Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2009;**38**:838–845

Advance Access publication 4 March 2009
doi:10.1093/ije/dyn372

Opening the Black Box: a motivation for the assessment of mediation

Danella M Hafeman and Sharon Schwartz*

Accepted 18 December 2008

Recent criticism of epidemiologic methods has focused on the limitations of 'black box' epidemiology, a pejorative label given to the simple identification of exposure–disease relationships. The assessment of mediation is an important tool for addressing this criticism. By using mediation analysis to open the black box, underlying mechanisms of the observed associations can be described and causal inference improved. An explicit theoretical motivation for such an analysis has been missing from the epidemiological literature. To provide this motivation, we integrate literature from epidemiology and other social sciences to describe the reasons that an investigator might want to assess mediation. We then describe the connections between these reasons and specific measures of indirect and direct effects that have been previously described.

Keywords Mediation, direct and indirect effects, counterfactual, causal inference, epidemiologic methods

Introduction

Since the Second World War, epidemiologists have largely focused on the identification of risk factors for various diseases, using an approach that has been pejoratively labelled as 'black box' epidemiology. While many discoveries in epidemiology have been made using this approach, it has limitations. Some critics cite the many associations that have been reported in one study, only to be refuted by another.[1] Others have pointed out that the 'black box' approach leads to the identification of a list of risk factors, but not an explanatory theory for how disease arises. In the absence of such theory, the ability to effectively predict the effect of interventions is limited.[2]

Based on these critiques, some authors have argued that we should discard the 'black box' approach altogether, in favour of a systems-based strategy.[3] However, methods in epidemiology are extremely well developed for the identification of risk factors, while methods for a systems-based approach are still in their infancy. Thus other authors have suggested something slightly less radical: that, instead of discarding the black box, we use existing methods to open it.[4,5]

We propose that the assessment of mediation is one very practical way to open the black box, leading to improved causal inference. Mediation is defined as the totality of processes that explain an observed relationship between exposure and disease.[6] By testing mediational hypotheses, investigators can, in a sense, view the inner workings of this box. Of course, a single mediator cannot fully explicate the observed relationship between exposure and disease. However, the assessment of mediation can be thought of as a useful first step to addressing the limitations of risk factor epidemiology.

Indeed, in other disciplines, mediation is regularly assessed as a way to test the mechanisms by which exposure, as measured, leads to the outcome. For example, the assessment of mediation is a crucial element of statistical analysis in psychology. The seminal paper on the assessment of mediation in psychology, written by Baron and Kenny in 1986,[7] has been referenced over 8500 times (Social Science Citation Index). Mediation also plays a central role in the classic validity scheme described by Shadish *et al.*,[8] in the context of 'causal explanation'; that is, explaining why and how an exposure has the effect that it does. By probing the mechanisms that explain

Department of Epidemiology, Mailman School of Public Health, Columbia University, USA.
* Corresponding author. E-mail: sbs5@columbia.edu

a relationship between exposure and disease, causal hypotheses can be more rigorously tested and etiologic relationships better understood.[9–11] Such knowledge is often a prerequisite for the successful assessment of public health interventions, because they help us to more effectively test the interventions and better predict the circumstances under which they will work.

The concept of causal explanation was discussed in Susser's 1973 textbook,[12] but has largely been ignored in the epidemiologic textbooks since that time (with some recent exceptions[13]). In fact, the dominant concept of mediation (or the separation of direct and indirect effects) in epidemiologic literature is not based on the exploration of mechanism, but rather additional intervention (on the mediator). For example, the blocked direct effect is defined as the effect of exposure if the mediator were blocked for every individual in the population.[14] While this answers a specific interventionist question (what would happen if we could intervene on both the exposure and the mediator?), it does not specifically assess the mechanisms by which exposure is causing disease; that is, it does not address the questions of causal explanation.

In this article, we integrate the concept of 'causal explanation' from the psychology literature with the extant epidemiology and statistics literature on mediation. Through an incorporation of this literature, we describe the reasons for conducting a mediation analysis in epidemiology. We next give several operational definitions of indirect and direct effects, and show how the correct definition depends on the particular reason for mediation analysis. By providing a conceptual foundation for the assessment of mediation, we hope to illustrate why opening the black box is important, and how mediation analysis represents an important tool for opening this box.

