Forthcoming, Preventive Science.

TECHNICAL REPORT
R-379
April 2011

The Causal Mediation Formula – A Guide to the Assessment of Pathways and Mechanisms

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

April 28, 2011

Abstract

Recent advances in causal inference have given rise to a general and easy-to-use formula for assessing the extent to which the effect of one variable on another is mediated by a third. This so-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' minds is: How and why does the intervention produce the effect, 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 works" 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, or an intermediate process between X and Z (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 and estimate the regression coefficient R_{YX}, to be equated with the *total effect*. Second, include Z in the regression and estimate the partial regression coefficient $R_{YX \cdot Z}$ when Z is "controlled for" (or "conditioned on" or "adjusted for"). The difference between the two slopes, $R_{YX} - R_{YX \cdot Z}$, would then measure the reduction in the total effect due to controlling for Z and should quantify the effect mediated through Z.

The intuition behind this scheme is demonstrated in Fig. 1(a) which shows a linear

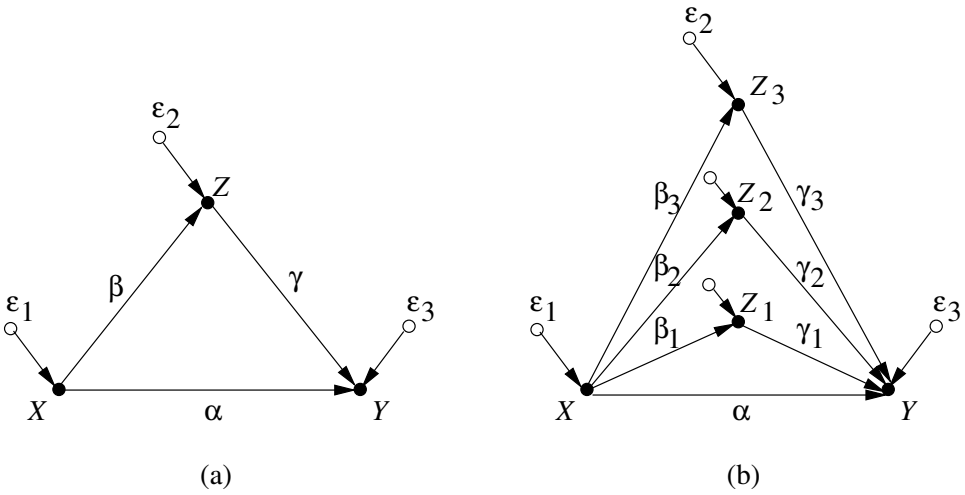

Figure 1: (a) A single mediator Z contributing $\beta \times \gamma$ to the overall effect. (b) Multiple mediators, each contributing $\beta_i \times \gamma_i$. The error terms, $\epsilon_1, \epsilon_2, \epsilon_3$ represent factors omitted from the analysis.

structural 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

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

and the product $\beta\gamma$ 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 represents an opportunity to intervene on three 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 and both methods would produce biased results. 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 τ 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 or α. In other words, the regression coefficient $R_{YZ \cdot X}$ would no longer equal γ and the difference $R_{YX} - R_{ZX}R_{YZ \cdot X}$ would no longer equal α. 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 (see Pearl 1998; Cole and Hernán 2002; VanderWeele and Vansteelandt 2009). Such correlations cannot be detected by statistical means and, so, regardless of whether the error terms are independent, the difference-in-coefficients and product-of-coefficients methods always yield the same (biased) result.

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 may not be appropriate, for example, to define the indirect effect in terms of the "difference" in the total effect, with and without control. Nor would 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, 2011a).

This paper describes a recently developed method that 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.[1] Our focus in the sequel however 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).

[1] 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 dependencies, whenever possible, both by covariate adjustment and through the use of instrumental variables (Pearl, 2009; Tian and Shpitser, 2010). These techniques are directly applicable to the analysis of mediations (Pearl, 2011a; Shpitser and VanderWeele, 2011), but are beyond the scope of this paper.

