Title
The Causal Mediation Formula – A practitioner guide to the assessment of causal pathways

Author
Pearl, Judea

Publication Date
2011-02-01

Peer reviewed

Submitted.

TECHNICAL REPORT
R-379
February 2011

The Causal Mediation Formula – A practitioner guide to the assessment of causal pathways

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

February 1, 2011

Abstract

Recent advances in causal inference have given rise to a general and easy-to-use estimator for assessing the extent to which the effect of one variable on another is mediated by a third. This estimator, called Mediation Formula, is applicable to nonlinear models with both discrete and continuous variables, and permits the evaluation of path-specific effects with minimal assumptions regarding the data-generating process. We demonstrate the use of the Mediation Formula in simple examples and illustrate why parametric methods of analysis yield distorted results, even when parameters are known precisely. We stress the importance of distinguishing between the necessary and sufficient interpretations of "mediated-effect" and show how to estimate the two components in nonlinear systems with continuous and categorical variables.

Keywords: Effect decomposition, direct and indirect effects, structural equation models, percentage explained

1 Introduction

Consider a randomized clinical trial in which an intervention X shows a significant effect on an outcome Y. A question that invariably comes to investigators' mind is: How and why does the intervention produce the change, or, more specifically, can the effect of X on Y be attributed to a change in some intermediate variable Z standing between the two? The reasons we are concerned with such questions are both scientific and practical. Scientifically, mediation tells us "how nature work" and, practically, it enables us to predict behavior under a rich variety of conditions and interventions. For example, an investigator interested in preventing Y may wish to assess the extent to which Y could be prevented by changing an intermediate variable, Z, standing between X and Y (MacKinnon, 2008, Ch. 2).

For the past few decades the analysis of mediation has been dominated by linear regression paradigms, most notably the one advanced by Baron and Kenny (1986), which can be stated

as follows: To test the contribution of a given mediator Z to the effect of X on Y, first regress Y on X to get the *total* effect, and, then, assess the reduction in this effect when we adjust for (or "condition on" or "control for") Z.

The appeal of this scheme is demonstrated in Fig. 1(a) which shows a linear structural

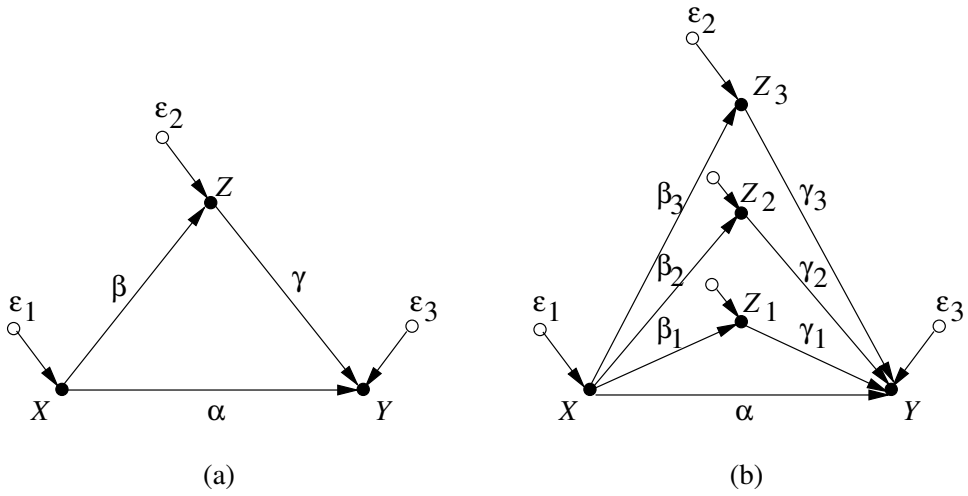

Figure 1: (a) A single mediator Z contributing $\beta \times \gamma$ to the overall effect. (b) Multiple mediators, each contributing $\beta_i \times \gamma_i$.

equation model governing the causal relationships between X, Y, and Z. If the total effect of X on Y through both pathways is $\tau = \alpha + \beta\gamma$, by adjusting for Z, we sever the Z-mediated path and the effect will be reduced to α. The difference between the two regression slopes gives

$$\tau - \alpha = \beta\gamma \tag{1}$$

which is what we expect the z-mediated effect to be.