To date, most of the epidemiological literature on mediation has focused on the problems that arise when assessing direct and indirect effects from the data.[15–17] As with any epidemiologic analysis, care must be taken when making causal inference based on observational data. For example, when assessing mediation, the possibility of confounding of the mediator–disease relationship must be considered, even in the context of a randomized trial.[14,15,18] The focus of this article is to establish a conceptual foundation for the assessment of mediation, thus making it worthwhile to solve the associated methodological challenges.

Motivation for mediation analysis

Based on an integration of the relevant psychology and epidemiology literature, we propose three reasons why an epidemiologist might want to open the 'black box' using mediation (summarized in the first column of Table 1).

Table 1 Why assess mediation: three reasons for assessing mediation, and their corresponding effect measures

Reason	Effects of 1° interest
1) Strengthen main effect hypothesis	TIE, PIE
2) Test pathway specific hypotheses	TIE, PDE
3) Evaluate and improve intervention	PIE

Strengthen evidence that the main effect is causal

The identification of a hypothesized mediator can provide evidence for a causal interpretation of the observed relationship between exposure and disease. First, the specified mediator demonstrates the pathway by which exposure causes disease, and thus leads to increased biological plausibility of the theory that is being tested. For example, the Endogenous Hormones Group[19] found an association between BMI and breast cancer that was fully mediated by free estradiol levels. The latter finding increases the biological plausibility of the theory that high BMI causes breast cancer, because the theoretical pathway was successfully tested.

Second, the identification of a hypothesized mediator decreases the likelihood that a given finding is spurious or caused by confounding of the main effect.[13] For instance, Carney *et al*.[20] found that the observed relationship between depression and mortality after myocardial infarction (MI) was partially mediated by low heart rate variability, an indicator of abnormal cardiac autonomic function. This finding decreased the likelihood that the observed association (at least the mediated portion) was spurious or due to a common cause of depression and death from MI. Specifically, a common cause of depression and MI could explain the main effect. But to be a viable explanation for the observed results, a confounder would not only have to cause both depression and mortality after MI, but also low heart rate variability. Of course, there are always alternative explanations for patterns of observations and we cannot rule out these possibilities completely. However, by making such alternative explanations less likely, this mediation analysis provides evidence that the main effect is causal.

This logic can be viewed as a modification of Pearl's forward identification of causal effects, as noted by Winship and Harding.[21–23] Sometimes the main effect is not identifiable by standard methods, because of an unmeasured common cause of the exposure and disease. In this case, we cannot interpret the observed relationship between exposure and outcome as causal. However, if a mediator has been measured, it is possible that this effect could be identifiable by the front-door criterion. That is, the quantity of interest would simply be the indirect effect (through the mediator of interest). To provide a valid quantification of the exposure–disease relationship, the measured mediator must intercept all directed pathways from exposure to disease; that is, it must be a full mediator of the exposure–disease relationship. But even if this stringent requirement is not met, the demonstration of a hypothesized indirect effect strengthens evidence that the main effect is, at least partly, causal.

Test a pathway-specific hypothesis

Often the primary hypothesis of interest is not whether there is an association between the exposure, as measured, and the disease. Instead, investigators want to know whether a specific mechanism of exposure is associated with the outcome. There are two ways that this question can be addressed.

First, many research hypotheses focus on explaining an association that has been observed many times, but is poorly understood. For example, it has been repeatedly demonstrated that socioeconomic status (SES) is a predictor of many health outcomes. However, the debate in this field centres on the mechanisms that explain this effect. Interesting questions include the following: (i) Is the relationship between SES and a given health outcome completely explained by known risk factors, or is there a 'direct' effect of SES on this outcome?[24,25] and (ii) is the relationship between SES and a given health outcome due to material factors or psychosocial and autonomic factors (or both)?[26–28] The main goal of these investigators is not simply to document the SES-related disparities in health outcomes, but rather to explain how they arise; mediation analysis is a core tool for answering such questions.

Second, researchers are often not interested in the total effect of the measured exposure on the outcome, but rather the effect of a particular aspect of exposure: the hypothesized 'active ingredient'. However, this ingredient might be difficult to measure, or not be available in a given dataset. In this case, a mediation analysis can be used to statistically determine the effect of interest, by adjusting for the aspects of exposure not included in the theory.[29] For example, Salas-Salvado *et al*.[30] tested the hypothesis that certain nutrients in the Mediterranean diet decrease markers of inflammation. While they measured the total association between Mediterranean diet and inflammatory markers, the investigators were not interested in certain mechanisms through which this diet might influence these markers, such as decreased BMI or reduced incidence of hypertension. Thus the authors adjusted for these variables in the analysis, leaving (hopefully) the 'direct' anti-inflammatory effects.