2 Total, direct and indirect effects

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

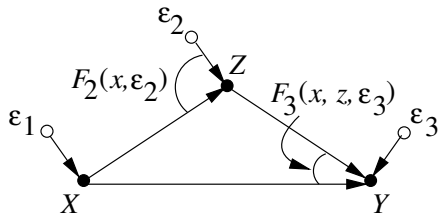

Figure 2: A generic model depicting mediation through Z with no confounders.

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 considered here, the easiest to define and estimate is the *total effect*, $TE_{0,1}$ 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:

$$TE_{0,1}(\epsilon_2, \epsilon_3) = F_3[1, F_2(1, \epsilon_2), \epsilon_3] - F_3[0, F_2(0, \epsilon_2), \epsilon_3]$$

At the population level, we will define the total effect $TE_{0,1}$ to be the expectation of the difference above taken over ϵ_2 and ϵ_3, which (assuming independent errors) gives:

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

regardless of the functional form of F_2 and F_3 (see, for example, Pearl (2009, p. 72)). 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 Controlled and natural 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 $X \to Z \to Y$. The resulting estimand is called the "controlled direct effect" (Robins and Greenland, 1992; Pearl, 2001):

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

which is the regression slope of Y on X keeping Z constant at z.[2]

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 for 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?

One can report, of course, the value of $CDE(z)$ for each level $Z = z$, and let the user choose the value that matches the intervention policy under consideration. In many cases, however, the policy informed by the direct effect is not one where Z is set to a uniform level for all units in the population but, rather, one where the sensitivity of Z to X is suppressed or enhanced, not Z itself. Taking the enzyme example above, a policy maker may be interested in the benefit of developing a cheaper drug, identical to the one studied, but lacking the potential to stimulate enzyme secretion. Absent the mediating effect of Z, the the efficacy of the new drug will be determined by whatever level Z attains naturally in the population, varying from individual to individual, not set uniformly by external control.

Under such settings, it is more meaningful to define a notion of direct effect called *natural*, that does not require setting Z uniformly over the population, but lets it vary from individual to individual. This notion, denoted $NDE_{x,x'}(Y)$ 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 *would have obtained* under $X = x$, before the transition from x to x' (Robins and Greenland, 1992; Pearl, 2001).[3] This definition of direct-effect invokes the phrase: "at

[2]The general causal expression for $CDE(z)$, which does not assume error-independence is given by:

$$CDE(z) = E[Y|do(X = 1, Z = z)] - E[Y|do(X = 0, Z = z)]$$

(see (Pearl, 2009, p. 127)) or, using the structural equations of Eq. (2),

$$CDE(z) = E[F_3(1, z, \epsilon_3)] - E[F_3(0, z, \epsilon_3)]$$

A necessary and sufficient condition for estimating $CDE(z)$ in observational studies (in the presence of unobserved confounders) is given in Tian and Shpitser (2010).

[3]Using the structural model of Eq. (2), the formal definition of the natural direct effect reads:

$$NDE_{x,x'}(Y) = E[F_3(x', F_2(x, \epsilon_2), \epsilon_3)] - E[F_3(x, F_2(x, \epsilon_2), \epsilon_3)]$$

Robins and Greenland (1992) called this notion of direct effect "Pure" while Pearl called it "Natural," to

whatever value they would have obtained" which is counterfactual; there is no way to rerun history and measure subjects response under conditions they have not actually experienced. It has been shown nevertheless (Pearl, 2001) that, for the confounding-free model of Fig. 2, the natural direct effect can be estimated from population data[4] and is given by:

$$NDE_{x,x'}(Y) = \sum_z [E(Y|X = x', Z = z) - E(Y|X = x, Z = z)]P(Z = z|X = x)$$

or, using a short-hand notation, we write:

