journal of Epidemiology, 2013; 28:779-784.

24. VanderWeele TJ, Tchetgen Tchetgen EJ. Attributing effects to interactions. Epidemiology, in press.

25. VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology, 2009;20:880-883.

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. 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; 175:1013-1020.

32. Truong T, Hung RJ, Amos CI, et al. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. J Natl Cancer Inst. 2010;102(13):959-971.

33. Miller DP, Liu G, De Vivo I, et al. Combinations of the variant genotypes of GSTP1, GSTM1, and p53 are associated with an increased lung cancer risk. Cancer Res. 2002;62(10):2819-2823.

34. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. Epidemiology, 2013; 24:24:224-232.

35. MacKinnon, D. P. (2008). Introduction to Statistical Mediation Analysis. New York: Erlbaum.

36. VanderWeele TJ, Mukherjee B, Chen J. Sensitivity analysis for interactions under unmeasured confounding. Statistics in Medicine, 2012;31:2552-2564.

37. Valeri, L., Lin, X., and VanderWeele, T.J., Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. Technical Report.

38. Valeri L. and VanderWeele TJ. The estimation of direct and indirect causal effects in the presence of a misclassified binary mediator. Biostatistics.

39. 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.

40. Didelez V, Dawid AP, Geneletti S. Direct and indirect effects of sequential treatments. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, Arlington, VA: AUAI Press; 2006:138-146.

eAppendix for "A unification of mediation and interaction: a four-way decomposition" by Tyler J. VanderWeele

1. Continuous Outcomes and Linear Regression Models

1.1 Continuous Outcome, Continuous Mediator

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] and VanderWeele[34] showed that the average controlled direct effect, the pure indirect effect, and the mediated interaction conditional on covariates $C = c$ were given by:

$$\begin{aligned} E[CDE(m^*)|c] &= (\theta_1 + \theta_3 m^*)(a - a^*) \\ E[PIE|c] &= (\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*) \\ E[INT_{med}|c] &= \theta_3\beta_1(a - a^*)(a - a^*). \end{aligned}$$

They also showed that the pure direct effect was given by $E[PDE|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$. The reference interaction is then given by difference between the the pure direct effect and the controlled direct effect:

$$\begin{aligned} E[INT_{ref}(m^*)|c] &= \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*) - (\theta_1 + \theta_3 m^*)(a - a^*) \\ &= \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c - m^*)\}(a - a^*). \end{aligned}$$

Standard errors for these expressions could be derived using the delta method along the lines of the derivations in VanderWeele and Vansteelandt[4] or by using bootstrapping.

1.2 Continuous Outcome, Binary Mediator

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

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

Valeri and VanderWeele[16] show that the average controlled direct effect and the average pure indirect effect are given by:

$$\begin{aligned} E[CDE(m^*)|c] &= (\theta_1 + \theta_3 m^*)(a - a^*) \\ E[PIE|c] &= (\theta_2 + \theta_3 a^*)\{\frac{exp[\beta_0 + \beta_1 a + \beta_2' c]}{1 + exp[\beta_0 + \beta_1 a + \beta_2' c]} - \frac{exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1 + exp[\beta_0 + \beta_1 a^* + \beta_2' c]}\}. \end{aligned}$$

The reference interaction is given by the difference between the pure direct effect and the controlled direct effect, which were both given by Valeri and VanderWeele[16]:

$$\begin{aligned} E[INT_{ref}(m^*)|c] &= \{\theta_1(a-a^*)\} + \{\theta_3(a-a^*)\}\frac{exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]} - (\theta_1+\theta_3 m^*)(a-a^*) \\ &= \theta_3(a-a^*)\left(\frac{exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]} - m^*\right) \end{aligned}$$

The mediated interaction is given by the difference between the total indirect effect and the pure indirect effect, which were also both given by Valeri and VanderWeele[16]:

$$\begin{aligned} E[INT_{med}|c] &= (\theta_2+\theta_3 a)\{\frac{exp[\beta_0+\beta_1 a + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a + \beta_2^{'}c]} - \frac{exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}\}. \\ &\quad -(\theta_2+\theta_3 a^*)\{\frac{exp[\beta_0+\beta_1 a + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a + \beta_2^{'}c]} - \frac{exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}\} \\ &= \theta_3(a-a^*)\{\frac{exp[\beta_0+\beta_1 a + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a + \beta_2^{'}c]} - \frac{exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}{1+exp[\beta_0+\beta_1 a^* + \beta_2^{'}c]}\}. \end{aligned}$$

2. Decomposition on a Ratio Scale and Logistic Regression Models

2.1. Four-way Decomposition on a Ratio Scale

From Proposition 1 in the text we have $Y_a - Y_{a^*}$

$$\begin{aligned} &= (Y_{am^*} - Y_{a^*m^*}) + \sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*})1(M_{a^*} = m) \\ &\quad + \sum\nolimits_m (Y_{am} - Y_{a^*m})\{1(M_a = m) - 1(M_{a^*} = m)\} + (Y_{a^*M_a} - Y_{a^*M_{a^*}}). \end{aligned}$$