Alternatively, one can venture to estimate β and γ independently of τ. This is done by first estimating the regression slope of Z on X to get β, then estimating the regression slope of Y on Z controlling for X, which gives us γ; multiplying the two slopes together gives us the mediated effect $\beta\gamma$. The scheme generalizes naturally to multi-path models, as shown in Fig. 1(b) which would represent an opportunity to intervene on four mediating variables, or any subset thereof. The difference between total effect τ and the effect measured after adjusting for mediator Z_i gives the extent to which the indirect path through Z_i contributes to the overall effect, τ. Again, this can be estimated either by the difference-in-coefficients or product-of-coefficients method.

The validity of these two methods depends of course on the assumption that the error terms, ϵ_1, ϵ_2, and ϵ_3, are uncorrelated for, otherwise, some of the structural parameters α, β and γ would not be estimable by regression methods. In randomized trials, where ϵ_1 can be identified with the randomization device, we are assured that ϵ_1 is uncorrelated with ϵ_2 and ϵ_3 and, so, the regressional estimates of α and β will be unbiased. However, randomization does not remove correlations between ϵ_2 and ϵ_3 and, if such exist, adjusting for Z will create spurious correlation between X and Y which will be added to τ and would prevent the

proper estimate of γ. This follows from the fact that "controlling" or "adjusting" for Z in the analysis (by including Z in the regression equation) does not physically disable the paths going through Z, it merely matches samples with equal Z values, and thus induces spurious correlations among other factors in the analysis. Still, regardless of whether the error terms are independent, the difference-in-coefficients and product-of-coefficients methods always yield the same result.[1]

This approach to mediation (often associated with Baron and Kenny) has two major drawbacks. One (mentioned above) is its reliance on the untested assumption of uncorrelated errors, and the second is its reliance on linearity and, in particular, on a property of linear systems called "effect constancy" (or "no interaction"): The effect of one variable on another is independent on the level at which we hold a third. This property does not extend to nonlinear systems; the level at which we control Z would in general modify the effect of X on Y. For example, if the output Y requires both X and Z to be present, then holding Z at zero would disable the effect of X on Y, while holding Z at a high value would enable the latter.

As a consequence, additions and multiplications are not self-evident in nonlinear systems. It would not be appropriate, for example, to estimate the indirect effect by subtracting the direct effect from the total – the relation between the three need not be additive. Nor will it be appropriate to multiply the effect of X on Z by that of Z on Y (keeping X at some level) – multiplicative compositions demand their justifications. Indeed, all attempts to define mediation by generalizing the difference and product strategies to nonlinear system have resulted in distorted and irreconcilable results (MacKinnon et al., 2007a,b; Pearl, 2010b).

This paper removes these nonlinear barriers and avails mediation analysis to a large space of new applications, especially those involving categorical data and highly nonlinear processes. The first limitation, the requirement of error independence (or "no unmeasured confounders," as it is often called) will remain intact, and should be kept in mind throughout our discussion.[2] Our focus in the sequel will be on crossing the linear-to-nonlinear barrier, using the same causal assumptions that support the standard linear analysis of Baron and Kenny (1986).

2 Total, direct and indirect effects

Consider the nonlinear version of the mediation model, as depicted in Fig. 2. In the most

[1] It is important to note that the equality $\tau - \alpha = \beta\gamma$ expressed in (1) is a universal identity among regressional coefficients of any three variables, and has nothing to do with causation or mediation. It will continue to hold regardless of whether confounders are present, whether the structural parameters are identifiable, whether the underlying model is linear or nonlinear and regardless of whether the arrows in the model of Fig. 1(a) point in the right direction. Moreover, the equality will hold among the OLS estimates of these parameters, regardless of sample size. Therefore, the failure of certain parameters in nonlinear regression to obey similar equalities should not be construed as an indication of faulty standardization, as suspected by MacKinnon et al. (2007a,b).