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

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

2.3 Indirect effects

Remarkably, the counterfactual definition of the natural 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. Thus, whereas in formalizing the notion of "direct effect" one has a choice between the controlled and natural interpretations (Section 2.2), the indirect effect has no "controlled" interpretation.[5]

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 $X = x'$. Formally, this counterfactual definition reads:

$$IE_{x,x'}(Y) = E[F_3(x, F_2(x', \epsilon_2), \epsilon_3)] - E[F_3(x, F_2(x, \epsilon_2), \epsilon_3)] \tag{7}$$

which is similar to the definition of the natural direct effect (footnote 3) save for exchanging x with x' in the first term.

Assuming again the confounding-free model of Fig. 2, the indirect effect defined in (7) can be reduced to an estimable expression (Pearl, 2001), given by :

$$IE_{x,x'}(Y) = \sum_z E(Y|x, z)[P(z|x') - P(z|x)]. \tag{8}$$

---

stress the natural, unperturbed distribution of values, $Z = F_2(x, \epsilon_2)$ at which we "freeze" Z while changing X from $X = x$ to $X = x'$. For discussions regarding policy implications of NDE versus CDE, see (Pearl, 2001; Robins, 2003; Joffe et al., 2007; Hafeman and Schwartz, 2009; Pearl, 2009, p. 132; Kaufman, 2010; Robins and Richardson, 2011; Albert and Nelson, 2011).

[4]In the presence of measured and unmeasured confounders, the general conditions under which NDE is estimable from population data are somewhat more stringent than those needed for CDE (footnote 3). For details see Pearl (2001); Avin et al. (2005); Petersen et al. (2006); Robins (2003); VanderWeele (2009); Kaufman (2010); Robins and Richardson (2011); Shpitser and VanderWeele (2011).

[5]But see Robins and Richardson (2011) for ways of simulating the deactivation of direct paths by hypothetical interventions on auxiliary intermediate variables along that path.

The intuition here is quite different and unveils a nonparametric version of the product-of-coefficients strategy. The term $E(Y|x,z)$ plays the role of γ in Fig. 1(a), for it specifies how Y responds to Z for any 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 (8) 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, 2010).

Not surprising, owed to the nonlinear nature of the model, the relationship between the total, direct and indirect effects is non-additive. 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,

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

where $IE_{x',x}(Y)$ stands for the indirect effect of the transition from $X = x'$ to $X = x$. In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have $IE_{x,x'} = \beta\gamma(x' - x) = -I_{x',x}$ and the standard additive formula prevails:

$$TE_{x,x'}(Y) = NDE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{10}$$

Moreover, 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. In general, however, additivity is a rare occurrence and it is the difference formula in Eq. (9) that governs the relation between the total direct and indirect effects of the transition from x to x'.

In the rest of the paper we will drop the letter 'N' from the acronym NDE, with the understanding that DE stands for the natural direct effect estimand given by Eq. (6). When no ambiguity arises, we will also drop the subscripts x, x' and from effect acronyms and use TE, DE and IE for the total, direct and indirect effects, respectively.

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

This subsection demonstrates how the Mediation Formula of Eq. (8) 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 (see footnote 4), that Z may represent a vector of variables, and that integrals should replace summations when dealing with continuous variables (Imai et al., 2010a).

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. (A response is "owed" to a path if it would not have occurred were it not for the mechanism represented by that path.) These two groups are not necessarily mutually exclusive as can be seen in our enzyme example; individuals who respond *only* in the presence of both the enzyme and the drug should owe their response to both the direct and indirect paths.

