To appear in C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), Causality: Statistical Perspectives and Applications. Forthcoming, 2011.

TECHNICAL REPORT
R-363
March 2011

The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models

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

March 30, 2011

Abstract

Mediation analysis aims to uncover causal pathways along which changes are transmitted from stimulus to response. 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, thus setting a causally-sound standard for mediation analysis of empirical data. 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

*Portions of this paper are adapted from Pearl (2000, 2009a,b, 2010); I am indebted to Elias Bareinboim, Adam Glynn, Donald Green, Booil Jo, Marshall Joffe, Helena Kraemer, David MacKinnon, Stephen Morgan, Patrick Shrout, Arvid Sjölander, Dustin Tingley, Tyler VanderWeele, Scott Weaver, Christopher Winship, Teppei Yamamoto, and readers of the UCLA Causality Blog (http://www.mii.ucla.edu/causality/) for discussions leading to these explorations. Send comments to <judea@cs.ucla.edu>. This research was supported in parts by NIH #1R01 LM009961-01, NSF #IIS-0914211 and #IIS-1018922, and ONR #N000-14-09-1-0665 and #N00014-10-1-0933.

systems with categorical variables, including logistic, probit, and nonparametric regressions.

Keywords: Effect decomposition, indirect effects, structural equation models, graphical models, counterfactuals, causal effects, potential-outcome, structural causal models, surrogate endpoints.

1 Mediation: Direct and Indirect Effects

1.1 Direct versus Total Effects

The target of many empirical studies in the social, behavioral, and health sciences is the *causal effect*, here denoted $P(y|do(x))$, which measures the *total* effect of a manipulated variable (or a set of variables) X on a response variable Y. In many cases, this quantity does not adequately represent the target of investigation and attention is focused instead on the *direct* effect of X on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

From a policy making viewpoint, an investigator may be interested in decomposing effects to quantify the extent to which weakening or strengthening specific causal pathways would impact the overall effect of X on Y. For example, the extent to which minimizing racial disparity in education would reduce racial disparity in earning. Or, taking a health-related example, the extent to which efforts to eliminate side-effect of a given treatment are likely to weaken or enhance the efficacy of that treatment. More often, however, the decomposition of effects into their direct and indirect components carries theoretical scientific importance, for it tells us "how nature works" and, therefore, enables us to predict behavior under a rich variety of conditions and interventions.

Structural equation models provide a natural language for analyzing path-specific effects and, indeed, considerable literature on direct, indirect and total effects has been authored by SEM researchers (Alwin and Hauser (1975), Graff and Schmidt (1982), Sobel (1987), Bollen (1989)), for both recursive and nonrecursive models. This analysis usually involves sums of powers of coefficient matrices, where each matrix represents the path coefficients associated with the structural equations.

Yet despite its ubiquity, the analysis of mediation has long been a thorny issue in the empirical sciences (Judd and Kenny, 1981; Baron and Kenny, 1986; Muller et al., 2005; Shrout and Bolger, 2002; MacKinnon et al., 2007a) primarily because structural equation modeling in those sciences were deeply entrenched in linear analysis, where the distinction between causal parameters and their regressional interpretations can easily be conflated (as in Holland, 1995; Sobel, 2008). The difficulties were further amplified in nonlinear models, where sums and products are no longer applicable. As demands grew to tackle problems involving binary and categorical variables, researchers could no longer define direct and indirect effects in terms of structural or regressional coefficients, and all attempts to extend the linear paradigms of effect decomposition to nonlinear systems produced distorted results (MacKinnon et al., 2007b). These difficulties have accentuated the need to redefine and derive causal effects from first principles, uncommitted to distributional assumptions or a particular parametric form of the equations. The structural methodology presented in this paper adheres to this philosophy and it has produced indeed a principled solution to the mediation problem, based on the counterfactual reading of structural equations (Balke and Pearl, 1994a,b; Pearl, 2009a, Chapter 7). The following subsections summarize the method and its solution, while Section 2 introduces the Mediation Formula, exemplifies its behavior, and demonstrates its usage in simple examples, including linear, quasi-linear, logistic, probit and nonparametric models. Finally, Section 4 compares the Mediation Formula to other methods proposed for effect decomposition and explains the difficulties that those methods have encountered in defining and assessing mediated effects.

1.2 Controlled Direct Effects

A major impediment to progress in mediation analysis has been the lack of notational facility for expressing the key notion of "holding the mediating variables fixed" in the definition of direct effect. Clearly, this notion must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, regression conditioning, stratification matching or adjustment. For example, consider the simple mediation models of Fig. 1(a), where the error terms (not shown explicitly) are assumed to be mutually independent. To measure the direct effect of X on Y it is sufficient to measure their association conditioned on the

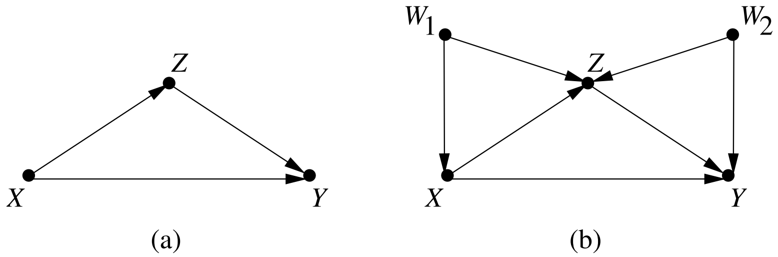

Figure 1: (a) A generic model depicting mediation through Z with no confounders, and (b) with two confounders, W_1 and W_2.

mediator Z. In Fig. 1(b), however, where the error terms are dependent, it will not be sufficient to measure the association between X and Y for a given level of Z because, by conditioning on the mediator Z, we create spurious associations between X and Y through W_2, even when there is no direct effect of X on Y (Pearl, 1998; Cole and Hernán, 2002).

Using the $do(x)$ notation, enables us to correctly express the notion of "holding Z fixed" and obtain a simple definition of the *controlled direct effect* of the transition from $X = x$ to $X = x'$ (Pearl, 2009a, p. 128):

$$CDE \triangleq E(Y|do(x'), do(z)) - E(Y|do(x), do(z)) \tag{1}$$

or, equivalently, using counterfactual notation:

$$CDE \triangleq E(Y_{x'z}) - E(Y_{xz})$$

where Z is the set of all mediating variables.[1] Readers can easily verify that, in linear systems, the controlled direct effect reduces to the path coefficient of the link $X \rightarrow Y$ regardless of whether confounders are present (as in Fig. 1(b)) and regardless of whether the error terms are correlated or not.

This separates the task of definition from that of identification, and thus circumvents many pitfalls in this area of research (Pearl, 2009b). The identification of CDE would depend, of course, on whether confounders are present and whether they can be neutralized by adjustment, but these do not alter its definition. Nor should trepidation about infeasibility of the action

[1] Readers not familiar with this notation can consult (Pearl, 2009a,b, 2010). Conceptually, $P(y|do(x))$ stands for the probability of $Y = y$ in a randomized experiment where treatment level is set to $X = x$, while $Y_x(u)$ stands for the value that Y would attain in unit u, had X been x. Formally, $P(y|do(x))$ and $Y_x(u)$ are defined, respectively, as the probability and value of variable Y in a modified structural model, in which the equation for X is replaced by a constant $X = x$).

$do(gender = male)$ enter the definitional phase of the study. Definitions apply to symbolic models, not to human biology.[2]

Graphical identification conditions for expressions of the type $E(Y|do(x), do(z_1), do(z_2), \ldots, do(z_k))$ in the presence of unmeasured confounders were derived by Pearl and Robins (1995) and invoke sequential application of the back-door condition (Pearl, 2009a, pp. 252–254), which is somewhat more powerful than G-computation (Robins, 1986). Tian and Shpitser (2010) have further derived a necessary and sufficient condition for this task, and thus resolved the identification problem for controlled direct effects (Eq. 1).