Such motivations are often implicit. For example, Salas-Salvado *et al*. did not state their reasons for adjusting for these probable consequences of exposure. While adjustment makes sense in some cases, it should be carefully considered and justified. Just as the validity of main effects depends on the assumption of no unmeasured confounding between the exposure and the disease, the validity of mediation analysis depends on the assumption of no unmeasured confounding of the mediator–disease relationship. Thus, when assessing direct effects, it is necessary to measure and adjust for common causes of the mediator and disease.[14,15]

Evaluate and improve an intervention

Mediation analysis can be used to evaluate whether an intervention is working the way we expect it to; in other words, we want to know whether the treatment causes (or prevents) the outcome through the proposed mediator(s). Assuming that the mediator is perfectly measured, there are two possible reasons that mediation would not be observed. First, it is possible that the treatment changes the mediator, but that the mediator has no effect on the outcome. In this case, future interventions might focus less on components that impact this mediator, because it is not an important predictor of the outcome. Second, it is possible that a mediator has an effect on the outcome status, but that the treatment does not change the mediator. In this case, we might want to focus efforts on improving the component of the program that is supposed to change this particular mediator.

For example, Bernat *et al*.[31] found that the 'Early Risers' program, a behavioural intervention for at-risk grade school students, prevented oppositional defiant disorder (ODD) symptoms. They hypothesized that the effects of this program would be mediated by three intermediate variables: increased social skills, more effective discipline and academic achievement. The authors found that the program's effects were mediated by the former two constructs, but not the latter. Although the program led to changes in academic achievement, academic achievement did not decrease ODD symptoms. Based on this analysis, future interventions might more efficiently impact the outcome of interest (decreased ODD symptoms) by focusing on the other two components.

We have presented three compelling and practical reasons for the assessment of mediation. Underlying all of these reasons for mediation analysis is the benefit of testing a more complex and explanatory theory, and thus gaining a deeper understanding of the particular exposure–disease relationship.[8,13,32] However, the differences among them are also important since each corresponds to a particular definition of the indirect (or direct) effect. Explicating the reasons for mediation analysis therefore also illuminates the conceptual differences among these definitions

Mediational effect definitions

When assessing mediation, the investigator might be interested in either the indirect or direct effect.

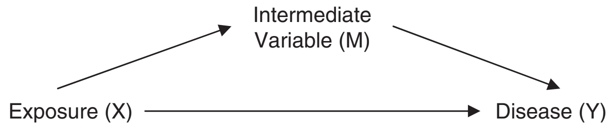

Figure 1 Simple diagram for mediation

The indirect effect is characterized by the pathway that goes through the measured intermediate variable ($X \rightarrow M \rightarrow$Y), while the direct effect does not go through the intermediate variable ($X \rightarrow Y$) (Figure 1). In all cases, when we say 'direct' effect, we mean that the effect is direct relative to measured variables; typically, there are other (unmeasured) intermediate variables that would mediate the direct effect. Throughout this paper, we assume that exposure (X), mediator (M) and outcome (Y) are dichotomous variables.

Two classes of direct and indirect effects have been described in the epidemiologic literature: 'natural' and 'controlled' effects. 'Natural' effects are descriptive. They partition the total effect according to naturally occurring values of the mediator, preserving the observed relationship between exposure and mediator.[33] In contrast, 'controlled' effects are prescriptive. They are defined based on the effect of exposure if the mediator were fixed at a particular value. Controlled effects do not preserve the relationship between exposure and mediator, but rather set the value for the mediator according to a postulated intervention.[14,33,34]

Controlled effects are of interest if the focus of inquiry is to determine the effect of exposure on disease, given an intervention that universally blocks (or assigns) the mediator. In contrast, natural effects are of interest if the investigator wants to explore the role of individual pathways in an observed relationship.[14,35] The reasons for mediation analysis discussed above, grounded in a theory of 'causal explanation', correspond to the latter intention; thus we focus our attention on natural effects.

The natural effects are further subdivided into 'pure' and 'total' effects, distinguished by the differential inclusion of interaction between exposure and mediator.[14,18] In the presence of statistical interaction between the exposure and mediator, the pure and total indirect (or direct) effects differ. The choice of which natural effect to use (pure or total) depends on the reason for mediation analysis. To link the natural effects to the reasons described, we introduce a counterfactual framework to describe their causal meaning.