[2] We should mention here that the management of confounding has gone through a major development in the past decade, in both linear and nonparametric models, and a complete set of techniques is now available for neutralizing error correlations, whenever possible, both by covariate adjustment and through the use of instrumental variables (Pearl, 2009). These techniques are applicable to the analysis of mediation.

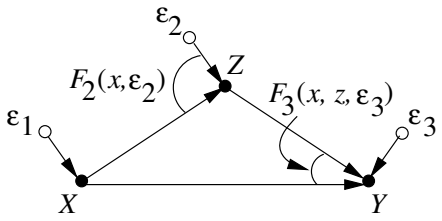

Figure 2: A generic model depicting mediation through Z with no confounders.

general case, the corresponding structural equations would have the form:

$$\begin{aligned} x &= F_1(\epsilon_1) \\ z &= F_2(x, \epsilon_2) \\ y &= F_3(x, z, \epsilon_3) \end{aligned} \tag{2}$$

where X, Y, Z are discrete or continuous random variables, F_1, F_2, and F_3 are arbitrary functions, and $\epsilon_1, \epsilon_2, \epsilon_3$ represent omitted factors which are assumed to be mutually independent yet arbitrarily distributed. Since the functions F_1, F_2, and F_3 are unknown to investigators, mediation analysis commences by first defining total, direct and indirect effects in terms of those functions and, then, expressing them in terms of the available data, which we assume is given in the form of random samples (x, y, z) drawn from the joint distribution $P(x, y, z)$.

2.1 Total Effect

Among the three types of effects consider here, the easiest to define and estimate is the *total effect*, which measures the change in Y produced by a unit change in X, say from $X = 0$ to $X = 1$. The status of Z need not be specified in this definition, since Z is allowed to track the changes in X and, so, we have for the total effect:

$$Y(X = 1) - Y(X = 0) = F_3[1, F_2(1, \epsilon_2), \epsilon_3] - F_3[0, F_2(0, \epsilon_2), \epsilon_3]$$

At the population level, we will define the total effect TE to be the expectation of the difference above taken over ϵ_2 and ϵ_3, which (assuming independent errors) gives:

$$TE = E(Y|X = 1) - E(Y|X = 0). \tag{3}$$

This difference is none other but the regression slope of Y on X, commonly estimated by OLS. More generally, however, if we are interested in the total effect of a transition from $X = x$ to $X = x'$, where x and x' are any two levels of X (say two dosage levels of a drug), we write:

$$TE_{x,x'} = E(Y|X = x') - E(Y|X = x). \tag{4}$$

Clearly, in nonlinear systems, both the baseline $X = x$ and the endpoint $X = x'$ may play a role in affecting the change of Y.

2.2 Direct Effects

The idea of estimating the direct effect of X on Y by controlling for Z is applicable to nonlinear models as well since, assuming ϵ_2 and ϵ_3 are independent, conditioning on Z simulates the physical action of "fixing" or "setting" Z at a constant value, z, thus preventing X from transmitting its change along the mediating path $X \rightarrow Z \rightarrow Y$. The resulting estimator is called the "controlled direct effect" (Robins and Greenland, 1992; Pearl, 2001):

$$CDE = E(Y|X = 1, Z = z) - E(Y|X = 0, Z = z) \tag{5}$$

which is the regression slope of Y on X keeping Z constant at z.

However, the question arises: at what value should we set Z? As remarked earlier, different settings of Z would yield different results. For example, assume that X stands of a drug taken to cure a disease Y. As a side effect, X also stimulates the secretion of an enzyme Z that hastens the process through which the drug acts on the disease. If we fix Z at a high level, the drug will appear highly efficacious, while if we fix Z at a low level, the drug will have only a meager effect. The question remains therefore, at what value of Z should we conduct our analysis if we wish to evaluate the direct effect of the drug on the disease, unmediated by Z?

Moreover, by conditioning on $(X = 1, Z = 0)$ and $(X = 0, Z = 0)$ respectively, as instructed by Eq. (5), we are comparing different types of subjects. The former represents subjects who end up with low Z despite taking the drug, while the latter represents subjects who have low Z before taking the drug. Taking the difference in $E(Y)$ in these two subpopulations does not capture the idea of measuring the direct effect of X on Y while holding Z constant for every individual.