Taking expectations conditional on $C = c$ gives: $E(Y_a - Y_{a^*}|c)$

$$\begin{aligned} &= E(Y_{am^*} - Y_{a^*m^*}|c) + \sum\nolimits_m E[(Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*})1(M_{a^*} = m)|c] \\ &\quad + \sum\nolimits_m E[(Y_{am} - Y_{a^*m})\{1(M_a = m) - 1(M_{a^*} = m)\}|c] + E(Y_{a^*M_a} - Y_{a^*M_{a^*}}|c). \end{aligned}$$

Under assumption (iv) this is: $E(Y_a - Y_{a^*}|c)$

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

and dividing by $E(Y_{a^*}|c)$ gives:

$$RR_c^{TE} - 1 = \kappa\left[RR_c^{CDE}(m^*) - 1\right] + \kappa RR_c^{INT_{ref}}(m^*) + \kappa RR_c^{INT_{med}} + (RR_c^{PIE} - 1)$$

where $RR_c^{TE} = \frac{E(Y_a|c)}{E(Y_{a^*}|c)}$, $\kappa = \frac{E(Y_{a^*m^*}|c)}{E(Y_{a^*}|c)}$, and

$$\begin{aligned} RR_c^{CDE}(m^*) &= \frac{E(Y_{am^*}|c)}{E(Y_{a^*m^*}|c)} \\ RR_c^{INT_{ref}}(m^*) &= \sum\nolimits_m RERI(a^*, m^*)P(M_{a^*} = m|c) \\ RR_c^{INT_{med}} &= \sum\nolimits_m RERI(a^*, m^*)\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\ RR_c^{PIE} &= \frac{E(Y_{a^*M_a}|c)}{E(Y_{a^*M_{a^*}}|c)} \end{aligned}$$

with $RERI(a^*, m^*) = \left(\frac{E(Y_{am}|c)}{E(Y_{a^*m^*}|c)} - \frac{E(Y_{a^*m}|c)}{E(Y_{a^*m^*}|c)} - \frac{E(Y_{am^*}|c)}{E(Y_{a^*m^*}|c)} + 1\right)$. Under assumptions (i)-(iii) we also have $E(Y_a|c) = E(Y|a, c)$, $E(Y_{am}|c) = \sum_m E[Y|a, m, c]P(m|a, c)$ and thus and $P(M_a = m|c) = P(M = m|a, c)$ and thus the right hand side of the equalities above would be identified from the data. VanderWeele[34] also showed that $\kappa RR_c^{INT_{med}} = \kappa \sum_m RERI(a^*, m^*)\{P(M_a = m|c) - P(M_{a^*} = m|c)\} = \left(\frac{E[Y_{aM_a}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{aM_{a^*}}|c]}{E[Y_{a^*M_{a^*}}|c]} - \frac{E[Y_{a^*M_a}|c]}{E[Y_{a^*M_{a^*}}|c]} + 1\right)$ and called this latter term $RERI_{mediated}$.

Note also under assumption (iv), $(RR_c^{PIE} - 1)$ can be rewritten as

$$\begin{aligned} (RR_c^{PIE} - 1) &= \left(\frac{E(Y_{a^*M_a}|c)}{E(Y_{a^*}|c)} - \frac{E(Y_{a^*}|c)}{E(Y_{a^*}|c)}\right) = \frac{\kappa}{E(Y_{a^*m^*}|c)}\{E(Y_{a^*M_a}|c) - E(Y_{a^*}|c)\} \\ &= \frac{\kappa}{E(Y_{a^*m^*}|c)} \sum\nolimits_m \{E[Y_{a^*m}|c] - E[Y_{a^*m^*}|c]\}\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\ &= \kappa \sum\nolimits_m \left(\frac{E(Y_{a^*m}|c)}{E(Y_{a^*m^*}|c)} - 1\right)\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\ &= \kappa \sum\nolimits_m \frac{E(Y_{a^*m}|c)}{E(Y_{a^*m^*}|c)}\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \end{aligned}$$

The proportion attributable to each of the four components is then obtained by simply dividing each of the four components in the display equation above by their sum as in Table 2. A similar decomposition could likewise be carried out on an additive scale using hazard ratios.

By similar arguments to those above but applied to Propositions 2 and 4, if assumption (iv) did not hold but assumptions (i)-(iii) all did hold, we would have that $(RR_c^{TE} - 1)$ decomposed into

the product of κ and the sum of:

$$RR_c^{CDE}(m^*) - 1 = \frac{E[Y|a, m^*, c]}{E[Y|a^*, m^*, c]} - 1$$

$$\int RERI(a^*, m^*) dP(M_{a^*}|c)$$

$$= \int \{\frac{E[Y|a, m, c]}{E[Y|a^*, m^*, c]} - \frac{E[Y|a^*, m, c]}{E[Y|a^*, m^*, c]} - \frac{E[Y|a, m^*, c]}{E[Y|a^*, m^*, c]} + 1\} dP(m|a^*, c)$$

$$\int RERI(a^*, m^*)\{dP(M_a|c) - dP(M_{a^*}|c)$$

$$= \int \{\frac{E[Y|a, m, c]}{E[Y|a^*, m^*, c]} - \frac{E[Y|a^*, m, c]}{E[Y|a^*, m^*, c]}\}\{dP(m|a, c) - dP(m|a^*, c)\}$$

$$\int \frac{E[Y_{a^*m}|c]}{E[Y_{a^*m^*}|c]}\{dP(M_a|c) - dP(M_{a^*}|c)\} = \int \frac{E[Y|a^*, m, c]}{E[Y|a^*, m^*, c]}\{dP(m|a, c) - dP(m|a^*, c)\}.$$