Counterfactual framework for mediation

Mediation can be described as a two-stage process: (i) the M-stage includes processes that cause the mediator and (ii) the Y-stage includes processes that cause the disease.[18] Two relevant potential outcomes for the M-stage of mediation can be defined: (i) mediator status if a given individual were exposed ($M_1 = 0$ or $M_1 = 1$) and (ii) mediator status if a given individual were unexposed ($M_0 = 0$ or $M_0 = 1$). The notation M_{x} denotes mediator status for a particular individual ($M = 0$ or $M = 1$) that would be observed if exposure were assigned to be present ($X = 1$) or absent ($X = 0$).

Table 2 Natural indirect and direct effects

Natural effect	Potential outcomes
Pure direct effect (PDE)	$\mathrm{P}(Y_{1M_0} = 1) - \mathrm{P}(Y_{0M_0} = 1)$
Pure indirect effect (PIE)	$\mathrm{P}(Y_{0M_1} = 1) - \mathrm{P}(Y_{0M_0} = 1)$
Total direct effect (TDE)	$\mathrm{P}(Y_{1M_1} = 1) - \mathrm{P}(Y_{0M_1} = 1)$
Total indirect effect (TIE)	$\mathrm{P}(Y_{1M_1} = 1) - \mathrm{P}(Y_{1M_0} = 1)$

Mediational effects can be defined in terms of potential outcomes.

There are four relevant potential outcomes that can be defined for the Y-stage, depending on the presence of exposure ($X = 1$, $X = 0$) and the presence of the mediator ($M = 1$, $M = 0$). The potential outcomes take the general form Y_{xm}, where Y_{xm} is the potential outcome that would be observed for a particular individual under a defined exposure ($X = \mathrm{x}$) and mediator ($M = \mathrm{m}$) condition. For example, Y_{10} is the disease status ($Y_{10} = 1$ or $Y_{10} = 0$) that would be observed if an individual were assigned to be exposed ($X = 1$) and mediator-negative ($M = 0$).

Finally, the M-stage and the Y-stage can be combined to produce compound potential outcomes. The compound potential outcomes, proposed by Pearl[33] (with notation from Petersen *et al.*[35]), represent an integration of the M-stage and the Y-stage. There are four compound potential outcomes, defined by (i) exposure status ($X = 1$, $X = 0$) and (ii) mediator given a specified exposure status (M_1, M_0). For example, Y_{0M_0} is the disease status (present or absent) that would be observed if an individual were unexposed ($X = 0$) and had the mediator that he or she would have if unexposed ($M_0 = 0$ or $M_0 = 1$); this is equivalent to the outcome that would be observed if this individual were unexposed ($Y_{0M_0} = Y_0$). Y_{1M_0}, on the other hand, is the disease status (present or absent) that would be observed if an individual were exposed ($X = 1$), but had the mediator that he or she would have had if unexposed ($M_0 = 1$ or $M_0 = 0$). While the latter potential outcome is more difficult to estimate from observed data, it is crucial for defining indirect and direct effects.

Using these potential outcomes, the causal meaning of previously described mediational effects can be described (summarized in Table 2). There are two natural indirect effects, referred to as the pure indirect effect (PIE) and total indirect effect (TIE). The pure indirect effect (PIE) is the effect of exposure if its only action were to cause the mediator. This would be a comparison between (i) the risk of disease if everyone were unexposed, but had the mediator they would have had were they exposed $[\mathrm{P}(Y_{0M_1}=1)]$ and (ii) the risk if everyone were unexposed $[\mathrm{P}(Y_{0M_0}=1)]$. The total indirect effect (TIE) is the effect of exposure due to the fact that it causes the mediator. This would be the difference between (i) the risk if everyone were exposed $[\mathrm{P}(Y_{1M_1}=1)]$ and (ii) the risk if everyone were exposed, but had the mediator they would have had were they unexposed $[\mathrm{P}(Y_{1M_0}=1)]$.

In parallel, there are two natural direct effects: the pure direct effect (PDE) and total direct effect (TDE). The pure direct effect (PDE) is the effect that the exposure would have if exposure did not cause the mediator. This would be operationalized as a comparison between (i) the risk of disease if everyone were exposed, but had the mediator they would have had were they unexposed $[\mathrm{P}(Y_{1M_0}=1)]$ and (ii) the risk of disease if everyone were unexposed $[\mathrm{P}(Y_{0M_0}=1)]$. The total direct effect (TDE) is the effect that the exposure would have if lack of exposure did not prevent the mediator. The TDE would be the difference between (i) the risk of disease if everyone were exposed $[\mathrm{P}(Y_{1M_1}=1)]$ and (ii) the risk of disease if everyone were unexposed, but had the mediator they would have were they exposed $[\mathrm{P}(Y_{0M_1}=1)]$.