1.3 Natural Direct Effects

In linear systems, the direct effect is fully specified by the path coefficient attached to the link from X to Y; therefore, the direct effect is independent of the values at which we hold Z. In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, it is not uncommon to find employers who prefer males for the high-paying jobs (i.e., high z) and females for low-paying jobs (low z). Focusing on one of these values of Z, or averaging over all values would not capture the underlying pattern of discrimination.

When the direct effect is sensitive to the levels at which we hold Z, it is often more meaningful to define the direct effect relative to some "natural" base-line level that may vary from individual to individual, and represents the level of Z just before the change in X. Conceptually, we can define the natural direct effect $DE_{x,x'}(Y)$ 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 $do(x)$. This hypothetical change, which Robins and Greenland (1992) conceived and called "pure" and Pearl (2001) formalized and analyzed under the rubric "natural," mirrors what lawmakers instruct us to consider in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In *Carson versus Bethlehem Steel Corp.*, 70 FEP Cases 921, 7th Cir. (1996)).

Thus, whereas the controlled direct effect measures the effect of X on Y while holding Z fixed, at a uniform level (z) for all units,[3] the natural direct

[2] In reality, it is the employer's perception of applicant's gender and his/her assessment of gender-job compatibility that render gender a "cause" of hiring – manipulation of gender is not needed.

[3] In the hiring discrimination example, this would amount, for example, to testing gender

effect allows z to vary from individual to individual and be fixed at the level that each individual held naturally, just before the change in X.

Pearl (2001) gave the following definition for the "natural direct effect":

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{2}$$

Here, Y_{x',Z_x} represents the value that Y would attain under the operation of setting X to x' and, simultaneously, setting Z to whatever value it would have obtained under the setting $X = x$. For example, if one were to estimate that the natural direct effect of gender on hiring equals 20% of the total effect, one can infer that 20% of the current gender-related disparity in hiring can be eliminated by making hiring decision gender-blind, while keeping applicants qualifications at their current values (which may be gender dependent).

We see from (2) that $DE_{x,x'}(Y)$, the natural direct effect of the transition from x to x', involves probabilities of *nested counterfactuals* and cannot be written in terms of the $do(x)$ operator. Therefore, the natural direct effect cannot in general be identified or estimated, even with the help of ideal, controlled experiments – a point emphasized in Robins and Greenland (1992).[4] However, aided by the formal definition of Eq. (2) and the notational power of nested counterfactuals, Pearl (2001) was nevertheless able to derive conditions under which the natural direct effect can be expressed in terms of the $do(x)$ operator, implying identifiability from controlled experiments. For example, if a set W exists that deconfounds Y and Z, the natural direct effect can be reduced to[5]

$$DE_{x,x'}(Y) = \sum_{z,w} [E(Y|do(x',z),w) - E(Y|do(x,z),w)]P(z|do(x),w). \tag{3}$$

The intuition is simple; the W-specific natural direct effect is the weighted average of the controlled direct effect, using the causal effect $P(z|do(x),w)$ as

bias while marking all application forms with the same level of schooling and other skill-defining attributes.

[4] The reason being that we cannot rerun history and test individuals' response both before and after the intervention. Robins (2003) elaborates on the differences between the assumptions made in (Pearl, 2001) and the weaker assumptions made in Robins and Greenland (1992), which prevented the latters from identifying natural effects even in the simple case of no-confounding. (Fig. 1(a)).

[5] The key condition for this reduction is the existence of a set W of covariates satisfying $Y_{xz} Z_{x'} \perp\!\!\!\perp W$, which simply states that W blocks all back-door paths from Z to Y (see Pearl (2009a, p. 101)). More refined counterfactual conditions for identification are derived in Petersen et al. (2006), Imai et al. (2010c), and Robins and Richardson (2011). However, none matches the clarity of the back-door condition above, and all are equivalent in the graphical language of non-parametric structural equations (Shpitser and VanderWeele, 2011).

a weighing function.[6]

In particular, it can be shown (Pearl, 2001) that the natural direct effect is identifiable in Markovian models (i.e., recursive equations with no unobserved confounders) where each *do*-expression can be reduced to a "*do*-free" expression by covariate adjustments (Pearl, 2009a) and then estimated by regression. For example, for the model in Fig. 1(b), $DE_{x,x'}(Y)$ reduces to:

$$DE_{x,x'}(Y) = \sum_{z}\sum_{w_2} P(w_2)[E(Y|x',z,w_2)) - E(Y|x,z,w_2))] \sum_{w_1} P(z|x,w_1,w_2)P(w_1). \tag{4}$$

while for the confounding-free model of Fig. 1(a) we have:

$$DE_{x,x'}(Y) = \sum_{z}[E(Y|x',z) - E(Y|x,z)]P(z|x). \tag{5}$$

Both (4) and (5) can be estimated by a regression.

When Z consists of multiple interconnected mediators, affected by an intricate network of observed and unobserved confounders, the adjustment illustrated in Eq. (4) must be handled with some care. Theorems 1 and 2 of Pearl (2001) can then be used to reduce $DE_{x,x'}(Y)$ to a *do*-expression similar to (3) (see footnote 5). Once reduced, the machinery of *do*-calculus (Pearl, 1995) can be invoked, and the methods of Pearl and Robins (1995), Tian and Shpitser (2010), and Shpitser and VanderWeele (2011) can select the proper set of covariates and reduce the natural direct effect (3) to an expression estimable by regression, whenever such reduction is feasible. For example, if in Fig. 1(b) W_1 is unobserved and another observed covariate, W_3, mediates the path $X \rightarrow Z$, $E(Y|do(x,z),w_2)$ is then identifiable through the front-door formula (Pearl, 1995, 2009a), thus rendering $DE_{x,x'}(Y)$ estimable by regression. This demonstrates that neither "ignorability" (Rosenbaum and Rubin, 1983) nor "sequential ignorability" (Imai et al., 2010a) is necessary for securing the identification of direct effects; transparent graph-based criteria are sufficient for determining when and how confounding can be controlled. See (Pearl, 2009a, pp. 341–344) for graphical interpretation of "ignorability" assumptions.

1.4 Indirect effects

Remarkably, the definition of the natural direct effect (2) 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

[6] Throughout this paper we will use summation signs with the understanding that integrals should be used whenever the summed variables are continuous.

variables in the model, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The *natural 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 to whatever value it would have attained had X been set to $X = x'$. Formally, this reads (Pearl, 2001):

$$IE_{x,x'}(Y) \triangleq E[(Y_{x,Z_{x'}}) - E(Y_x)], \tag{6}$$

which is almost identical to the direct effect (Eq. (2)) save for exchanging x and x' in the first term.

Invoking the same conditions that led to the experimental identification of the direct effect, Eq. (3), we obtain a parallel formula for the indirect effect:

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

The intuition here is somewhat different, and represents a nonlinear version of the "product-of-coefficients" strategy in linear models (MacKinnon, 2008); the $E(Y|do(x,z),w)$ term encodes the effect of Z on Y for fixed $X = x$ and $W = w$, while the $[P(z|do(x'),w) - P(z|do(x),w)]$ encodes the effect of X on Z. We see that what was a simple product-of-coefficients in linear models turns into a convolution type operation, involving all values of Z.

In non-experimental studies, the *do*-operator need be reduced to regression type expression using covariate adjustment or instrumental variable methods. For example, for the model in Fig. 1(b), Eq. (7) reads:

$$IE_{x,x'}(Y) = \sum_z \sum_{w_2} P(w_2)[E(Y|x,z,w_2)) \sum_{w_1}[P(z|x',w_1,w_2) - P(z|x,w_1,w_2)P(w_1)]. \tag{8}$$

while for the confounding-free model of Fig. 1(a) we have

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)] \tag{9}$$

which, like Eq. (5) can be estimated by a two-step regression.