2.2 Binary Outcome, Continuous Mediator

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

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

with M normally distribution conditional on (A, C) with variance σ^2. Suppose that the outcome is rare so that odds ratios approximate risk ratios. VanderWeele and Vansteelandt[5] derived expressions for the controlled direct effect, the pure indirect effect, and the pure direct effect, all on the risk ratio scale. The total effect, controlled direct effect, and pure indirect effect were given approximately by:

$$\begin{aligned} RR_c^{TE} &\approx \exp[\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + \frac{1}{2}\theta_3^2\sigma^2(a^2 - a^{*2})] \\ RR_c^{CDE}(m^*) &\approx \exp[(\theta_1 + \theta_3 m^*)(a - a^*)] \\ RR_c^{PIE} &\approx \exp[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)] \end{aligned}$$

where the approximations (here and below) hold to the extent that the outcome is rare. We have that $\kappa = \frac{E(Y_{a^*m^*}|c)}{E(Y_{a^*}|c)}$ is given by:

$$\begin{aligned} \kappa &= \frac{E(Y_{a^*m^*}|c)}{E(Y_{a^*}|c)} = \frac{E[Y|a^*, m^*, c]}{\int E[Y|a^*, m, c] dP(m|a^*, c)} \\ &\approx \frac{\exp(\theta_0 + \theta_1 a^* + \theta_2 m^* + \theta_3 a^* m^* + \theta_4' c)}{\exp\{\theta_0 + \theta_1 a^* + \theta_4' c\} \int \exp\{(\theta_2 + \theta_3 a^*)m\} dP(m|a^*, c)} \\ &= \frac{\exp(\theta_2 m^* + \theta_3 a^* m^*)}{\exp\{(\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a^*)^2\sigma^2\}} \\ &= e^{\theta_2 m^* + \theta_3 a^* m^* - (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) - \frac{1}{2}(\theta_2 + \theta_3 a^*)^2\sigma^2}. \end{aligned}$$

We have $\int \frac{E[Y|a,m,c]}{E[Y|a^*,m^*,c]} dP(m|a^\dagger, c)$

$$
\begin{aligned}
&\approx \int \exp(\theta_1 a + \theta_2 m + \theta_3 am - \theta_1 a^* - \theta_2 m^* - \theta_3 a^* m^*) dP(m|a^\dagger, c) \\
&= \exp\{\theta_1(a-a^*) - \theta_2 m^* - \theta_3 a^* m^*\} \int \exp\{(\theta_2 + \theta_3 a)m\} dP(m|a^\dagger, c) \\
&= \exp\{\theta_1(a-a^*) - \theta_2 m^* - \theta_3 a^* m^*\} \exp\{(\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^\dagger + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2 \sigma^2\} \\
&= e^{\theta_1(a-a^*) - \theta_2 m^* - \theta_3 a^* m^* + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^\dagger + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2 \sigma^2}.
\end{aligned}
$$

The reference interaction is thus given by:

$$
\begin{aligned}
RR_c^{INT_{ref}}(m^*) &= \int \{\frac{E[Y|a,m,c]}{E[Y|a^*,m^*,c]} - \frac{E[Y|a^*,m,c]}{E[Y|a^*,m^*,c]} - \frac{E[Y|a,m^*,c]}{E[Y|a^*,m^*,c]} + 1\} dP(m|a^*, c) \\
&= e^{\theta_1(a-a^*) - \theta_2 m^* - \theta_3 a^* m^* + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2 \sigma^2} \\
&\quad - e^{-\theta_2 m^* - \theta_3 a^* m^* + (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a^*)^2 \sigma^2} - e^{(\theta_1 + \theta_3 m^*)(a-a^*)} + 1
\end{aligned}
$$

and the component due to the reference interaction $\kappa RR_c^{INT_{ref}}(m^*)$ by:

$$
\begin{aligned}
&e^{\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 \sigma^2)\}(a-a^*) + \frac{1}{2}\theta_3^2 \sigma^2 (a^2 - a^{*2})} - 1 \\
&- e^{\theta_1(a-a^*) + \theta_2 m^* + \theta_3 a m^* - (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) - \frac{1}{2}(\theta_2 + \theta_3 a^*)^2 \sigma^2} \\
&+ e^{\theta_2 m^* + \theta_3 a^* m^* - (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) - \frac{1}{2}(\theta_2 + \theta_3 a^*)^2 \sigma^2}
\end{aligned}
$$