3.1 Estimating mediation effects:

The Mediation Formula (8) 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 estimate the conditional expectation

$$g(x, z) = E(Y|x, z) \tag{11}$$

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

$$IE_{x,x'}(Y) = E_{Z|X}[g_x(z)|x'] - E_{Z|X}[g_x(z)|x]. \tag{12}$$

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 may arise to use parametric or semi-parametric approximations. We can then choose an appropriate 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 (8) 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):

$$\begin{aligned} x &= a_0 + \epsilon_1 \\ z &= b_0 + \beta x + \epsilon_2 \\ y &= c_0 + \alpha x + \gamma z + \epsilon_3 \end{aligned} \tag{13}$$

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 (8) gives

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

and yields

$$\begin{aligned} IE_{x,x'}(Y) &= \sum_z (c_0 + \alpha x + \gamma z)[P(z|x') - P(z|x)] \\ &= \gamma[E(Z|x') - E(Z|x)] \qquad (14) \\ &= (x' - x)(\beta\gamma) \qquad (15) \\ &= (x' - x)(\tau - \alpha) \qquad (16) \end{aligned}$$

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 16) or as a product $\beta\gamma$ of two regression coefficients (equation 15) (see MacKinnon et al., 2007b). These two strategies do not generalize to nonlinear systems (Pearl, 2011a) 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 (Judd and Kenny, 1981; Jo, 2008; Kraemer et al., 2008; MacKinnon, 2008). Further assume that we correctly account for this added term and, through diligent analysis on a large data set, 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.[6] In linear analysis, the former fraction is captured by the product $\beta\gamma/\tau$ (Eq. 15), the latter by the difference $(\tau - \alpha)/\tau$ (Eq. 16) 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 (8) and assuming $x = 0$ and $x' = 1$, yields the following decomposition:

$$\begin{aligned} DE_{0,1} &= \alpha + b_0\delta \\ IE_{0,1} &= \beta\gamma \\ TE_{0,1} &= \alpha + b_0\delta + \beta(\gamma + \delta) \\ &= DE_{0,1} + IE_{0,1} + \beta\delta \end{aligned}$$

[6] By "explain" we mean "sufficient to sustain even in the absence of direct effect." By "owed to" we mean "would not occur absent of mediation." These interpretations follow from the counterfactual definitions formulated in Section 2, of which Eqs. (6) and (8) are a statistical estimands.

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

$$IE_{0,1} = \beta\delta$$

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

$$TE_{0,1} - DE_{0,1} = \beta\gamma + \beta\delta$$

These conclusions are not readily discernible from the structural equations without the guidance of the Mediation Formula and, indeed, they have not been deduced by previous analyses (Judd and Kenny, 1981; Jo, 2008; Kraemer et al., 2008; MacKinnon, 2008).

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 way 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 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 $E(Y|x,z) = g_{x,z}$ and $E(Z|x) = h_x$, needed for (6), (8), and (11), can be readily

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

$$DE = (g_{1,0} - g_{0,0})(1 - h_0) + (g_{1,1} - g_{0,1})h_0 \tag{17}$$
$$IE = (h_1 - h_0)(g_{0,1} - g_{0,0}) \tag{18}$$
$$TE = g_{1,1}h_1 + g_{1,0}(1 - h_1) - [g_{0,1}h_0 + g_{0,0}(1 - h_0)] \tag{19}$$

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 all omitted factors (ϵ_2 and ϵ_3 in Fig. 2) are judged to be independent. Our research question is whether Z transmits the action of X on Y, or is merely a catalyst that accelerates the direct action of X on Y.

Treatment X	Enzyme present Z	Percentage cured $g_{x,z} = E(Y\|x,z)$
YES	YES	$g_{1,1} = 80\%$
YES	NO	$g_{1,0} = 40\%$
NO	YES	$g_{0,1} = 30\%$
NO	NO	$g_{0,0} = 20\%$

Table 2:

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

Table 3:

Substituting this data into Eqs. (17)–(19) yields:

$$DE = (0.40 - 0.20)(1 - 0.40) + (0.80 - 0.30)0.40 = 0.32$$
$$IE = (0.75 - 0.40)(0.30 - 0.20) = 0.035$$
$$TE = 0.80 \times 0.75 + 0.40 \times 0.25 - (0.30 \times 0.40 + 0.20 \times \cancel{0.10}\ 0.60) = 0.46$$
$$IE/TE = 0.07 \qquad DE/TE = 0.696 \qquad 1 - DE/TE = 0.304$$