For this reason, it is more meaningful to define

a notion of direct effect that does not require setting Z uniformly over the population, but let it vary from individual to individual. This notion, denoted $DE_{x,x'}(Y)$ is defined as the expected change in Y induced by changing X from x to x' while keeping all mediating factors constant at whatever value they *obtained* before the transition from x to x' (Robins and Greenland, 1992; Pearl, 2001).[3] This definition of direct-effect invokes the phrase: "at whatever value they obtained" which is counterfactual; there is no way to rerun history and measure subjects response under conditions they have not actually experienced. Pearl (2001) showed however that, for the confounding-free model of Fig. 2, the natural direct effect can be estimated from population data and is given by:

$$DE_{x,x'}(Y) = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \tag{6}$$

The intuition is simple, the natural direct effect is the weighted average of the controlled direct effect, using the pre-intervention distribution $P(z|x)$ as a weighing function. Equation (6) can easily be estimated by a two-step regression, as will be shown in the sequel.

[3]Robins and Greenland (1992) called this notion of direct effect "Pure" while Pearl called it "Natural," denoted NDE, to be distinguished from the "controlled direct effect" (Eq. 5) which is specific to one level of the mediator Z. We will delete the letter "N" from the acronyms of both the direct and indirect effect, and use DE and IE, respectively.

2.3 Indirect Effects

Remarkably, the counterfactual definition of the direct effect can be turned around and provide an operational definition for the *indirect effect*—a concept shrouded in mystery and controversy, because it is impossible, by controlling any of the variables in the model, to selectively disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The *indirect effect*, IE, of the transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z (for each individual) to whatever value it would have attained had X been set to $X = x'$. Going through the counterfactual algebra of this nested expression, Pearl (2001) showed that, for the confounding-free model of Fig. 2, the indirect effect can also be reduced to an estimable expression, given by:

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)]. \tag{7}$$

The intuition here is quite different and unveils a nonlinear version of the product-of-coefficients strategy. The term $E(Y|x,z)$ plays the role of γ in Fig. 1(a), for it captures the effect of Z on Y for fixed x, and the difference $P(z|x') - P(z|x)$ plays the role of β, for it captures the impact of the transition from x to x' on the probability of Z. We see that what was a simple product operation in linear systems is here replaced by a composition operator that involves summation over all values of Z.

Equation (7) provides a general formula for mediation effects, applicable to any nonlinear system, any distribution, and any type of variables. Moreover, the formula is readily estimable by regression. Owing to its generality and ubiquity, I have referred to this expression as the "Mediation Formula" (Pearl, 2009, 2010b).

Not surprising, owed to the nonlinear nature of the model, the relationship between the total, direct and indirect effects is non-additive. Indeed, the total effect TE of a transition has been shown to be the *difference* (not the *sum*) between the direct effect and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{8}$$

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{9}$$

Since each term above is based on an independent operational definition, this equality constitutes a formal justification for the additive formula used routinely in linear systems.

3 The Mediation Formula: A Simple Solution to a Thorny Problem

This subsection demonstrates how the Mediation Formula of Eq. (7) can be applied in assessing mediation effects in nonlinear models. We will use the standard mediation model of

Fig. 2, where all error terms are assumed to be mutually independent, with the understanding that adjustment for appropriate sets of covariates W may be necessary to achieve this independence, that Z may represent a vector of variables, and that integrals should replace summations when dealing with continuous variables (Imai et al., 2010).

The Mediation Formula represents the average increase in the outcome Y that the transition from $X = x$ to $X = x'$ is expected to produce absent any direct effect of X on Y. Though based on solid causal principles, it embodies no causal assumption other than the generic mediation structure of Fig. 2. When the outcome Y is binary (e.g., recovery, or hiring) the ratio $(1 - IE/TE)$ represents the fraction of responding individuals who owe their response to direct paths, while $(1 - DE/TE)$ represents the fraction who owe their response to Z-mediated paths.