The mediated interaction is given by:

$$
\begin{aligned}
RR_c^{INT_{med}} &= \int \{\frac{E[Y|a,m,c]}{E[Y|a^*,m^*,c]} - \frac{E[Y|a^*,m,c]}{E[Y|a^*,m^*,c]}\} \{dP(m|a,c) - dP(m|a^*,c)\} \\
&\approx e^{\theta_1(a-a^*) - \theta_2 m^* - \theta_3 a^* m^* + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2 \sigma^2} \\
&\quad - e^{-\theta_2 m^* - \theta_3 a^* m^* + (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a^*)^2 \sigma^2} \\
&\quad - e^{\theta_1(a-a^*) - \theta_2 m^* - \theta_3 a^* m^* + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2 \sigma^2} \\
&\quad + e^{-\theta_2 m^* - \theta_3 a^* m^* + (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a^*)^2 \sigma^2}.
\end{aligned}
$$

and the component due to the mediated interaction $\kappa RR_c^{INT_{med}}$ by:

$$
\begin{aligned}
&e^{\{\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2 \sigma^2)\}(a-a^*) + \frac{1}{2}\theta_3^2 \sigma^2 (a^2 - a^{*2})} \\
&- e^{(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a-a^*)} - e^{\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 \sigma^2)\}(a-a^*) + \frac{1}{2}\theta_3^2 \sigma^2 (a^2 - a^{*2})} + 1.
\end{aligned}
$$

We also have that the component due to controlled direct effect is:

$$
\begin{aligned}
\kappa \left[RR_c^{CDE}(m^*) - 1\right] &= \kappa[e^{(\theta_1 + \theta_3 m^*)(a-a^*)} - 1] \\
&= e^{\theta_1(a-a^*) + \theta_2 m^* + \theta_3 a m^* - (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) - \frac{1}{2}(\theta_2 + \theta_3 a^*)^2 \sigma^2} \\
&\quad - e^{\theta_2 m^* + \theta_3 a^* m^* - (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) - \frac{1}{2}(\theta_2 + \theta_3 a^*)^2 \sigma^2}
\end{aligned}
$$

and the component due to the pure indirect effect is:

$$\begin{aligned}(RR_c^{PIE}-1) &= \kappa\int_m \frac{E(Y_{a^*m}|c)}{E(Y_{a^*m^*}|c)}\{dP(m|a,c)-dP(m|a^*,c)\}\\ &= \kappa\{e^{-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}\\ &\quad -e^{-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}\}\\ &= e^{(\theta_2\beta_1+\theta_3\beta_1 a^*)(a-a^*)}-1.\end{aligned}$$

Standard errors for these various expressions 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.

2.3 Binary Outcome, Binary Mediator

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

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

Valeri and VanderWeele[16] show that the average total effect, controlled direct effect and the average pure indirect effect conditional on $C=c$ are given approximately by:

$$\begin{aligned}RR_c^{TE} &\approx \frac{\exp(\theta_1 a)\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a)\}}{\exp(\theta_1 a^*)\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*)\}}\\ RR_c^{CDE}(m^*) &\approx \exp\{(\theta_1+\theta_3 m)(a-a^*)\}\\ RR_c^{PIE} &\approx \frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^*)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*)\}}\end{aligned}$$

where the approximations (here and below) hold to the extent that the outcome is rare. We have that $\kappa=\frac{E(Y_{a^*m^*}|c)}{E(Y_{a^*}|c)}$ is given by:

$$\begin{aligned}\kappa &= \frac{E(Y_{a^*m^*}|c)}{E(Y_{a^*}|c)}=\frac{E[Y|a^*,m^*,c]}{\int E[Y|a^*,m,c]dP(m|a^*,c)}\\ &\approx \frac{\exp(\theta_0+\theta_1 a^*+\theta_2 m^*+\theta_3 a^* m^*+\theta_4' c)}{\exp\{\theta_0+\theta_1 a^*+\theta_4' c\}\int\exp\{(\theta_2+\theta_3 a^*)m\}dP(m|a^*,c)}\\ &= \frac{\exp(\theta_2 m^*+\theta_3 a^* m^*)}{\frac{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*)}{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)}}\\ &= \frac{e^{\theta_2 m^*+\theta_3 a^* m^*}\{1+e^{\beta_0+\beta_1 a^*+\beta_2' c}\}}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}}.\end{aligned}$$

We also have $\int \frac{E[Y|a,m,c]}{E[Y|a^*,m^*,c]} dP(m|a^\dagger, c)$

$$
\begin{aligned}
&\approx \int \exp(\theta_1 a + \theta_2 m + \theta_3 am - \theta_1 a^* - \theta_2 m^* - \theta_3 a^* m^*) dP(m|a^\dagger, c) \\
&= \exp\{\theta_1(a-a^*) - \theta_2 m^* - \theta_3 a^* m^*\} \int \exp\{(\theta_2 + \theta_3 a)m\} dP(m|a^\dagger, c) \\
&= \frac{e^{\theta_1(a-a^*)-\theta_2 m^* - \theta_3 a^* m^*}}{1 + e^{\beta_0 + \beta_1 a^\dagger + \beta_2' c}} (1 + e^{\beta_0 + \beta_1 a^\dagger + \beta_2' c + \theta_2 + \theta_3 a}) \\
&\quad \frac{e^{\theta_1(a-a^*)-\theta_2 m^* - \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a^\dagger + \beta_2' c + \theta_2 + \theta_3 a})}{1 + e^{\beta_0 + \beta_1 a^\dagger + \beta_2' c}}.
\end{aligned}
$$