Under the condition of perfect additivity, the pure and total indirect effects will be equal (PIE = TIE); similarly, the pure and total direct effects will be equal (PDE = TDE). This is because, given perfect additivity, the effect through the mediator is the same in the exposed $[\mathrm{P}(Y_{1M_1}=1)-\mathrm{P}(Y_{1M_0}=1)]$ and the unexposed $[\mathrm{P}(Y_{0M_1}=1)-\mathrm{P}(Y_{0M_0}=1)]$. Similarly, the direct effect of exposure is the same regardless of mediator status. However, in the absence of perfect additivity, these equalities will not hold. When these effects diverge (i.e. PDE $\neq$ TDE, PIE $\neq$ TIE), the investigator must decide which quantity to estimate. This decision will depend, in part, on the reason for conducting a given mediation analysis.

Assessment of mediational effects

Methods for estimating the pure and total effects based on the observed data have been discussed previously.[18,33,35] Briefly, the pure and total effects can be calculated by estimating the following quantities: $[\mathrm{P}(Y_{0M_0}=1)]$, $[\mathrm{P}(Y_{1M_1}=1)]$, $[\mathrm{P}(Y_{0M_1}=1)]$ and $[\mathrm{P}(Y_{1M_0}=1)]$. The risk of disease if everyone were unexposed and had the mediator they would have if unexposed $[\mathrm{P}(Y_{0M_0}=1)]$ can be estimated by the observed risk in the unexposed $[\mathrm{P}(Y=1|X=0)]$. Similarly, the risk of the outcome if everyone were exposed and had the mediator they would have if exposed $[\mathrm{P}(Y_{1M_1}=1)]$ can be estimated by the observed risk in the exposed $[\mathrm{P}(Y=1|X=1)]$.

The risk of disease if everyone were unexposed but had the mediator that they would have had were they exposed $[\mathrm{P}(Y_{0M_1}=1)]$ is estimated as a weighted average of the observed risk in the unexposed, mediator-positive $[\mathrm{P}(Y=1|X=0, M=1)]$ and the risk in the unexposed, mediator-negative $[\mathrm{P}(Y=1|X=0, M=0)]$; the weights are based on the probability of the mediator in the exposed $\mathrm{P}(M=1|X=1)$.[33]

$$\mathrm{P}(Y_{0M_1}=1)=\mathrm{P}(M=1|X=1)\times\mathrm{P}(Y=1|X=0,M=1)$$
$$+\,\mathrm{P}(M=0|X=1)\times\mathrm{P}(Y=1|X=0,M=0)$$

Analogously, the risk of disease if everyone were exposed but had the mediator they would have had were they unexposed $[\mathrm{P}(Y_{1M_0}=1)]$ is estimated as a weighted average of the observed risk in the exposed, mediator-positive $[\mathrm{P}(Y=1|X=1, M=1)]$ and the risk in the exposed, mediator-negative $[\mathrm{P}(Y=1|X=1, M=0)]$; the weights are based on the probability of the mediator in the unexposed $\mathrm{P}(M=1|X=0)$.[33]

$$\mathrm{P}(Y_{1M_0}=1)=\mathrm{P}(M=1|X=0)\times\mathrm{P}(Y=1|X=1,M=1)$$
$$+\,\mathrm{P}(M=0|X=0)\times\mathrm{P}(Y=1|X=1,M=0)$$

The estimates of the pure and direct effects, based upon the above approximations, are given in Table 3. Note that these estimates can be made conditional on measured confounders of the exposure–disease or mediator–disease relationship. To adjust for a common cause of the mediator and disease that is also a consequence of exposure, more sophisticated methods are necessary.[14,36–38] For unbiased estimation of direct and indirect effects, we must assume that there is no unmeasured confounding of the mediator-disease relationship.[14,15,18]

Mediational effects of interest

We now describe the direct and/or indirect effect(s) that most directly correspond to each of the reasons