1.5 Effect decomposition

Not surprisingly, owed to the nonlinear nature of the model, the relationship between the total, direct and indirect effects is non-additive. Indeed, it can be shown that, in general, the total effect TE of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the *reverse* transition. Formally,

$$TE_{x,x'}(Y) \overset{\Delta}{=} E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{10}$$

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

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{11}$$

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.[7]

Note that, although it cannot in general be expressed in *do*-notation, the indirect effect has clear policy-making implications. For example: in the hiring discrimination context, a policy maker may be interested in predicting the gender mix in the work force if gender bias is eliminated and all applicants are treated equally—say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent.

More generally, a policy maker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully selecting the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*, that is, the effect of X on Y through a selected set of paths (Avin et al., 2005). Avin et al. (2005), with all other paths *deactivated*. The operation of disabling a path can be expressed in nested counterfactual notation, as in Eqs. (2) and (6).

In all these cases, the policy intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Pearl (2001) has suggested therefore that *signal sensing* is more fundamental to the notion of causation than *manipulation*; the latter being but a crude way of stimulating the former in experimental setup. The mantra "No causation without manipulation" must be rejected. (See (Pearl, 2009a, Section 11.4.5).)

[7] Some authors (e.g., VanderWeele (2009); Vansteelandt (2011), (Chapter [VANSTEELANDT]) take Eq. (11) as the definition of indirect effect (see footnote 8), which ensures additivity by definition, but presents a problem of interpretation; the resulting indirect effect, aside from being redundant, does not represent the same transition, from x to x', as do the total and direct effects. This prevents us from comparing the effect attributable to mediating paths with that attributable to unmediated paths, under the same conditions.

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

2.1 Mediation in non-parametric models

This subsection demonstrates how the solution provided in equations (5) and (9) can be applied in assessing mediation effects in non-parametric, possibly nonlinear models. We will use the simple mediation model of Fig. 1(a), where all error terms (not shown explicitly) are assumed to be mutually independent, with the understanding that adjustment for appropriate sets of covariates W may be necessary to achieve this independence (as in (4) and (8)) and that integrals should replace summations when dealing with continuous variables (Imai et al., 2010c).

Combining (5), (9), and (10), the expressions for the direct (DE), indirect (IE) and total (TE) effects, IE become:

$$DE_{x,x'}(Y) = \sum_z [E(Y|x',z) - E(Y|x,z)]P(z|x) \tag{12}$$

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)] \tag{13}$$

$$TE_{x,x'}(Y) = E(Y|x') - E(Y|x) \tag{14}$$

These three equations provide general formulas for mediation effects, applicable to any nonlinear system, any distribution, and any type of variables. Moreover, the formulas are readily estimable by regression. Owed to their generality and ubiquity, I have referred to these expressions as the "Mediation Formula" (Pearl, 2009b).

The Mediation Formula (13) 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. 1(a). 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.[8]

The Mediation Formula tells us that IE depends only on the expectation of the counterfactual Y_{xz}, not on its functional form $f_Y(x, z, u_Y)$ or its distribution $P(Y_{xz} = y)$. It calls therefore for a two-step regression which, in principle, can

[8] For simplicity and clarity, we remove the subscripts from TE, DE, and IE, whenever no ubiquity arises. Robins (2003) and Hafeman and Schwartz (2009) refer to $TE - IE$ and $TE - DE$ as "total direct" and "total indirect" effects, respectively.

be performed non-parametrically. In the first step we regress Y on X and Z, and obtain the estimate

$$g(x, z) \triangleq E(Y|x, z)$$

for every (x, z) cell. In the second step we fix x and estimate the conditional expectation of $g(x, z)$ with respect to z, conditional on $X = x'$ and $X = x$, respectively, and take the difference:

$$IE_{x,x'}(Y) = E_{Z|x'}[g(x, z)] - E_{Z|x}[g(x, z)]$$

Non-parametric 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, quasi-linear logit, probit), estimate the parameters separately (e.g., by regression or maximum likelihood methods), insert the parametric approximation into (13) and estimate its two conditional expectations (over z) to get the mediated effect.

The power of the Mediation Formula was recognized by Petersen et al. (2006); Glynn (2009); Hafeman and Schwartz (2009); Mortensen et al. (2009); VanderWeele (2009); Kaufman (2010); Imai et al. (2010c). Imai et al. (2010a) have further shown that nonparametric identification of mediation effects under the no-confounding assumption (Fig. 1a) 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.

In the next section this power will be demonstrated on linear and nonlinear models, with the aim of explaining the distortions produced by conventional methods of parametric mediation analysis, and how they are rectified through the Mediation Formula.

2.2 Mediation effects in linear, logistic, and probit models

The linear case: Difference versus product estimation

Let us examine what the Mediation Formula yields when applied to the linear version of model 1(a), which reads:

$$
\begin{aligned}
x &= u_X \\
z &= \gamma_{xz} x + u_Z \\
y &= \gamma_{xy} x + \gamma_{zy} z + u_Y
\end{aligned}
\tag{15}
$$

Computing the conditional expectation in (13) gives

$$g(x,z) = E(Y|x,z) = E(\gamma_{xy} x + \gamma_{zy} z + u_Y) = a_0 + \gamma_{xy} x + \gamma_{zy} z$$

and yields

$$
\begin{aligned}
DE_{x,x'} &= \sum_z [(a_0 + \gamma_{xy} x' + \gamma_{zy} z) - (a_0 + \gamma_{xy} x + \gamma_{zy} z)] P(z|x) \\
&= \gamma_{xy}(x' - x) && (16) \\
IE_{x,x'}(Y) &= \sum_z (a_0 + \gamma_{xy} x + \gamma_{zy} z)[P(z|x') - P(z|x)]. \\
&= \gamma_{zy}[E(Z|x') - E(Z|x)] \\
&= (\gamma_{zy}\gamma_{xz})(x' - x) && (17) \\
&= (\beta_{xy} - \gamma_{xy})(x' - x) && (18)
\end{aligned}
$$

where β_{xy} is the regression coefficient $\beta_{xy} = \frac{\partial}{\partial x} E(Y|x) = \gamma_{xy} + \gamma_{xz}\gamma_{zy}$

$$
\begin{aligned}
TE_{x,x'}(Y) &= (E(Y|x') - E(Y|x)) \\
&= \sum_z E(Y|x', z) P(z|x') + \sum_z E(Y|x, z) P(z|x) \\
&= \sum_z (a_0 + \gamma_{zy} x' + \gamma_{zy} z) P(z|x') - \sum_z (a_0 + \gamma_{xy} x + \gamma_{zy} z) P(z|x) \\
&= \gamma_{xy}(x' - x) + \gamma_{zy} E(Z|x') - \gamma_{zy} E(Z|x) \\
&= (\gamma_{xy} + \gamma_{zy}\gamma_{xz})(x' - x) && (19)
\end{aligned}
$$

We thus obtained the standard expressions for effects in linear systems. In particular, we see that the indirect effect can be estimated either as a difference in two regression coefficients (Eq. 18) or a product of two regression coefficients (Eq. 17), with Y regressed on both X and Z.[9] When generalized to nonlinear systems, however, these two strategies yield conflicting results (MacKinnon and Dwyer, 1993; MacKinnon et al., 2007b) and much controversy has developed as to which strategy should be used in assessing the size of mediation effects (MacKinnon and Dwyer, 1993; Freedman et al., 1992; Molenberghs et al., 2002; MacKinnon et al., 2007b; Glynn, 2009; Green et al., 2010).