The reference interaction is thus given by: $RR_c^{INT_{ref}}(m^*) =$

$$
\begin{aligned}
&\int \{\frac{E[Y|a,m,c]}{E[Y|a^*,m^*,c]} - \frac{E[Y|a^*,m,c]}{E[Y|a^*,m^*,c]} - \frac{E[Y|a,m^*,c]}{E[Y|a^*,m^*,c]} + 1\} dP(m|a^*, c) \\
= &\frac{e^{\theta_1(a-a^*)-\theta_2 m^* - \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c}} - \frac{e^{-\theta_2 m^* - \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c}} \\
&- e^{(\theta_1 + \theta_3 m^*)(a-a^*)} + 1
\end{aligned}
$$

and the component due to the reference interaction $\kappa RR_c^{INT_{ref}}(m^*)$ by:

$$
\begin{aligned}
= &\frac{e^{\theta_1(a-a^*)} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*}} - 1 \\
&- \frac{e^{\theta_1(a-a^*) + \theta_2 m^* + \theta_3 a m^*} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*}} e^{(\theta_1 + \theta_3 m^*)(a-a^*)} + \frac{e^{\theta_2 m^* + \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*}}
\end{aligned}
$$

The mediated interaction is given by: $RR_c^{INT_{med}} =$

$$
\begin{aligned}
&\int \{\frac{E[Y|a,m,c]}{E[Y|a^*,m^*,c]} - \frac{E[Y|a^*,m,c]}{E[Y|a^*,m^*,c]}\} \{dP(m|a,c) - dP(m|a^*,c)\} \\
= &\frac{e^{\theta_1(a-a^*)-\theta_2 m^* - \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a + \beta_2' c + \theta_2 + \theta_3 a})}{1 + e^{\beta_0 + \beta_1 a + \beta_2' c}} - \frac{e^{-\theta_2 m^* - \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a + \beta_2' c + \theta_2 + \theta_3 a^*})}{1 + e^{\beta_0 + \beta_1 a + \beta_2' c}} \\
&- \frac{e^{\theta_1(a-a^*)-\theta_2 m^* - \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c}} + \frac{e^{-\theta_2 m^* - \theta_3 a^* m^*} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c}}
\end{aligned}
$$

and the component due to the mediated interaction $\kappa RR_c^{INT_{med}}$ by:

$$
\begin{aligned}
= &\frac{e^{\theta_1(a-a^*)} (1 + e^{\beta_0 + \beta_1 a + \beta_2' c + \theta_2 + \theta_3 a})(1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c})}{(1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*})(1 + e^{\beta_0 + \beta_1 a + \beta_2' c})} - \frac{(1 + e^{\beta_0 + \beta_1 a + \beta_2' c + \theta_2 + \theta_3 a^*})(1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c})}{(1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*})(1 + e^{\beta_0 + \beta_1 a + \beta_2' c})} \\
&- \frac{e^{\theta_1(a-a^*)} (1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a})}{(1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*})} + 1
\end{aligned}
$$

We also have that the component due to controlled direct effect is:

$$\begin{aligned}\kappa\left[RR_c^{CDE}(m^*)-1\right] &= \kappa[e^{(\theta_1+\theta_3 m^*)(a-a^*)}-1] \\ &= \frac{e^{\theta_1(a-a^*)+\theta_2 m^*+\theta_3 a m^*}(1+e^{\beta_0+\beta_1 a^*+\beta_2' c})}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}} - \frac{e^{\theta_2 m^*+\theta_3 a^* m^*}(1+e^{\beta_0+\beta_1 a^*+\beta_2' c})}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}}\end{aligned}$$

and the component due to the pure indirect effect is:

$$\begin{aligned}&\kappa\int_m \frac{E(Y_{a^* m}|c)}{E(Y_{a^* m^*}|c)}\{dP(m|a,c)-dP(m|a^*,c)\} \\ =\ &\kappa\left(\frac{e^{-\theta_2 m^*-\theta_3 a^* m^*}(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^*})}{1+e^{\beta_0+\beta_1 a+\beta_2' c}} - \frac{e^{-\theta_2 m^*-\theta_3 a^* m^*}(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*})}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c}}\right) \\ =\ &\frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^*)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*)\}}-1.\end{aligned}$$

Standard errors for these expressions could be derived using the delta method along the lines of the derivations in the Online Appendix of Valeri and VanderWeele[16] or by using bootstrapping.

3. SAS Code for the 4-Way Decomposition

3.1. Continuous Outcome, Continuous Mediator

To estimate the components of the 4-way decomposition for the effect of exposure A on a continuous outcome Y with continuous mediator M under the regression models in Section 1.1, one can use the code below. Suppose we have a dataset named 'mydata' with outcome variable 'y', exposure variables 'a' and mediator 'm' and three covariates 'c1', 'c2' and 'c3'. If there were more or fewer covariates the user would have to modify the second, third, fourth, fifth and tenth lines of the code below to include these covariates.

The user must input in the third line of code the two levels of A ('a1=' and 'a0=') that are being compared (these are exposure levels 1 and 0 in the code below but this could be modified for an ordinal or continuous exposure) and the level of $M = m^*$ ('mstar=') at which to compute the controlled direct effect and the remainder of the decomposition (it is assumed in the code below that the mediator is fixed to the value $M = m^* = 0$ but this could be modified). The user must also input in the third line of the code the value of the covariates C at which the effects are to be calculated ('cc1=', 'cc2' and 'cc3='). Alternatively the mean value of these covariates in the sample could be inputted on this line as a summary measure. The code below on line 3 specifies these as 10, 10, and 20 which should be altered according to the covariate values in the application of interest.