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 develop a cheaper drug, identical to the one studied, but lacking the potential to stimulate enzyme secretion would face a reduction of 30.7% in recovery cases. More decisively, proposals to substitute the drug with one that merely mimics its stimulant action on Z but has no direct effect on Y are bound for failure; the drug evidently has a beneficial effect on recovery that is independent of, though enhanced by enzyme stimulation.

4 Relations to Traditional 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, nor do they agree on a principle for defining those effects when the process is known. 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, 2011a). 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. (17)–(19).

Attempts to extend the difference and product heuristics to nonparametric analysis have encountered ambiguities that the Mediation Formula can now 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,

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

but does not specify on what value we should condition X. Equation (18) resolves this ambiguity: C_γ should be conditioned on $X=0$ in order for the product $C_\beta C_\gamma$ to yield the correct mediation measure, IE.

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

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

and subtract it from the total effect, but does not specify on what value we should condition Z. Equation (17) 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 (Pearl, 2011a; Imai et al., 2010a). 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. (8) explicates this target quantity formally, and casts it in terms of estimable quantities.

The power of the Mediation Formula was recognized by Petersen et al. (2006); Glynn (2009); VanderWeele and Vansteelandt (2009); Hafeman and Schwartz (2009); Mortensen et al. (2009); VanderWeele (2009); Kaufman (2010); Imai et al. (2010a). Imai et al. (2010c) have further shown that nonparametric identification of mediation effects under the no-confounding assumption (Fig. 2) allows for a flexible estimation strategy and illustrate this with various nonlinear models, quantile regressions, and generalized additive models. Imai et al. (2010b) describe an implementation of these extensions using a convenient R package. Sjölander (2009) provides bound on DE in cases where the confounders between Z and Y cannot be controlled.

The ability of the Mediation Formula to carry us across the linear-nonlinear barrier may suggest that all mediation-related questions can now be answered nonparametrically and, more specifically, that, similar to traditional path analysis in linear systems, we can now assess the mediating effect of *any* chosen path or a bundle of paths in a causal diagram (Alwin and Hauser, 1975; Bollen, 1989). This turned out not to be the case. Avin et al. (2005) showed that there are many bundles of paths (i.e., subgraphs) in a graph G whose mediation effects cannot be assessed from either observational or experimental studies, even in the absence of unobserved confounders. They proved that the mediation effect of a subgraph SG is estimable if and only if it contains no "broken fork," that is, a path p_1 from X to some vertex W, and two paths, p_2 and p_3, from W to Y, such that p_1 and p_2 are in SG and p_3 is in G but not in SG.

Clearly, a broken fork condition cannot occur in the graph of Fig. 1(b), and this enables us to assess the mediation effect of any subset of $\{Z_1, Z_2, Z_3\}$. However, if we add the arrow $Z_3 \rightarrow Z_2$, then the effect contributed by the path $X \rightarrow Z_3 \rightarrow Z_2 \rightarrow Y$ would not be estimable, because the path $p_3 : Z_3 \rightarrow Y$ has been removed from the evaluated subgraph, thus creating a "broken fork".

Methodologically, the Mediation Formula brings with it all the advantages of a framework that allows one to define concepts such as mediation without resorting to parametric models.[7]

[7]The philosophical basis for this methodology has been expounded and debated elsewhere (e.g., Pearl (2000, 2009, 2010, 2011b)) and has led to the unification of counterfactual logic, structural equations and graphical models.

This allows us to separate the process of defining an estimand and statistical model that represent the target causal quantity and knowledge about the data generating process from the estimation procedure. In high dimensional problems, we can then draw on the large literature on non-parametric (and when appropriate) semi-parametric estimation.