Table 3 Effect estimates

Effect	Estimated quantity
PDE	$\mathrm{P}(M=1\|X=0)\times\mathrm{P}(Y=1\|X=1, M=1)+\mathrm{P}(M=0\|X=0)\times\mathrm{P}(Y=1\|X=1, M=0)-\mathrm{P}(Y=1\|X=0)$
PIE	$\mathrm{P}(M=1\|X=1)\times\mathrm{P}(Y=1\|X=0, M=1)+\mathrm{P}(M=0\|X=1)\times\mathrm{P}(Y=1\|X=0, M=0)-\mathrm{P}(Y=1\|X=0)$
TDE	$\mathrm{P}(Y=1\|X=1)-[\mathrm{P}(M=1\|X=1)\times\mathrm{P}(Y=1\|X=0, M=1)+\mathrm{P}(M=0\|X=1)\times\mathrm{P}(Y=1\|X=0, M=0)]$
TIE	$\mathrm{P}(Y=1\|X=1)-[\mathrm{P}(M=1\|X=0)\times\mathrm{P}(Y=1\|X=1, M=1)+\mathrm{P}(M=0\|X=0)\times\mathrm{P}(Y=1\|X=1, M=0)]$

Natural effects can be estimated based on observable proportions, given the assumptions discussed in the text.

for mediation analysis described above. Results are summarized in Table 1.

Strengthen evidence that the main effect is causal

The identification of a hypothesized mediator decreases the likelihood that the association between exposure and outcome is completely explained by confounding. In this case, the crucial question is whether such an indirect effect exists; the magnitude of the indirect effect is much less important. Generally, the PIE and TIE will lead to the same qualitative conclusions, regarding whether or not an indirect effect exists.[18] Thus the decision concerning which particular indirect effect is of interest, PIE vs TIE, becomes much less relevant.

That said, we still might consider which indirect measure will best reflect the quantity of interest: the effect of exposure that is due to the measured pathway (and not due to a confounder of the exposure–disease relationship). Both the TIE and PIE represent this quantity of interest, under slightly different circumstances. The PIE is the effect of exposure explained through the mediator, if all variables (including any confounder of the exposure–disease relationship) were at their unexposed values. The TIE, in contrast, is the effect of exposure explained through the mediator, if all variables (including any confounder of the exposure–disease relationship) were at their exposed values. If we want to identify the indirect effect of exposure in the exposed population, the TIE is thus the effect of interest.

Test a pathway-specific hypothesis

Depending on whether the measured intermediate variable is on the pathway of interest, a path-specific hypothesis can be tested by either quantifying the indirect or direct effect. When the indirect effect is of interest, the investigator generally wants to quantify the effect of exposure that is explained by the pathway through the intermediate variable. This quantity is equivalent to the effect of exposure that would be prevented if the exposure did not cause the mediator, which is the difference between the total effect of exposure $[\mathrm{P}(Y_{1M_1}=1)-\mathrm{P}(Y_{0M_0}=1)]$ and the effect of exposure if it did not cause the mediator $[\mathrm{P}(Y_{1M_0}=1)-\mathrm{P}(Y_{0M_0}=1)]$. This difference is equivalent to the TIE $[\mathrm{P}(Y_{1M_1}=1)-\mathrm{P}(Y_{1M_0}=1)]$; thus the TIE quantifies the desired effect.

When the direct effect is of interest, the investigator generally wants to know the effect of exposure if the exposure did not cause the mediator. This effect is best quantified by the PDE, a comparison between (i) the risk in the exposed, if they had the mediator they would have had were they unexposed $[\mathrm{P}(Y_{1M_0}=1)]$ and (ii) the unexposed $[\mathrm{P}(Y_{0M_0}=1)]$.

Evaluate and improve an intervention

An intervention often consists of many components, each of which impact different intermediate variables. Given an observed treatment effect, the investigator might want to know which components of the intervention are 'active ingredients'; that is, which components explain the observed relationship between treatment and outcome. The purpose of a mediation analysis is to assess the impact of each putative 'active ingredient', given in isolation. In other words, we attempt to determine the effect that the intervention would have had if its only action were to impact a particular mediator. This effect is quantified by the PIE: a comparison between (i) the risk in the unexposed, if they had the mediator they would have had were they exposed $[\mathrm{P}(Y_{0M_1}=1)]$ and (ii) the unexposed $[\mathrm{P}(Y_{0M_0}=1)]$.

Mediation analysis can also be used to evaluate the reasons why a tested intervention did not produce a hypothesized outcome. First, a treatment might fail to impact a hypothesized mediator, reflecting the fact that the relevant component has not been effectively implemented. In response, attempts might be made to improve this component. Second, a hypothesized mediator might fail to impact the desired outcome; to make the intervention more efficient, the relevant component might be eliminated. To evaluate and improve a treatment, it is thus crucial to quantify not only the PIE, but also its parts: (i) the effect of treatment on the mediator and (ii) the effect of the mediator on the outcome.