The output will include estimates and confidence intervals for the total effect as well as the four components of the total effect, i.e. the controlled direct effect, the reference interaction, the mediated interaction, and the pure indirect effect; the output will also include estimates and confidence intervals for the proportion of the total effect due to each of the four components; and estimates and confidence intervals for the overall proportion mediated, the overall proportion due to interaction, and the overall proportion of the effect that would be eliminated if the mediator M were fixed to the value m^*, specified by the user.


```
proc nlmixed data=mydata;
```

```
parms t0=0 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=0 b1=0 bc1=0 bc2=0 bc3=0 ss_m=1 ss_y=1;
a1=1; a0=0; mstar=0; cc1=10; cc2=10; cc3=20;
mu_y=t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2 + tc3*C3;
mu_m =b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3;
ll_y= -((y-mu_y)**2)/(2*ss_y)-0.5*log(ss_y);
ll_m= -((m-mu_m)**2)/(2*ss_m)-0.5*log(ss_m);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
cde = (t1 + t3*mstar)*(a1-a0);
intref = t3*(b0 + b1*a0 + bcc - mstar)*(a1-a0);
intmed = t3*b1*(a1-a0)*(a1-a0);
pie = (t2*b1 + t3*b1*a0)*(a1-a0);
te = cde + intref + intmed + pie;
estimate 'Total Effect' te;
estimate 'CDE' cde;
estimate 'INTref' intref;
estimate 'INTmed' intmed;
estimate 'PIE' pie;
estimate 'Proportion CDE' cde/te;
estimate 'Proportion INTref' intref/te;
estimate 'Proportion INTmed' intmed/te;
estimate 'Proportion PIE' pie/te;
estimate 'Overall Proportion Mediated' (pie+intmed)/te;
estimate 'Overall Proportion Attributable to Interaction' (intref+intmed)/te;
estimate 'Overall Proportion Eliminated' (intref+intmed+pie)/te;
run;
```

3.2. Continuous Outcome, Binary Mediator

To estimate the components of the 4-way decomposition for the effect of exposure A on a continuous outcome Y with binary mediator M under the regression models in Section 1.2, one can use the code below. The explanation of the code follows that presented in Section 3.1 above.

```
proc nlmixed data=mydata;
parms t0=0 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=1 b1=0 bc1=0 bc2=0 bc3=0 ss_y=1;
a1=1; a0=0; mstar=0; cc1=10; cc2=10; cc3=20;
mu_y=t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2 + tc3*C3;
p_m=(1+exp(-(b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3)))**-1;
ll_y= -((y-mu_y)**2)/(2*ss_y)-0.5*log(ss_y);
ll_m= m*log (p_m)+(1-m)*log(1-p_m);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
cde = (t1 + t3*mstar)*(a1-a0);
intref = t3*(a1-a0)*(exp(b0+b1*a0+bcc)/(1+exp(b0+b1*a0+bcc)) - mstar);
intmed = t3*(a1-a0)*(exp(b0+b1*a1+bcc)/(1+exp(b0+b1*a1+bcc))-exp(b0+b1*a0+bcc)/(1+exp(b0+b1*a0+bcc)));
pie = (t2 + t3*a0)*(exp(b0+b1*a1+bcc)/(1+exp(b0+b1*a1+bcc))-exp(b0+b1*a0+bcc)/(1+exp(b0+b1*a0+bcc)));
te = cde + intref + intmed + pie;
estimate 'Total Effect' te;
estimate 'CDE' cde;
estimate 'INTref' intref;
estimate 'INTmed' intmed;
estimate 'PIE' pie;
estimate 'Proportion CDE' cde/te;
```

```
estimate 'Proportion INTref' intref/te;
estimate 'Proportion INTmed' intmed/te;
estimate 'Proportion PIE' pie/te;
estimate 'Overall Proportion Mediated' (pie+intmed)/te;
estimate 'Overall Proportion Attributable to Interaction' (intref+intmed)/te;
estimate 'Overall Proportion Eliminated' (intref+intmed+pie)/te;
run;
```

3.3. Binary Outcome, Continuous Mediator

To estimate the components of the 4-way decomposition on the ratio scale for the effect of exposure A on a binary outcome Y with continuous mediator M under the regression models in Section 2.2, one can use the code below. Suppose we have a dataset named 'mydata' with outcome variable 'y', exposure variables 'a' and mediator 'm' and three covariates 'c1', 'c2' and 'c3'. If there were more or fewer covariates the user would have to modify the second, third, fourth, fifth and tenth lines of the code below to include these covariates.

The user must input in the third line of code the two levels of A ('a1=' and 'a0=') that are being compared (these are exposure levels 1 and 0 in the code below but this could be modified for an ordinal or continuous exposure) and the level of $M = m^*$ ('mstar=') at which to compute the controlled direct effect and the remainder of the decomposition (it is assumed in the code below that the mediator is fixed to the value $M = m^* = 0$ but this could be modified). The user must also input in the third line of the code the value of the covariates C at which the effects are to be calculated ('cc1=', 'cc2' and 'cc3='). Alternatively the mean value of these covariates in the sample could be inputted on this line as a summary measure. The code below on line 3 specifies these as 58.57, 1.44, and 0.34 which should be altered according to the covariate values in the application of interest.