We now show that neither of these strategies generalizes to nonlinear systems; direct application of (13) is necessary. Moreover, we will see that, though yielding identical results in linear systems, the two strategies represent legitimate intuitions in pursuits of two distinct causal quantities. The difference-in-coefficients method seeks to estimate $TE - DE$, while the product-of-coefficients method seeks to estimate IE. The former represents the reduction in TE if indirect paths were deactivated, while the latter represents the portion of TE that would remain if the direct path were deactivated. The choice between $TE - DE$ and IE depends of course on the specific decision making objectives that the study aims to inform. If the policy evaluated aims to prevent the outcome Y by ways of manipulating the mediating pathways, the target of analysis should be the difference $TE - DE$, which measures the highest prevention effect of any such manipulation. If, on the other hand, the policy aims to prevent the outcome by manipulating the direct pathway, the target of analysis should shift IE, for $TE-IE$ measures the highest preventive impact of this type of manipulations.

In the hiring discrimination example, $TE-DE$ gives the maximum reduction in racial earning disparity that can be expected from programs aiming to achieve educational parity. $TE - IE$ on the other hand measures the maximum reduction in earning disparity that can be expected from eliminating hiring discrimination by employers. The difference-in-coefficients strategy is motivated by the former types of problems while the product-of-coefficients by the latter.

The next section illustrates how nonlinearities bring about the disparity between IE and $TE - DE$.

[9] Note that the equality $\beta_{xy} - \gamma_{xy} = \gamma_{xz}\gamma_{zy}$ established in (18) 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 therefore 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 parameters in nonlinear regression to obey similar equalities should not be construed as an indication of faulty standardization, as suggested by (MacKinnon et al., 2007a,b).

The logistic case

To see how the Mediation Formula facilitates nonlinear analysis, let us consider the logistic and probit models treated in MacKinnon et al. (2007b).[10] To this end, let us retain the linear model of (15) with one modification: the outcome of interest will be a threshold-based indicator of the linear outcome Y in (15). In other words, we regard

$$Y^* = \gamma_{xy} x + \gamma_{zy} z + u_Y \tag{20}$$

as a latent variable, and define the outcome Y as

$$Y = \begin{cases} 1 & \text{if } \gamma_0 + \gamma_{xy} x + \gamma_{zy} z + u_Y > 0 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where γ_0 is some unknown threshold level. We will assume that the error U_Y is governed by the logistic distribution

$$P(U_Y < u) = L(u) \triangleq \frac{1}{1 + e^{-u}} \tag{22}$$

and, consequently, $E(Y|x, z)$ attains the form:

$$\begin{aligned} E(Y|x,z) &= \frac{1}{1 + e^{-(\gamma_0 + \gamma_{xy} x + \gamma_{zy} z)}} && (23) \\ &= L(\gamma_0 + \gamma_{xy} x + \gamma_{zy} z) && (24) \end{aligned}$$

We will further assume that U_Z is normal with zero mean and infinitesimal variance $\sigma_z^2 << 1$.

Given this logistic model and its parameter set $(\gamma_0, \gamma_{xy}, \gamma_{zy}, \gamma_{xz}, \sigma_z^2)$, we will now compute the direct (DE), indirect (IE) and total (TE) effects associated with the transition from $X = 0$ to $X = 1$. From the Mediation Formula

[10] Pearl (2010) analyzes Boolean models with Bernoulli noise.

((12)–(14)), we obtain:

$$\begin{aligned}
DE &= \int_{z=-\infty}^{\infty} [L(\gamma_0 + \gamma_{xy} + \gamma_{zy}) - L(\gamma_0 + \gamma_{zy} z)] f_{Z|X}(z|X=0) dz \\
&= L(\gamma_0 + \gamma_{xy}) - L(\gamma_0) + 0(\sigma_z^2) \qquad (25) \\
IE &= \int_{z=-\infty}^{\infty} [L(\gamma_0 + \gamma_{zy} z)[f_{Z|X}(z|X=1) - f_{Z|X}(z|X=0)] dz \\
&= L(\gamma_0 + \gamma_{zy}\gamma_{xz}) - L(\gamma_0) + 0(\sigma_z^2) \qquad (26) \\
TE &= E(Y|X=1) - E(Y|X=0) = \int_{z=\infty}^{\infty} E(Y|X=1,z) f_{Z|X}(z|X=1) dz \\
&\quad - \int_{z=\infty}^{\infty} E(Y|X=0,z) f_{Z|x}(z|X=0) dz \\
&= L(\gamma_0 + \gamma_{xy} x + \gamma_{zy} z) - L(\gamma_0) + 0(\sigma_z^2) \qquad (27)
\end{aligned}$$

where $0(\sigma_z^2) \to 0$ as $\sigma_z \to 0$.

It is clear that, due to the nonlinear nature of $L(u)$, none of these effects coincides with its corresponding effect in the linear model of Eq. (15). In other words, it would be wrong to assert the equalities:

$$\begin{aligned}
DE_{0,1} &= \gamma_{xy} \\
IE_{0,1} &= \gamma_{zy}\gamma_{xz} \\
TE_{0,1} &= \gamma_{xy} + \gamma_{zy}\gamma_{xz} = \beta_{xy}
\end{aligned}$$

as is normally assumed in the mediation literature (Prentice, 1989; Freedman et al., 1992; MacKinnon and Dwyer, 1993; Fleming and DeMets, 1996; Molenberghs et al., 2002; MacKinnon et al., 2007b). In particular, the mediated fractions $1 - DE/TE$, and IE/TE may differ substantially from the fractions $\gamma_{xz}\gamma_{zy}/(\gamma_{xy} + \gamma_{xz}\gamma_{zy}), 1 - \gamma_{xy}/\beta_{xy}$, or $\gamma_{xz}\gamma_{zy}/\beta_{xy}$ that have been proposed to evaluate mediation effects by traditional methods. The latters are heuristic ratios informed by the linear portion of the model, while the formers are derived formally from the counterfactual specifications of the target quantities, as in (2) and (6).

Figure 2 depicts DE, IE, and TE as a function of γ_0, the threshold coefficient that dichotomizes the outcome (as in Eq. (21)). These were obtained analytically, from Eqs. (25)-(27), using the values $\gamma_{xz} = \gamma_{xy} = \gamma_{zy} = 0.5$ for illustrative purposes. We see that all three measures vary with γ_0 and deviate substantially from the assumptions that equate DE with $\gamma_{xy} = 0.50$, IE with $\gamma_{xz}\gamma_{zy} = 0.25$ and TE with $\gamma_{xy} + \gamma_{xz}\gamma_{zy} = 0.75$ (MacKinnon and Dwyer, 1993; MacKinnon et al., 2007b).