3.1 Estimating mediation effects:

The Mediation Formula tells us that IE depends only on the conditional expectation of Y, not on its distribution. It calls therefore for a two-step regression which, in principle, can be performed nonparametrically. In the first step we regress Y on X and Z, and obtain the estimate

$$g(x, z) = E(Y|x, z) \tag{10}$$

for every (x, z) cell. In the second step we fix x and regard $g(x, z)$ as a function $g_x(z)$ of Z. We now estimate the conditional expectation of $g_x(z)$, conditional on $X = x'$ and $X = x$, respectively, and take the difference

$$IE_{x,x'}(Y) = E_{Z|X}[g_x(z)|x'] - E_{Z|X}[g_x(z)|x]. \tag{11}$$

Nonparametric estimation is not always practical. When Z consists of a vector of several mediators, the dimensionality of the problem might prohibit the estimation of $E(Y|x, z)$ for every (x, z) cell, and the need arises to use parametric approximation. We can then choose any convenient parametric form for $E(Y|x, z)$ (e.g., linear, logit, probit), estimate the parameters separately (e.g., by regression or maximum likelihood methods), insert the parametric approximation into (7) and estimate its two conditional expectations (over z) to get the mediated effect (VanderWeele, 2009).

3.2 The linear case

Let us examine what the Mediation Formula yields when applied to the linear version of our model, shown in Fig. 1(a):

$$\begin{aligned} x &= a_0 + \epsilon_1 \\ z &= b_0 + \beta x + \epsilon_2 \\ y &= c_0 + \alpha x + \gamma z + \epsilon_3 \end{aligned} \tag{12}$$

with ϵ_1, ϵ_2, and ϵ_3 uncorrelated, zero-mean error terms and a_0, b_0, c_0 the corresponding regression intercepts. Computing the conditional expectation in (7) gives

$$E(Y|x, z) = c_0 + \alpha x + \gamma z$$

and yields

$$\begin{aligned} IE_{x,x'}(Y) &= \sum_z (c_0 + \alpha x + \gamma z)[P(z|x') - P(z|x)] \\ &= \gamma[E(Z|x') - E(Z|x)] \qquad (13) \\ &= (x' - x)(\beta\gamma) \qquad (14) \\ &= (x' - x)(\tau - \alpha) \qquad (15) \end{aligned}$$

where τ is the slope of the total effect;

$$\tau = (E(Y|x') - E(Y|x))/(x' - x) = \alpha + \beta\gamma.$$

We thus obtained the standard expressions for indirect effects in linear systems, which can be estimated either as a difference $\tau - \alpha$ of two regression coefficients (equation 15) or as a product $\beta\gamma$ of two regression coefficients (equation 14) (see MacKinnon et al., 2007b). These two strategies do not generalize to nonlinear systems (Pearl, 2010a) as will be shown next.

3.3 Linear models with interaction

To understand the difficulty, assume that the correct model behind the data contains a product term xz in the equation for y:

$$y = c_0 + \alpha x + \gamma z + \delta xz + \epsilon_3,$$

a nonlinear model explored by many researchers (Jo, 2008; Kraemer et al., 2008; MacKinnon, 2008). Further assume that we correctly account for this added term and, through elaborate regression analysis, we obtain accurate estimates of all parameters in this model. It is still not clear what combinations of parameters measure the direct and indirect effects of X on Y, or, more specifically, how to assess the fraction of the total effect that is *explained* by mediation and the fraction that is *owed* to mediation. In linear analysis, the former fraction is captured by the product $\beta\gamma/\tau$ (Eq. 14), the latter by the difference $(\tau - \alpha)/\tau$ (Eq. 15) and the two quantities coincide. In the presence of interaction, however, each fraction demands a separate analysis, as dictated by the Mediation Formula.

To witness, substituting the nonlinear equation in (4), (6) and (7) and assuming $x = 0$ and $x' = 1$, yields the following decomposition:

$$\begin{aligned} DE &= \alpha + b_0\delta \\ IE &= \beta\gamma \\ TE &= \alpha + b_0\delta + \beta(\gamma + \delta) \\ &= DE + IE + \beta\delta \end{aligned}$$

We therefore conclude that the portion of output change for which mediation would be *sufficient* is

$$IE = \beta\delta$$

while the portion for which mediation would be *necessary* is

$$TE - DE = \beta\gamma + \beta\delta$$

We note that, due to interaction, a direct effect can be sustained even when the parameter α vanishes and, moreover, a total effect can be sustained even when both the direct and indirect effects vanish. This illustrates that estimating parameters in isolation tells us little about the effect of mediation and, more generally, mediation and moderation are intertwined and cannot be assessed separately.