The output will include estimates and confidence intervals for the total effect risk ratio, the excess relative risk (i.e. the relative risk minus 1) as well as the four components of the excess relative risk, i.e. the excess relative risks due to the controlled direct effect, to the reference interaction, to the mediated interaction, and to the pure indirect effect; the output will also include estimates and confidence intervals for the proportion of the excess relative risk due to each of the four components; and estimates and confidence intervals for the overall proportion mediated, the overall proportion due to interaction, and the overall proportion of the effect that would be eliminated if the mediator M were fixed to the value m^*, specified by the user.

```
proc nlmixed data=mydata;
parms t0=1 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=0 b1=0 bc1=0 bc2=0 bc3=0 ss_m=1;
a1=1; a0=0; mstar=0; cc1=58.57; cc2=1.44; cc3=0.34;
p_y=(1+exp(-(t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2 + tc3*C3)))**-1;
mu_m =b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3;
ll_m= -((m-mu_m)**2)/(2*ss_m)-0.5*log(ss_m);
ll_y= y*log (p_y)+(1-y)*log(1-p_y);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
CDE_comp = exp( t1*(a1-a0)+t2*mstar + t3*a1*mstar - (t2+t3*a0)*(b0+b1*a0+bcc)
   -  (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m )
   - exp(t2*mstar + t3*a0*mstar - (t2+t3*a0)*(b0+b1*a0+bcc) -  (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m );
INTref_comp = exp((t1+t3*(b0+b1*a0+bcc+t2*ss_m))*(a1-a0) + (1/2)*t3*t3*ss_m*(a1*a1-a0*a0)) - (1.0)
   -exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar-(t2+t3*a0)*(b0+b1*a0+bcc)- (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m)
   +exp(t2*mstar+t3*a0*mstar-(t2+t3*a0)*(b0+b1*a0+bcc)- (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m);
```

```
INTmed_comp = exp( (t1+t2*b1+t3*(b0+b1*a0+b1*a1+bcc+t2*ss_m))*(a1-a0)
    + (1/2)*t3*t3*ss_m*(a1*a1-a0*a0) )
   -exp( (t2*b1+t3*b1*a0)*(a1-a0) ) -exp( (t1+t3*(b0+b1*a0+bcc+t2*ss_m ))*(a1-a0)
    + (1/2)*t3*t3*ss_m*(a1*a1-a0*a0) ) + (1);
PIE_comp = exp( (t2*b1+t3*b1*a0)*(a1-a0) ) - (1);
terr=cde_comp+intref_comp+intmed_comp+pie_comp;
total = exp((t1 + t3*(b0+b1*a0+bcc + t2*ss_m))*(a1-a0)+(1/2)*t3*t3*ss_m*(a1*a1-a0*a0))
    *exp((t2*b1+t3*b1*a1)*(a1-a0));
estimate 'Total Effect Risk Ratio' total;
estimate 'Total Excess Relative Risk' total-1;
estimate 'Excess Relative Risk due to CDE' cde_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTref' intref_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTmed' intmed_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to PIE' pie_comp*(total-1)/terr;
estimate 'Proportion CDE' cde_comp/terr;
estimate 'Proportion INTref' intref_comp/terr;
estimate 'Proportion INTmed' intmed_comp/terr;
estimate 'Proportion PIE' pie_comp/terr;
estimate 'Overall Proportion Mediated' (pie_comp+intmed_comp)/terr;
estimate 'Overall Proportion Attributable to Interaction' (intref_comp+intmed_comp)/terr;
estimate 'Overall Proportion Eliminated' (intref_comp+intmed_comp+pie_comp)/terr;
run;
```

The code given above is applicable to cohort data. For case-control studies in which sampling is done on the outcome Y, if the outcome is rare, then the code above can be adapted by fitting the mediator regression only among the controls. This can be done by replacing the sixth line of code by: ll_m= -((m-mu_m)**2)/(2*ss_m)-0.5*log(ss_m)*(1-y);