The bias produced by such assumptions is further accentuated in Fig. 3, which compares several fractions (or proportions) proposed to measure

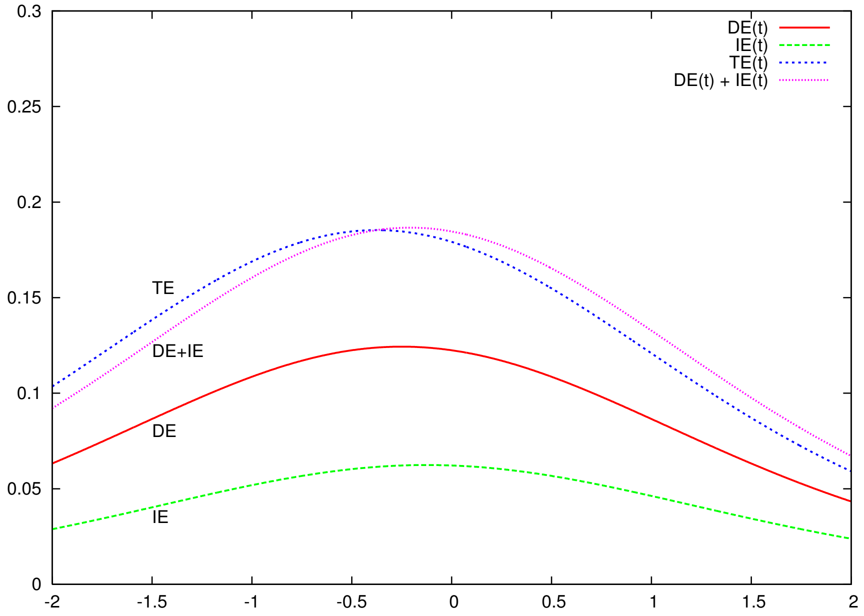

Figure 2: Direct (DE), indirect (IE) and total (TE) effects for the logistic model of Eq. (24) as a function of the threshold γ_0 that dichotomizes the outcome.

the relative contribution of mediation to the observed response. Recall that $1-DE/TE$ measures the extent to which mediation was *necessary* for the observed response, while IE/TE the extent to which it was *sufficient*. Figure 3 shows that the necessary fraction $(1-DE/TE)$ exceeds the sufficient fraction (IE/TE) as γ_0 becomes more negative. Indeed, in this region, both direct and indirect paths need be activated for Y^* to exceed the threshold of Eq. (22). Therefore, the fraction of responses for which mediation was necessary is high and the fraction for which mediation was sufficient is low. The disparity between the two will be revealed by varying the intercept γ_0, a parameter that is hardly paid noticed to in traditional analyses and which will be shown to be important for understanding the interplay between DE and IE, and the role they play in shaping mediated effects. The opposite occurs for positive γ_0 (negative threshold), where each path alone is sufficient for activating Y and it is unlikely therefore that the mediator becomes a necessary enabler of $Y=1$.

None of this dynamics is represented in the fixed fraction $\gamma_{xz}\gamma_{zy}/(\gamma_{xy}+\gamma_{xz}\gamma_{zy}) = 0.25/0.75 = 1/3$ which standard logistic regression would report as the fraction of cases "explained by mediation." Some of this dynamics is

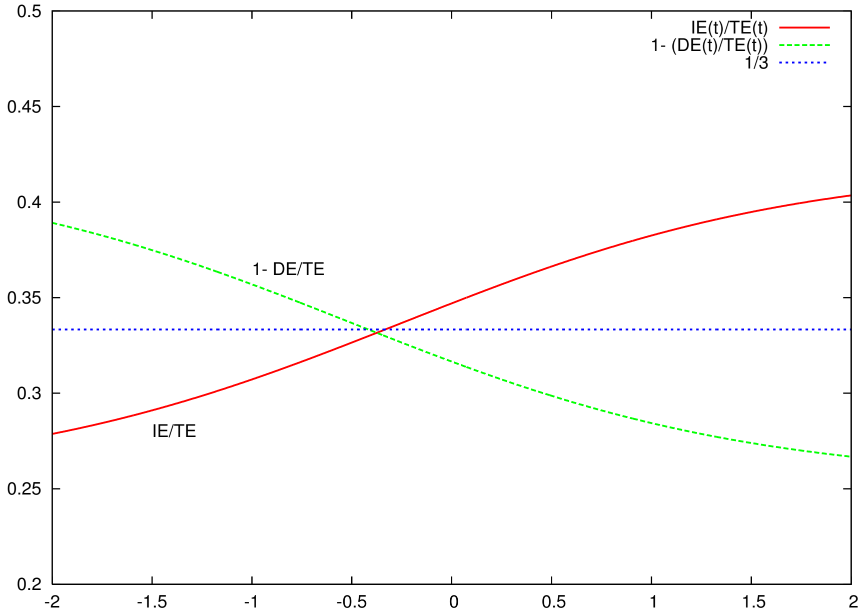

Figure 3: Necessary $(1 - DE/TE)$ and sufficient (IE/TE) mediation proportions for the logistic model of Eq. (24)

reflected in the fraction $\gamma_{xz}\gamma_{zy}/[E(Y|X = 1) - E(Y|X = 0)] = 0.25/TE$ (not shown in Fig. 3) which some researchers have recommended as a measure of mediation (MacKinnon and Dwyer, 1993; MacKinnon et al., 2007b). But this measure is totally incompatible with the correct fractions shown in Fig. 3. The differences are accentuated again for negative γ_0 (positive threshold), where both direct and indirect processes must be activated for Y^* to exceed the threshold, and the fraction of responses for which mediation is necessary $(1 - DE/TE)$ is high and the fraction for which mediation is sufficient (IE/TE) is low.

The probit case

Figure 4 displays the behavior of a probit model. It was computed analytically by assuming a probit distribution in Eq. (22), which leads to the same expressions in (21)–(24), with Φ replacing L. Noticeably, Figs. 4 and 5 reflect more pronounced variations of all effects with γ_0, as well as more pronounced deviation of these curve from the constant $\gamma_{xz}\gamma_{zy}/(\gamma_{xy} + \gamma_{xz}\gamma_{zy}) = 1/3$ that regression analysis defines as the "proportion mediated" measure (Sjölander, 2009). We speculate that the difference in behavior between the logistic and probit models is due to the latter sharper approach toward the asymptotic

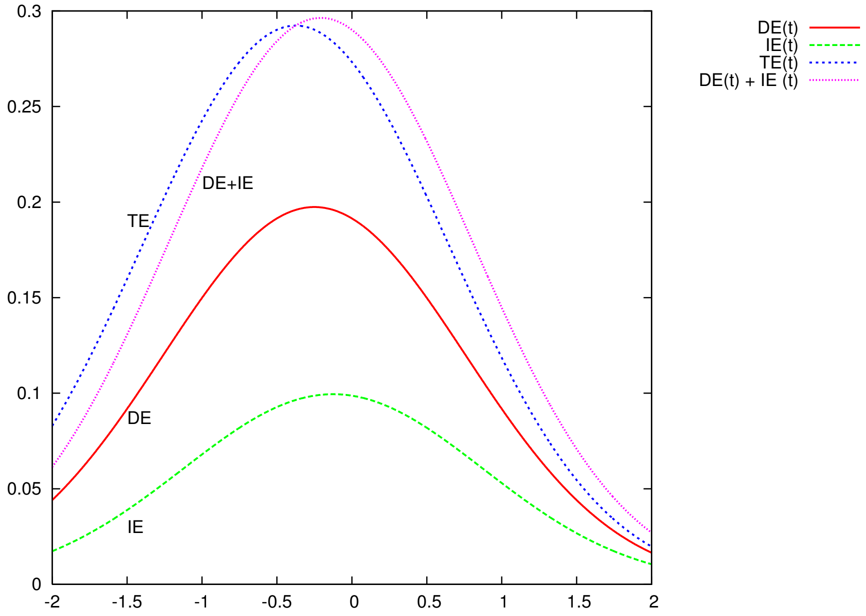

Figure 4: Direct (DE), indirect (IE) and total (TE) effects for a probit model.

limits.