If the policy evaluated aims to prevent the outcome Y by ways of weakening the mediating pathways, the target of analysis should be the difference $TE - DE$, which measures the highest prevention potential of any such policy. If, on the other hand, the policy aims to prevent the outcome by weakening the direct pathway, the target of analysis should shift to IE, for $TE - IE$ measures the highest preventive potential of this type of policy.

3.4 The binary case

The main power of the Mediation Formula shines in studies involving categorical variables, especially when we have no parametric model of the data generating process. To illustrate, consider the case where all variables are binary, still allowing for arbitrary interactions and arbitrary distributions of all processes. The low dimensionality of the binary case permits both a nonparametric solution and an explicit demonstration of how mediation can be estimated directly from the data. Generalizations to multi-valued variables are straightforward.

Assume that the model of Fig. 2 is valid and that the observed data is given by Table 1. The factors $E(Y|x,z) = g_{xz}$ and $E(Z|x) = h_x$, needed for (6), (7), and (10), can be readily

Number of Samples	X	Z	Y	$E(Y\|x,z) = \boldsymbol{g_{xz}}$	$E(Z\|x) = \boldsymbol{h_x}$
n_1	0	0	0	$\frac{n_2}{n_1+n_2} = g_{00}$	$\frac{n_3+n_4}{n_1+n_2+n_3+n_4} = h_0$
n_2	0	0	1		
n_3	0	1	0	$\frac{n_4}{n_3+n_4} = g_{01}$	
n_4	0	1	1		
n_5	1	0	0	$\frac{n_6}{n_5+n_6} = g_{10}$	$\frac{n_7+n_8}{n_5+n_6+n_7+n_8} = h_1$
n_6	1	0	1		
n_7	1	1	0	$\frac{n_8}{n_7+n_8} = g_{11}$	
n_8	1	1	1		

Table 1: Computing the Mediation Formula for the model in Fig. 2, with X, Y, Z binary.

estimated as shown in the two right-most columns of Table 1 and, when substituted in (6), (9), (7), yield

$$DE = (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0 \tag{16}$$

$$IE = (h_1 - h_0)(g_{01} - g_{00}) \tag{17}$$

$$TE = g_{11}h_1 + g_{10}(1 - h_1) - [g_{01}h_0 + g_{00}(1 - h_0)] \tag{18}$$

We see that logistic or probit regression is not necessary; simple arithmetic operations suffice to provide a general solution for any conceivable data set, regardless of the data-generating process.

3.5 Numerical example

To anchor these formulas in a concrete example, let us assume that $X = 1$ stands for a drug treatment, $Y = 1$ for recovery, and $Z = 1$ for the presence of a certain enzyme in a patient's blood which appears to be stimulated by the treatment. Assume further that the data described in Tables 2 and 3 was obtained in a randomized clinical trial and that our research question is whether Z mediates the action of X on Y, or is merely a catalyst that accelerates the action of X on Y.

Treatment X	Enzyme present Z	Percentage cured $g_{xz} = E(Y\|x, z)$
YES	YES	$g_{11} = 80\%$
YES	NO	$g_{10} = 40\%$
NO	YES	$g_{01} = 30\%$
NO	NO	$g_{00} = 20\%$

Table 2:

Treatment X	Percentage with Z present
NO	$h_0 = 40\%$
YES	$h_1 = 75\%$

Table 3:

Substituting this data into Eqs. (16)–(18) yields:

$$DE = (0.40 - 0.20)(1 - 0.40) + (0.80 - 0.30)0.40 = 0.32$$
$$IE = (0.75 - 0.40)(0.30 - 0.20) = 0.035$$
$$TE = 0.80 \times 0.75 + 0.40 \times 0.25 - (0.30 \times 0.40 + 0.20 \times 0.10) = 0.46$$
$$IE/TE = 0.07 \qquad DE/TE = 0.696 \qquad 1 - DE/TE = 0.304$$