Conclusions

The assessment of mediation represents an important way to address the critiques of 'black box' epidemiology by moving beyond the identification of simple exposure–disease relationships. The concept of 'causal explanation', well developed in psychology, strengthens the theoretical basis for mediation analysis in epidemiology. In addition, delineating the questions that mediation analysis can address illuminates the conceptual meaning of different definitions of direct and indirect effects.

The differences between the 'pure' and 'total' effects are due to the differential inclusion of statistical interaction (synergy and parallelism).[14,18] For some purposes, the inclusion of synergy is part of what we mean by mediation; in others, the inclusion of parallelism is appropriate. Thus, in contrast to previous work,[16] we find that the assessment of mediation is conceptually meaningful in the presence of interaction between exposure and mediator. Interaction between exposure and mediator is not a source of bias, but rather an additional complexity that can yield rich information. However, the investigator must determine the correct natural effect based, in part, on the reason for analysis.

Previous papers have addressed the issue of bias in mediation analysis.[6,14,15] In fact, the focus of the methodological literature on mediation in epidemiology has been to highlight the potential biases of the analysis. These biases are indeed challenging and warrant serious consideration. Nonetheless, the assessment of mediation allows us to move beyond the simple identification of exposure–disease associations, toward an explanation of these relationships. The reasons for assessing mediation in epidemiology are compelling, and can be directly linked to extant mediational effects. Mediation analysis is very useful for opening the 'black box' between exposure and disease in epidemiologic studies.

Acknowledgements

The authors thank Tyler VanderWeele, Maria Glymour and Ezra Susser for their helpful comments on this manuscript. Financial support for this work comes from the Medical Scientist Training Program and Department of Epidemiology, Columbia University.

Conflict of interest: None declared.

KEY MESSAGES

- The assessment of mediation is a valuable tool for opening the 'black box' in epidemiology.
- There are several compelling reasons to assess mediation.
- These reasons correspond to different measures of direct and indirect effects.

References

[1] Taubes G. Epidemiology faces its limits. *Science (New York), NY* 1995;**269:**164–69.

[2] Skrabanek P. The emptiness of the black box. *Epidemiology* 1994;5553–55.

[3] Koopman JS. Emerging objectives and methods in epidemiology. *Am J Public Health* 1996;**86:**630–32.

[4] Weed DL. Beyond black box epidemiology. *Am J Public Health* 1998;**88:**12–14.

[5] Susser M, Susser E. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *Am J Public Health* 1996;**86:**674–77.

[6] Hafeman D. *Opening the black box: A reassessment of mediation from a counterfactual perspective [Dissertation]*. New York: Columbia University, 2008.

[7] Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 1986;**51:**1173–82.

[8] Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company, 2002.

[9] Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol Methods* 2002;**7:**422–45.

[10] MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annu Rev Psychol* 2007;**58:**593–614.

[11] MacKinnon DP. Analysis of mediating variables in prevention and intervention research. *NIDA Res Monogr* 1994;**139:**127–53.

[12] Susser M. *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology*. New York: Oxford University Press, 1973.

[13] Susser E, Schwartz S, Morabia A, Begg M, Brome E. *Psychiatric Epidemiology: Searching for Causes of Mental Disorders*. New York: Oxford University Press, 2006.

[14] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;**3:**143–55.

[15] Cole SR, Hernan MA. Fallibility in estimating direct effects. *Int J Epidemiol* 2002;**31:**163–65.

[16] Kaufman JS, Maclehose RF, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiol Perspect Innov* 2004;**1:**4.

[17] Robins J. The control of confounding by intermediate variables. *Stat Med* 1989;**8:**679–701.

[18] Hafeman DM. A sufficient cause based approach to the assessment of mediation. *Eur J Epidemiol* 2008;**23:** 711–21.

[19] Endogenous Hormones Breast Cancer Collaborative G. Body Mass Index, Serum Sex Hormones, and Breast Cancer Risk in Postmenopausal Women. *J Natl Cancer Inst* 2003;**95:**1218–26.

[20] Carney RM, Blumenthal JA, Freedland KE *et al*. Low heart rate variability and the effect of depression on post-myocardial infarction mortality. *Arch Intern Med* 2005;**165:** 1486–91.

[21] Pearl J. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2000.