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 asymptotically ensures bias-free estimates while making no assumption on the distributional form of the underlying process.[8]

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 ordinary 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 (i.e., no unmeasured confounders), 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.

Acknowledgments

This paper has benefited from the comments of three anonymous reviewers and from discussions with Kosuke Imai, Booil Jo, Marshall Joffe, David Kenny, Helena Kraemer, David MacKinnon, Ilya Shpitser, Patrick Shrout, Steven Sussman, Dustin Tingley, Mark Vander-Laan, and Tyler VanderWeele. This research 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

Albert, J. M. and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, pages DOI: 10.1111/j.1541–0420.2010.01547.x.

Alwin, D. and Hauser, R. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 40:37–47.

[8]In practice, the estimates produced may still suffer from misspecification bias, finite-sample bias, and sample-selection bias (see Bareinboim and Pearl (2011)).

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, pages 357–363, Edinburgh, UK. Morgan-Kaufmann Publishers.

Bareinboim, E. and Pearl, J. (2011). Controlling selection bias in causal inference. Technical Report R-381, <http://ftp.cs.ucla.edu/pub/stat_ser/r381.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI-11)*.

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(6):1173–1182.

Bollen, K. (1989). *Structural Equations with Latent Variables*. John Wiley, New York.

Cole, S. and Hernán, M. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology*, 31(1):163–165.

Glynn, A. (2009). The product and difference fallacies for indirect effects. Technical report, Department of Government and The Institute for Quantitative Social Sciences, Harvard University. Submitted for Publication.

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 Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2010b). Causal mediation analysis using R. In Vinod, H., editor, *Advances in Social Science Research Using R*, pages 129 − −154, <http://imai.princeton.edu/research/mediationR.html>. Springer (Lecture Notes in Statistics), New York.

Imai, K., Keele, L., and Yamamoto, T. (2010c). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, 13(4):314–336.

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(1):74–97.

Judd, C. and Kenny, D. (1981). *Estimating the Effects of Social Interactions*. Cambridge University Press, Cambridge, England.

Kaufman, J. (2010). Invited commentary: Decomposing with a lot of supposing. *American Journal of Epidemiology*, 172(12):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.

Mortensen, L., Diderichsen, F., Smith, G., and Andersen, A. (2009). The social gradient in birthweight at term: Quantification of the mediating role of maternal smoking and body mass index. *Human Reproduction*, 24(10):2629–2635.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27(2):226–284.

Pearl, J. (2000). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):428–431.

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

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

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):DOI: 10.2202/1557–4679.1203, <http://www.bepress.com/ijb/vol6/iss2/7/>.

Pearl, J. (2011a). The mediation formula: A guide to the assessment of causal pathways in non-linear models. In Berzuini, C., Dawid, P., and Bernardinelli, L., editors, *Causal Inference: Statistical Perspectives and Applications*. Wiley and Sons. Forthcoming.

Pearl, J. (2011b). Principal stratification a goal or a tool? *The International Journal of Biostatistics*, 7. Article 20, DOI: 10.2202/1557-4679.1322.

Petersen, M., Sinisi, S., and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology*, 17(3):276–284.

Robins, J. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In Green, P., Hjort, N., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 70–81. Oxford University Press, Oxford.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Robins, J. and Richardson, T. (2011). Alternative graphical causal models and the identification of direct effects. In Shrout, P. E., Keyes, K. M., and Ornstein, K., editors, *Causality and Psychopathology, Finding the Determinants of Disorder and their Cures*, pages 103–158. Oxford University Press, New York.

Shpitser, I. and VanderWeele, T. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *The International Journal of Biostatistics*, 7(1):Article 16.

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, 28:558–571.

Tian, J. and Shpitser, I. (2010). On identifying causal effects. In Dechter, R., Geffner, H., and Halpern, J., editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 415–444. College Publications, UK.

VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26.

VanderWeele, T. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.