We conclude that 30.4% of those recovered owe their recovery to the capacity of the treatment to stimulate the secretion of the enzyme, while only 7% of recoveries would be sustained by enzyme stimulation alone. The enzyme seems to act more as a catalyst for the healing process of X than having a healing action of its own. The policy implication of such a study would be that efforts to substitute the drug with an alternative stimulant of the enzyme are not likely to be effective, the drug evidently has a beneficial effect on recovery that is independent of, though enhanced by enzyme stimulation.

4 Relations to Other Approaches

In comparing these results to those produced by conventional mediation analyses we should note that conventional methods do not define direct and indirect effects in a setting where the underlying process is unknown. MacKinnon (2008, Ch. 11), for example, analyzes categorical data using logistic and probit regressions and constructs effect measures using products and differences of the parameters in those regressional forms. This strategy is not compatible with the causal interpretation of effect measures, even when the parameters are precisely known; IE and DE may be extremely complicated functions of those regression coefficients (Pearl, 2010b). Fortunately, those coefficients need not be estimated at all; effect measures can be estimated directly from the data, circumventing the parametric analysis altogether, as shown in Eq. (16)–(18).

Attempts to extend the difference and product heuristics to nonparametric analysis have encountered ambiguities that conventional analysis fails to resolve.

The product-of-coefficients heuristic advises us to multiply the unit effect of X on Z

$$C_\beta = E(Z|X=1) - E(Z|X=0) = h_1 - h_0$$

by the unit effect of Z on Y given X,

$$C_\gamma = E(Y|X=x, Z=1) - E(Y|X=x, Z=0) = g_{x1} - g_{x0}$$

but does not specify on what value we should condition X. Equation (17) resolves this ambiguity by determining that C_γ should be conditioned on $X=0$; only then would. the product $C_\beta C_\gamma$ yield the correct mediation measure, IE.

The difference-in-coefficients heuristics instructs us to estimate the direct effect coefficient

$$C_\alpha = E(Y|X=1, Z=z) - E(Y|X=0, Z=z) = g_{1z} - g_{0z}$$

and subtract it from the total effect, but does not specify on what value we should condition Z. Equation (16) determines that the correct way of estimating C_α would be to condition on both $Z=0$ and $Z=1$ and take their weighted average, with $h_0 = P(Z=1|X=0)$ serving as the weighting function.

To summarize, the Mediation Formula dictates that, in calculating IE, we should condition on both $Z=1$ and $Z=0$ and average while, in calculating DE, we should condition on only one value, $X=0$, and no average need be taken.

The difference and product heuristics are both legitimate, with each seeking a different effect measure. The difference-in-coefficients heuristics, leading to $TE - DE$, seeks to measure the percentage of units for which mediation was *necessary*. The product-of-coefficients heuristics on the other hand, leading to IE, seeks to estimate the percentage of units for which mediation was *sufficient*. The former informs policies aiming to modify the direct pathway while the latter informs those aiming to modify mediating pathways.

In addition to providing causally sound estimates for mediation effects, the Mediation Formula also enables researchers to evaluate analytically the effectiveness of various parametric specifications relative to any assumed model. This type of analytical "sensitivity analysis" has been used extensively in statistics for parameter estimation but could not be

applied to mediation analysis, owing to the absence of an objective target quantity that captures the notion of indirect effect in both linear and nonlinear systems, free of parametric assumptions. The Mediation Formula of Eq. (7) explicates this target quantity formally, and casts it in terms of estimable quantities. It has been used by Imai et al. (2010) to examine the robustness of empirical findings to the possible existence of unmeasured confounders.