2.3 Special cases of mediation models

In this section we will discuss three special cases of mediation processes that lend themselves to simplified analysis,

Incremental causal effects

Consider again the logistic threshold model of Eq. (21), and assume we are interested in assessing the response Y to an incremental change in the treatment variable X, say from $X = x$ to $X = x + \delta$. In other words, our target quantities are the limits as $\delta \to 0$ of

$$DE_{inc}(x) = \frac{1}{\delta} DE_{x,x+\delta}$$
$$IE_{inc}(x) = \frac{1}{\delta} DE_{x,x+\delta}$$
$$TE_{inc}(x) = \frac{1}{\delta} DE_{x,x+\delta}$$

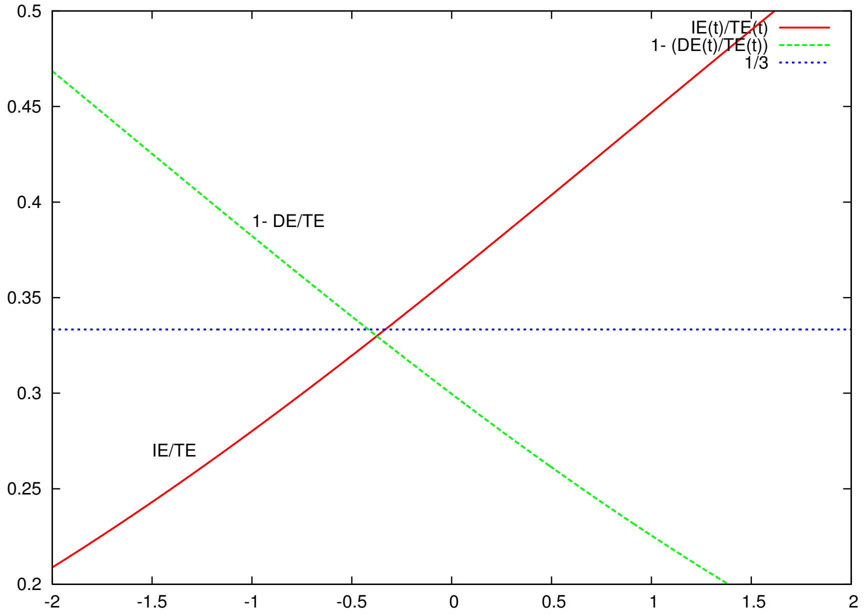

Figure 5: Necessary $(1 - DE/TE)$ and sufficient (IE/TE) mediation proportions for a probit model.

If we maintain the infinitesimal variance assumption $\sigma_z^2 << 1$, we obtain:

$$\begin{aligned} DE_{inc}(x) &= lim_{\delta\to 0}\frac{1}{\delta}DE_{x,x+\delta} \\ &= lim_{\delta\to 0}\frac{1}{\delta}\int [E(Y|x+\delta, z) - E(Y|x,z)]f_{Z|X}(z|x)dz \\ &= \frac{\partial}{\partial x}E(Y|x,z)|_{z=h(x)} + 0(\sigma_z^2) \end{aligned}$$

where $h(x) = E(Z|x)$.

Similarly, we have

$$
\begin{aligned}
IE_{inc}(x) &= \lim_{\delta\to 0}\frac{1}{\delta}IE_{x,x+\delta}\\
&= \lim_{\delta\to 0}\frac{1}{\delta}\int_z E(Y|x,z)[f(z|x+\delta)-f(z|x)]dz\\
&= \lim_{\delta\to 0}\frac{1}{\delta}E(Y|x,z)|_{z=h(x+\delta)} - E(Y|x,z)|_{z=h(x)}\\
&= \lim_{\delta\to 0}\frac{1}{\delta}[E(Y|x,h(x+\delta)) - E(Y|x,h(x))] + 0(\sigma_z^2)\\
&= \frac{\partial}{\partial z}E(Y|x,z)\frac{d}{dx}h(x)|_{z=h(x)}
\end{aligned}
$$

and

$$
TE_{inc}(x) = \lim_{\delta\to 0}\frac{1}{\delta}[E(Y|x+\delta) - E(Y|x)] = \frac{d}{dx}E(Y|x)
$$

Using the rule of partial differentiation, we have $TE_{inc} = DE_{inc} + IE_{inc}$ a result obtained in (Winship and Mare, 1983), though starting from a different perspective.

Linear outcome with binary mediator

It is interesting to inquire how effects are decomposed when we retain the linear form of the outcome process, but let the intermediate variable Z be a binary variable that is related to X through an arbitrary nonlinear process $P(Z=1|x)$.

Considering a transition from $X = x_0$ to $X = x_1$ and writing

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

we readily obtain:

$$
\begin{aligned}
DE &= \sum_z [(\alpha x_1 + \beta z + \gamma)] - [(\alpha x_0 + \beta z + \gamma)]P(z|x_0)\\
&= \alpha(x_1 - x_0) \qquad (28)\\
IE &= \sum_z (\alpha x_1 + \beta z + \gamma) - [P(z|x_1) - P(z|x_0)]\\
&= \beta[E(Z_0|x_1) - E(Z|x_0)] \qquad (29)\\
TE &= \sum_z E(Y|x_1,z)P(z|x_1) - E(Y|x_0,z)P(z|x_0)\\
&= \sum_z (\alpha x_1 + \beta z + \gamma)P(z|x_1) - \sum_z (\alpha x_0 + \beta z + \gamma)P(z|x_0)\\
&= \alpha(x_1 - x_0) + \beta[E(Z|x_1) - E(Z|x_0)] \qquad (30)
\end{aligned}
$$

Again, we have $TE = DE + IE$.

We see that as long as the outcome processes is linear, non linearities in the mediation process do not introduce any surprises; effects are decomposed into their direct and indirect components in a textbook-like fashion. Moreover, the distribution $P(Z|x)$ plays no role in the analysis; it is only the expectation $E(Z|x)$ that need be estimated.

This result was obtained by Li et al. (2007) who estimated IE using a difference-in-coefficients strategy. It follows, in fact, from a more general property of the Mediation Formula (13), first noted by VanderWeele (2009), which will be discussed next.

Semi-linear outcome process

Suppose $E(Y|x, z)$ is linear in Z, but not necessarily in X. We can then write

$$\begin{aligned} E(Y|x,z) &\stackrel{\Delta}{=} g(x,z) = f(x) + t(x)z \\ E(Z|x) &\stackrel{\Delta}{=} h(x) \end{aligned}$$

and the Mediation Formulas give (for the transition from $X = x_0$ to $X = x_1$):

$$\begin{aligned} DE &= \sum_z \{(f(x_1) + t(x_1)z) - (f(x_0) - t(x_0)z)\}P(z|x_0) \\ &= f(x_1) - f(x_0) + (t(x_1) - t(x_0))E(Z|x_0) \\ &= g(x_1, h(x_0)) - g(x_0, h(x_0)) \qquad (31) \\ IE &= \sum_z (f(x_0) + t(x_0)z)(P(z|x_1) - P(z|x_0)) \\ &= t(x_0)(E(Z|x_1) - E(Z|x_0)) \\ &= t(x_0)[h(x_1) - h(x_0)] \qquad (32) \\ TE &= \sum_z (f(x_1) + t(x_1)z)P(z|x_1) - \sum_z (f(x_0) + t(x_0)z)P(z|x_0) \\ &= f(x_1) - f(x_0) + t(x_1)E(Z|x_1) - t(x_0)E(Z|x_0) \\ &= g(x_1, h(x_1)) - g(x_0, h(x_0)) \qquad (33) \end{aligned}$$

We see again that only the conditional mean $E(Z|x)$ need enter the estimation of causal effects in this model, not the entire distribution $P(z|x)$. However, the equality $TE = DE + IE$ no longer holds in this case; the nonlinearities embedded in the interaction term $t(x)z$ may render Z an enabler or inhibitor of the direct path, thus violating the additive relationship between the three effect measures.