3.4. Binary Outcome, Binary Mediator

To estimate the components of the 4-way decomposition for the effect of exposure A on a binary outcome Y with binary mediator M under the regression models in Section 2.3, one can use the code below. The explanation of the code follows that presented in Section 3.3 above.

```
proc nlmixed data=mydata;
parms t0=1 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=0 b1=0 bc1=0 bc2=0 bc3=0;
a1=1; a0=0; mstar=0; cc1=58.57; cc2=1.44; cc3=0.34;
p_y=(1+exp(-(t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2 + tc3*C3)))**-1;
p_m =(1+exp(-(b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3)))**-1;
ll_y= y*log (p_y)+(1-y)*log(1-p_y);
ll_m= m*log (p_m)+(1-m)*log(1-p_m);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
CDE_comp = exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar)*(1+exp(b0+b1*a0+bcc))/(1+exp(b0+b1*a0+bcc+t2+t3*a0))
   - exp(t2*mstar+t3*a0*mstar)*(1+exp(b0+b1*a0+bcc))/(1+exp(b0+b1*a0+bcc+t2+t3*a0));
INTref_comp = exp(t1*(a1-a0))*(1+exp(b0+b1*a0+bcc+t2+t3*a1))/(1+exp(b0+b1*a0+bcc+t2+t3*a0)) - (1)
   -exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar)*(1+exp(b0+b1*a0+bcc))*exp((t1+t3*mstar)*(a1-a0))
    /(1+exp(b0+b1*a0+bcc+t2+t3*a0))
   + exp(t2*mstar+t3*a0*mstar)*(1+exp(b0+b1*a0+bcc))/(1+exp(b0+b1*a0+bcc+t2+t3*a0));
INTmed_comp = exp(t1*(a1-a0))*(1+exp(b0+b1*a1+bcc+t2+t3*a1))*(1+exp(b0+b1*a0+bcc))
    /( (1+exp(b0+b1*a0+bcc+t2+t3*a0))*(1+exp(b0+b1*a1+bcc))  )
   - (1+exp(b0+b1*a1+bcc+t2+t3*a0))*(1+exp(b0+b1*a0+bcc)) / ( (1+exp(b0+b1*a0+bcc+t2+t3*a0))
    *(1+exp(b0+b1*a1+bcc))  )
   - exp(t1*(a1-a0))*(1+exp(b0+b1*a0+bcc+t2+t3*a1))/(1+exp(b0+b1*a0+bcc+t2+t3*a0)) + (1);
```

```
PIE_comp = (1+exp(b0+b1*a0+bcc))*(1+exp(b0+b1*a1+bcc+t2+t3*a0)) / ( (1 + exp(b0+b1*a1+bcc))
    *(1+exp(b0+b1*a0+bcc+t2+t3*a0)) ) -(1);
terr=cde_comp+intref_comp+intmed_comp+pie_comp;
total = exp(t1*a1)*(1+exp(b0+b1*a0+bcc))*(1+exp(b0+b1*a1+bcc+t2+t3*a1))
    / ( exp(t1*a0)*(1 + exp(b0+b1*a1+bcc))*(1+exp(b0+b1*a0+bcc+t2+t3*a0)) );
estimate 'Total Effect Risk Ratio' total;
estimate 'Total Excess Relative Risk' total-1;
estimate 'Excess Relative Risk due to CDE' cde_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTref' intref_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTmed' intmed_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to PIE' pie_comp*(total-1)/terr;
estimate 'Proportion CDE' cde_comp/terr;
estimate 'Proportion INTref' intref_comp/terr;
estimate 'Proportion INTmed' intmed_comp/terr;
estimate 'Proportion PIE' pie_comp/terr;
estimate 'Overall Proportion Mediated' (pie_comp+intmed_comp)/terr;
estimate 'Overall Proportion Attributable to Interaction' (intref_comp+intmed_comp)/terr;
estimate 'Overall Proportion Eliminated' (intref_comp+intmed_comp+pie_comp)/terr;
run;
```

The code given above is applicable to cohort data. For case-control studies in which sampling is done on the outcome Y, if the outcome is rare, then the code above can be adapted by fitting the mediator regression only among the controls. This can be done by replacing the sixth line of code by: ll_m= m*log (p_m)+(1-m)*log(1-p_m)*(1-y);

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 4. Although several of the components of the four-way decomposition are not identified in this setting, alternative effects which randomly set M to a value chosen from the distribution of a particular exposure level can be identified. The discussion here will give a randomized interventional interpretation to Proposition 4 in the text and extend that result to settings such as Figure 4 in which there is a mediator-outcome confounder affected by the exposure.

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 $a^* = 0$. As in VanderWeele[34], 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 a randomized interventional analogue of the pure indirect effect. 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; this is a randomized interventional analogue of the pure direct effect. 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. As shown in VanderWeele[34] this third term 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. We have the three-way decomposition given in VanderWeele.[34] Moreover, for the analogue of the pure direct effect we have: $\{E(Y_{aG_{a^*|c}}|c) - E(Y_{a^*G_{a^*|c}}|c)\}$

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

i.e. the analogue of the pure direct effect is the sum of a controlled direct effect and the reference interaction term, $\sum_m E[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]P(M_{a^*} = m|c)$. We thus have a randomized interventional analogue of the four way decomposition.

To identify these effects the following conditions suffice: Assumptions (i) $Y_{am} \perp\!\!\!\perp A|C$ and (iii) $M_a \perp\!\!\!\perp A|C$ above, that conditional on C there is no unmeasured exposure-outcome or exposure-mediator confounding, along with an assumption (ii*) that $Y_{am} \perp\!\!\!\perp M|\{A, C, L\}$, 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 4. Under the three assumptions, each of these component are identified from data and it follows from the g-formula[39] that:

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

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

and

$$\begin{aligned} & \sum_m E[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_{a^*} = m|c)\} \\ = & \sum_{l,m}\{E[Y|a,l,m,c]P(l|a,c) - E[Y|a^*,l,m,c]P(l|a^*,c) - E[Y|a,l,m^*,c]P(l|a,c) \\ & + E[Y|a^*,l,m^*,c]P(l|a^*,c)\}P(m|a^*,c). \end{aligned}$$

Thus a randomized interventional analogue of the four-way decomposition holds and its components can be identified under assumptions (i), (ii*) and (iii). When Figure 3 is in fact the underlying causal diagram so the L can be chosen to be empty then assumption (ii*) simply becomes assumption (ii) in the text. And the identification results here simply reduce to those of Proposition 4 in the text. As in Proposition 4 in the text, the randomized interventional interpretation does not require the more controversial cross-world independence assumption, assumption (iv).