The derivation of the Mediation Formula was facilitated by taking seriously the graphical-counterfactual-structural symbiosis spawned by the structural interpretation of counterfactuals (Pearl, 2009, Ch. 7). In contrast, when the mediation problem is approached from an strict potential-outcome viewpoint, void of structural guidance, counterintuitive definitions ensue, carrying the label "principal strata" (Rubin, 2004, 2005), which are at variance with common understanding of direct and indirect effects (VanderWeele, 2008; Joffe et al., 2007). For example, the direct effect is definable only in units absent of indirect effects. This means that a grandfather would be deemed to have no direct effect on his grandson's behavior in families where he has had some effect on the father. This precludes from the analysis all typical families, in which a father and a grandfather have simultaneous, complementary influences on children's upbringing. In linear systems, to take an even sharper example, the direct effect would be undefined whenever indirect paths exist from the cause to its effect. The emergence of such paradoxical conclusions underscores the wisdom, if not necessity of a symbiotic analysis, in which counterfactuals are governed by their structural definition.[4]

Conclusions

Traditional methods of mediation analysis have been limited to linear models or semi-linear regression models, and have produced distorted estimates of "mediation effects" when applied to nonlinear models, or models with categorical variables. This paper offers a causally sound alternative that ensures bias-free estimates while making no assumption on the distributional form of the underlying process.

We distinguished between proportion of response cases for which mediation was *necessary* and those for which mediation would have been *sufficient*. Both measures play a role in mediation analysis, and are given here a formal representation through the Mediation Formula. This formula is estimable by simple regression and provides an objective measure of the extent to which an effect is mediated through a given mediating path, independent of the method chosen for estimating that effect. While the validity of the formulas rests on the same assumptions that are required for standard linear analysis, their general appeal to nonlinear systems, continuous and categorical variables, and arbitrary complex interactions render them a powerful tool for the assessment of causal pathways in many of the health related sciences.

[4]Such symbiosis is now standard in epidemiology research (Robins, 2001; Petersen et al., 2006; VanderWeele and Robins, 2007; Hafeman and Schwartz, 2009; Joffe and Green, 2009; VanderWeele, 2009; Kaufman, 2010) and is making its way slowly toward the social and behavioral sciences (Morgan and Winship, 2007; Imai et al., 2010).

Acknowledgments

This paper has benefited from discussions with Elias Barenboim, Peter Bentler, Ken Bollen, Jeffrey Hoyle, Booil Jo, Marshall Joffe, David Kaplan, David Kenny, Helena Kraemer, David MacKinnon, Rod McDonald, Steven Sussman, and Leland Wilkinson, and was supported in parts by grants from NIH #1R01 LM009961-01, NSF #IIS-0914211 and #IIS-1018922, and ONR #N000-14-09-1-0665 and #N00014-10-1-0933.

References

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51** 1173–1182.

Hafeman, D. and Schwartz, S. (2009). Opening the black box: A motivation for the assessment of mediation. *International Journal of Epidemiology* **3** 838–845.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* **25** 51–71.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods* **13** 314–336.

Joffe, M. and Green, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* 530–538.

Joffe, M., Small, D. and Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science* **22** 74–97.

Kaufman, J. (2010). Invited commentary: Decomposing with a lot of supposing. *American Journal of Epidemiology* **172** 1349–1351.

Kraemer, H., Kiernan, M., Essex, M. and Kupfer, D. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology* **27** S101–S108.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, New York.

MacKinnon, D., Fairchild, A. and Fritz, M. (2007a). Mediation analysis. *Annual Review of Psychology* **58** 593–614.

MacKinnon, D., Lockwood, C., Brown, C., Wang, W. and Hoffman, J. (2007b). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials* **4** 499–513.

Morgan, S. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. Cambridge University Press, New York, NY.

Pearl, J. (2001). Direct and indirect effects. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2010a). An introduction to causal inference. *The International Journal of Biostatistics* **6** DOI: 10.2202/1557–4679.1203, <http://www.bepress.com/ijb/vol6/iss2/7/>.

Pearl, J. (2010b). The mediation formula: A guide to the assessment of causal pathways in non-linear models. Tech. Rep. R-363, <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Petersen, M., Sinisi, S. and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology* **17** 276–284.

Robins, J. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12** 313–320.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31** 161–170.

Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322–331.

VanderWeele, T. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters* **78** 2957–2962.

VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20** 18–26.

VanderWeele, T. and Robins, J. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology* **18** 561–568.