This becomes more transparent when we examine the standard linear model to which a multiplicative term xz is added, as is done, for example, in the analyses of Kraemer et al. (2008), Jo (2008), and Preacher et al. (2007). In this model we have

$$\begin{aligned} E(Y|x,z) &\triangleq g(x,z)) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz \\ E(Z|x) &\triangleq h(x) = \gamma_0 + \gamma_1 x \end{aligned}$$

Substituting

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x \\ t(x) &= \beta_2 + \beta_3 x \\ h(x) &= \gamma_0 + \gamma_1 x \end{aligned}$$

in (31)–(33) and letting $x_1 - x_0 = 1$, gives

$$DE = \beta_1 + \beta_3(\gamma_0 + \gamma_1 x_0) \tag{34}$$

$$IE = \gamma_1(\beta_2 + \beta_3 x_0) \tag{35}$$

$$TE = \beta_1 + \beta_3\gamma_0 + \beta_2\gamma_1 + \beta_3\gamma_1(x_0 + x_1) \tag{36}$$

In particular the relationships between DE, IE, and TE becomes

$$TE = DE + IE + \gamma_1\beta_3$$

which clearly identifies the product $\gamma_1\beta_3$ as the culprit for the non-additivity $TE \neq TE + IE$. Indeed, when $\gamma_1\beta_3 \neq 0$, Z acts both as a moderator and a mediator, and both DE and IE are affected by the interaction term $\beta_3 xz$. Note further that the direct and indirect effects can both be zero and the total effect non-zero; a familiar nonlinear phenomenon that occurs when Z is a necessary enabler for the effect of X on Y. This dynamics has escaped standard analyses of mediation which focused exclusively on estimating? structural parameters, rather than effect measures, as in (34)–(36).

It is interesting to note that, due to interaction, a direct effect can exist even when β_1 vanishes, though β_1 is the path coefficient associated with the direct link $X \rightarrow Y$. This illustrates that estimating parameters in isolation tells us little about the problem until we understand the way they combine to form effect measures. More generally, mediation and moderation are inextricably intertwined and cannot be assessed separately, a position affirmed by Kraemer et al. (2008) and Preacher et al. (2007).

The binary case

To complete our discussion of models in which the mediation problem lends itself to a simple solution, we now address the case where all variables are binary, still allowing though 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 outcomes are straightforward.

Assume that the model of Fig. 1(a) is valid and that the observed data is given by Table 1. The factors $E(Y|x, z)$ and $P(Z|x)$ in Eqs. (12)–(14) can be readily estimated as shown in the two right-most columns of Table 1 and, when substituted in (12)–(14), yield:

	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.

$$DE = (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0 \tag{37}$$

$$IE = (h_1 - h_0)(g_{01} - g_{00}) \tag{38}$$

$$TE = g_{11}h_1 + g_{10}(1 - h_1) - [g_{01}h_0 + g_{00}(1 - h_0)] \tag{39}$$

We see that logistic or probit regression is not necessary, simple arithmetic operations suffice to provide a general solution for any conceivable dataset.

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. (37)–(39) yields:

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

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.

For completeness, we note that the controlled direct effects are (using (1)):

$$CDE_{z=0} = g_{10} - g_{00} = 0.40 - 0.20 = 0.20$$

and

$$CDE_{z=1} = g_{11} - g_{01} = 0.80 - 0.30 = 0.50$$

which are quite far apart. Their weighted average, governed by $P(Z = 1|X = 0) = h_0 = 0.40$, gives us $DE = 0.32$. These do not enter, however, into the calculation of IE, since the indirect effect cannot be based on controlling variables; it requires instead a path-deactivating operator, as mirrored in the definition of Eq. (6).

3 Relation to Other Methods

3.1 Methods based on differences and products

Attempts to compare these results to those produced by conventional mediation analyses encounter two obstacles. First, conventional methods do not define direct and indirect effects in a nonparametric setting, without committing to specific functional or distributional forms. 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. Section 2 demonstrates that this strategy is not compatible with the causal interpretation of effect measures, even when the parameters are known precisely; IE and DE may be extremely complicated functions of those regression coefficients (see Eqs. (25–26)). Fortunately, those coefficients need not be estimated at all; effect measures can be estimated directly from the data, circumventing the parametric representation altogether.

Second, 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 slope of Z on X

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

by the slope of Y on Z fixing X,

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

but does not specify at what value we should fix X. Equation (38) resolves this ambiguity by determining that X should be fixed to $X = 0$; 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 (37) 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 $h_0 = P(Z = 1|X = 0)$ as the weighting function.

To summarize, 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.

Reiterating the discussion of Section 2, 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.

3.2 Relation to Principal-Strata Direct Effect

The derivation of the Mediation Formula (Pearl, 2001) was made possible by the counterfactual interpretation of structural equations (see footnote 1) and the symbiosis between graphical and counterfactual analysis that this interpretation engenders.[11] In contrast, the structure-less approach of Rubin (1974) has spawned other definitions of direct effects, normally referred to as "principal-strata direct effect (PSDE)" (Frangakis and Rubin, 2002; Mealli and Rubin, 2003; Rubin, 2004, 2005; Egleston et al., 2010). Whereas the natural direct effect measures the average effect that would be transmitted in the population with all mediating paths (hypothetically) *deactivated*, the PSDE is defined as the effect transmitted in those units only for whom mediating paths *happened to be deactivated* in the study. This definition leads to unintended results that stand contrary to common usage of direct effects (Robins et al., 2007, 2009; VanderWeele, 2008), excluding from the analysis all individuals who are both directly and indirectly affected by the causal variable X (Pearl, 2009b). In linear models, as a striking example, a direct effect will be flatly undefined, unless β, the $X \rightarrow Z$ coefficient is zero. In some other cases, the direct effect of the treatment will be deemed to be nil, if a small subpopulation

[11] Such symbiosis is now standard in epidemiology research (Robins, 2001; Petersen et al., 2006; VanderWeele and Robins, 2007; Hafeman and Schwartz, 2009; VanderWeele, 2009; Albert and Nelson, 2011) and is making its way slowly toward the social and behavioral sciences, (e.g., Morgan and Winship (2007); Imai et al. (2010a); Elwert and Winship (2010); Chalak and White (2011)), despite islands of resistance (Wilkinson et al., 1999, p. 600; Sobel, 2008; Rubin, 2010; Imbens, 2010).

exists for which treatment has no effect on both Y and Z. In view of these definitional inadequacies we do not include "principal-strata direct effect" in our discussion of mediation, though they may well be suited for other applications[12], for example, when a stratum-specific property is genuinely at the focus of one's research.

Indeed, taking a "principal strata" perscpective, Rubin found the concept of mediation "ill-defined." In his words: "The general theme here is that the concepts of direct and indirect causal effects are generally ill-defined and often more deceptive than helpful to clear statistical thinking in real, as opposed to artificial problems" (Rubin, 2004). Conversely, attempts to define and understand mediation using the notion of "principal-strata direct effect" have encountered basic conceptual difficulties (Lauritzen, 2004; Robins et al., 2007, 2009; Pearl, 2009b), concluding that "it is not always clear that knowing about the presence of principal stratification effects will be of particular use" (VanderWeele, 2008). As a result, it is becoming widely recognized that the controlled, natural and indirect effects discussed in this paper are of greater interest, both for the purposes of making treatment decisions and for the purposes of explanation and identifying causal mechanisms (Joffe et al., 2007; Albert and Nelson, 2011; Mortensen et al., 2009; Imai et al., 2010a; Geneletti, 2007; Robins et al., 2007, 2009; Petersen et al., 2006; Hafeman and Schwartz, 2009; Kaufman, 2010; Cai et al., 2008).