[22] Winship C, Harding DJ. *A General Strategy for the Identification of Age, Period, Cohort Models: A Mechanism Based Approach*, 2004. http://www.qmp.isr.umich.edu/ASAMConference/Papers/WinshipHardingAPC.pdf (Accessed 20 June 2007).

[23] Morgan SL, Winship C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press, 2007.

[24] Bouchardy C, Verkooijen HM, Fioretta G. Social class is an important and independent prognostic factor of breast cancer mortality. *Int J Cancer* 2006;**119:**1145–51.

[25] Lynch JW, Kaplan GA, Cohen RD, Tuomilehto J, Salonen JT. Do cardiovascular risk factors explain the relation between socioeconomic status, risk of all-cause

mortality, cardiovascular mortality, and acute myocardial infarction? *Am J Epidemiol* 1996;**144:**934–42.

[26] Marmot M, Wilkinson RG. Psychosocial and material pathways in the relation between income and health: a response to Lynch *et al*. *Br Med J (Clinical Res Ed)* 2001;**322:**1233–36.

[27] Marmot MG, Bosma H, Hemingway H, Brunner E, Stansfeld S. Contribution of job control and other risk factors to social variations in coronary heart disease incidence. *Lancet* 1997;**350:**235–39.

[28] Lynch JW, Smith GD, Kaplan GA, House JS. Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions. *BMJ* 2000;**320:**1200–4.

[29] Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*. Gaithersburg, Maryland: Aspen Publishers, 2000.

[30] Salas-Salvado J, Garcia-Arellano A, Estruch R *et al*. Components of the mediterranean-type food pattern and serum inflammatory markers among patients at high risk for cardiovascular disease. *Eur J Clin Nutr* 2008;**62:**651–59.

[31] Bernat D, August G, Hektner J, Bloomquist M. The early risers preventive intervention: Testing for six-year outcomes and mediational processes. *J Abnorm Child Psychol* 2007;**35:**605–17.

[32] Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. *Eval Rev* 1981;**5:** 602–19.

[33] Pearl J. Direct and Indirect Effects. *Proceedings of the American Statistical Association Joint Statistical Meetings*. MN: MIRA Digital Publishing, 2001.

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

[35] Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology* 2006;**17:**276–84.

[36] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;**11:**550–60.

[37] Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D (eds). *Statistical Models in Epidemiology: The Environment and Clinical Trials*. New York: Springer-Verlag, 1999.

[38] van der Laan MJ, Petersen ML. Estimation of direct and indirect causal effects in longitudinal studies. *UC Berkeley Division of Biostatistics Working Paper Series 155* 2004 Berkley, CA: Berkley Electronic Press, 2004.

Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2009; all rights reserved. Advance Access publication 20 March 2009

International Journal of Epidemiology 2009;**38**:845–847
doi:10.1093/ije/dyp163

Commentary: Gilding the black box

Jay S Kaufman

Accepted 23 February 2009

Most epidemiology textbooks have the obligatory passage on ‘what is a cause?’ These discussions often start with Hume, pass reverently through Bradford-Hill and (if the book is of relatively recent vintage) end with Pearl. But as Hafeman and Schwartz[1] point out in their essay, few texts in our field go on to the question that really motivates these authors, which is ‘what is a causal structure?’ The closest thing I can find on my own bookcase might be Mervyn Susser's[2] ‘Causal Thinking in the Health Sciences’, now more than 35 years old and long out of print.

Despite the generally simplistic approach taken by our textbooks, the big breakthrough stories in biomedical research often sound like a sportscaster narrating the progress of a pinball machine game: ‘First the ball hits a lever, than bounces into a hole, where it triggers a sensor that opens a chute, and the ball slides down to the flipper ...’. The answer to a causal question such as ‘how did I just lose that ball down the side chute?’ can only be answered by referring to (i) multiple events that happened in sequence, (ii) the imagining of alternate ways that the sequence could have occurred instead and (iii) by having the whole process situated in a context with regular and comprehensible laws (such as the physical layout of the pinball machine, or the physiology and environment of a living organism). We naturally tend to conceive of ‘explanations’ as Rube Goldberg devices, where some exposure leads to a predictable cascade of events, and finally to the ‘outcome’—a cancer, or a death, or whatever. And we can easily imagine holding some intermediate gear still, which interrupts the natural flow of this device, and in this way we find that the outcome depended on that step. This is an analysis of mediation.

Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada. E-mail: jay.kaufman@mcgill.ca