The limitation of PSDE stems not from the notion of "principal-strata" per se, which is merely a classification of units into homogeneously reacting classes, and has been used advantageously by many researchers (Balke and Pearl, 1994a,b; Pearl, 1993; Balke and Pearl, 1997; Heckerman and Shachter, 1995; Pearl, 2000, p. 264; Lauritzen, 2004; Sjölander, 2009). Rather, the limitation results from strict adherence to an orthodox philosophy which prohibits one from regarding a mediator as a cause unless it is manipulable. This prohibition prevents one from defining the direct effect as it is commonly used in decision making and scientific discourse – an effect transmitted once all mediating paths are "deactivated" (Pearl, 2001; Avin et al., 2005; Albert and Nelson, 2011), and forces one to use statistical conditionalization instead. Path deactivation requires counterfactual constructs in which the mediator acts as an antecedent, as in Eqs. (1), (2) and (6), regardless of whether it is physically manipulable. After all, if our aim is to uncover causal mechanisms, it is hard to accept the PSDE restriction that nature's pathways should depend on whether we have

[12] Joffe and Green (2009) and Pearl and Bareinboim (2011) examine the adequacy of the "principal-strata" definition of surrogate outcomes; a notion related, though not identical to mediation. There, too, the restrictions imposed by the "principal-strata" framework lead to surrogacy criteria that are incompatible with the practical aims of surrogacy (see Pearl (2011)).

the technology to manipulate one variable or another. (See (Pearl, 2011) for in depth discussion of these issues.)

4 Conclusions

Traditional methods of mediation analysis produce distorted estimates of "mediation effects" when applied to nonlinear models or models with categorical variables. By focusing on parameters of logistic and probit estimators, instead of the target effect measures themselves, traditional methods produce consistent estimates of the former and biased estimates of the latter. This paper demonstrates that the bias can be substantial even in simple systems with all processes correctly parameterized, and only the outcome dichotomized. The 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 and effective estimation methods through the Mediation Formula.

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. For example, it would be straightforward to investigate the distortion created by assuming logistic model (as in (23)) when data is generated in fact by a probit distribution, or vice versa. This exercise would amount to finding the maximum-likelihood (ML) estimates of γ_0, γ_{xy}, and γ_{zy} in (24) for data generated by a probit distribution and compare the estimated effect measures computed through (25)–(27) with the true values of those measures, as dictated by the probit model.[13] This type of analytical "sensitivity analysis" has been used extensively in statistics for parameter estimation, but could not be adequately applied to mediation analysis, owed to the absence of an objective target quantity that captures the notion of indirect effect in nonlinear systems. MacKinnon et al. (2007b) for example evaluated sensitivity to misspecifications by comparing the estimated parameters against their true values, though disparities in parameters may not represent disparity in effect measures (i.e., ED or IE). By providing such objective measures of effects, the Mediation Formula of Eq. (13) enables us to measure directly the disparities

[13] An alternative would be to find the ML estimates of DE, IE, and TE directly, through (25), (26) and (27), rather than going through (23) (van der Laan and Rubin, 2006).

in the target quantities.[14]

While the validity of the Mediation Formulas rests on the same assumptions (i.e., no unmeasured confounders) that are standard requirement in linear mediation analysis, their appeal to general 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 social, behavioral and health-related sciences.

References

Albert, J. M. and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics* 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.

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*. Morgan-Kaufmann Publishers, Edinburgh, UK.

Balke, A. and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence 10* (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 46–54.

Balke, A. and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I. MIT Press, Menlo Park, CA, 230–237.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92** 1172–1176.

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.

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

[14] Sensitivity analysis using both analytical and simulation techniques are described in Imai et al. (2010a).

Cai, Z., Kuroki, M., Pearl, J. and Tian, J. (2008). Bounds on direct effect in the presence of confounded intermediate variables. *Biometrics* **64** 695–701.

Chalak, K. and White, H. (2011). Direct and extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics* **44** 1–51.

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

Egleston, B. L., Cropsey, K. L., Lazev, A. B. and Heckman, C. J. (2010). A tutorial on principal stratification-based sensitivity analysis: application to smoking cessation studies. *Clinical Trials* **7** 286–298.

Elwert, F. and Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (R. Dechter, H. Geffner and J. Halpern, eds.). College Publications, UK, 327–336.

Fleming, T. and DeMets, D. (1996). Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine* **125** 605–613.

Frangakis, C. and Rubin, D. (2002). Principal stratification in causal inference. *Biometrics* **1** 21–29.

Freedman, L., Graubard, B. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **8** 167–178.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B (Methodological)* **69** 199–215.

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

Graff, J. and Schmidt, P. (1982). A general model for decomposition of effects. In *Systems Under Indirect Observation: Causality, Structure, Prediction* (K. Jöreskog and H. Wold, eds.). North-Holland, Amsterdam, 131–148.

Green, D., Ha, S. and Bullock, J. (2010). Enough already about black box experiments: Studying mediation is more difficult than most scholars suppose. *Annals of the American Academy of Political and Social Science* **628** 200–208.

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

Heckerman, D. and Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research* **3** 405–430.

Holland, P. (1995). Some reflections on Freedman's critiques. *Foundations of Science* **1** 50–57.

Imai, K., Keele, L. and Tingley, D. (2010a). A general approach to causal mediation analysis. Tech. rep., Department of Politics, Princeton University.

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

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

Imbens, G. (2010). An economists perspective on of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods* **15** 47–55.

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.

Judd, C. and Kenny, D. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* **5** 602–619.

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.

Lauritzen, S. (2004). Discussion on causality. *Scandinavian Journal of Statistics* **31** 189–192.

Li, Y., Schneider, J. and Bennett, D. (2007). Estimation of the mediation effect with a binary mediator. *Statistics in Medicine* **26** 3398–3414.

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

MacKinnon, D. and Dwyer, J. (1993). Estimating mediated effects in prevention studies. *Evaluation Review* **4** 144–158.

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.

Mealli, F. and Rubin, D. (2003). Assumptions allowing the estimation of direct causal effects. *Journal of Econometrics* **112** 79–87.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T. and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* **23** 607–625.

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.

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** 2629–2635.

Muller, D., Judd, C. and Yzerbyt, V. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology* **89** 852–863.

Pearl, J. (1993). Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*. Tome LV, Book 1, Florence, Italy.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

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

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

Pearl, J. (2009b). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146, <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

Pearl, J. (2010). 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. (2011). Principal stratification a goal or a tool? Tech. Rep. R-382, <http://ftp.cs.ucla.edu/pub/stat_ser/r382.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, with discussion, the *International Journal of Biostatistics*.

Pearl, J. and Bareinboim, E. (2011). Transportability across studies: A formal approach. Tech. Rep. R-372, <http://ftp.cs.ucla.edu/pub/stat_ser/r372.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Pearl, J. and Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 444–453.

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

Preacher, K., Rucker, D. and Hayes, A. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research* **28** 185–227.

Prentice, R. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8** 431–440.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling* **7** 1393–1512.

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

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

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

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

Robins, J., Richardson, T. and Spirtes, P. (2009). On identification and inference for direct effects. Tech. rep., Harvard University, MA.

Robins, J., Rotnitzky, A. and Vansteelandt, S. (2007). Discussion of principal stratification designs to estimate input data missing due to death. *Biometrics* **63** 650–654.

Rosenbaum, P. and Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.

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.

Rubin, D. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods* **15** 38–46.

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

Shrout, P. and Bolger, N. (2002). Mediation in experimental and non-experimental studies: New procedures and recommendations. *Psychological Methods* **7** 422–445.

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

Sobel, M. (1987). Direct and indirect effects in linear structural equation models. *Sociological Methods & Research* **16** 1155–176.

Sobel, M. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* **33** 230–231.

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

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2** Article 11. Available at: http://www.bepress.com/ijb/vol2/iss1/11.

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.

Vansteelandt, S. (2011). Estimation of direct and indirect effects. In *Causal Inference: Statistical Perspectives and Applications* (C. Berzuini, P. Dawid and L. Bernardinelli, eds.). Wiley and Sons.

Wilkinson, L., the Task Force on Statistical Inference and *APA Board of Scientific Affairs* (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* **54** 594–604.

Winship, C. and Mare, R. (1983). Structural equations and path analysis for discrete data. *The American Journal of Sociology* **89** 54–110.