Mediation analysis in epidemiology: methods, interpretation and bias.

Mediation analysis in epidemiology: methods, interpretation and bias

Lorenzo Richiardi[1]*, Rino Bellocco[2 3] and Daniela Zugna[1]

[1]Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, CPO-Piemonte, Turin, Italy, [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden and [3]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
*Corresponding author. Cancer Epidemiology Unit, University of Turin, Via Santena 7, 10126 Turin, Italy. E-mail: lorenzo.richiardi@unito.it

Abstract
In epidemiological studies it is often necessary to disentangle the pathways that link an exposure to an outcome. Typically the aim is to identify the total effect of the exposure on the outcome, the effect of the exposure that acts through a given set of mediators of interest (indirect effect) and the effect of the exposure unexplained by those same mediators (direct effect). The traditional approach to mediation analysis is based on adjusting for the mediator in standard regression models to estimate the direct effect. However, several methodological papers have shown that under a number of circumstances this traditional approach may produce flawed conclusions. Through a better understanding of the causal structure of the variables involved in the analysis, with a formal definition of direct and indirect effects in a counterfactual framework, alternative analytical methods have been introduced to improve the validity and interpretation of mediation analysis. In this paper, we review and discuss the impact of the three main sources of potential bias in the traditional approach to mediation analyses: (i) mediator-outcome confounding;(ii) exposure-mediator interaction and (iii) mediator-outcome confounding affected by the exposure. We provide examples and discuss the impact these sources have in terms of bias.
Key words: Mediation analysis, direct effects, indirect effects

Introduction

The importance of mediation analysis in epidemiological studies relies on the need to disentangle the different pathways that could explain the effect of an exposure on an outcome. Mediation analysis is typically applied when a researcher wants to assess the extent to which the effect of an exposure is explained, or is not explained by a given set of hypothesized mediators (also called intermediate variables[1]). In this way, the total effect of an exposure on an outcome, the effect of the exposure that is explained by a given set of mediators (indirect effect) and the effect of the exposure unexplained by those same mediators (direct effect) can be defined. Intuitively, one expects that the total effect can be decomposed into direct and indirect effects. Suppose that the total effect of a binary exposure translates into a risk difference of 15%; if the direct risk difference is 10%, we would expect one-third of the total effect to be explained by the mediator, and the remaining two-thirds to be explained by alternative pathways.

In this paper, we will address the fact that this intuitive expectation of effect decomposition may not hold true. Also, the concept of the proportion of effect explained by a mediator can be cumbersome in some situations. For example, walking to work increases both the total amount of physical activity and the total levels of exposure to air pollution. In an epidemiological study of incidence of coronary heart disease (CHD), if exposure to air pollution were the mediator, walking to work could have a protective direct effect on CHD but a simultaneous harmful indirect effect on CHD. Theoretically, in this scenario we could observe no total effect (risk difference = 0) on CHD due to opposite direct and indirect effects.

Epidemiological studies often require the study of mediation: for example, in studies of molecular mechanisms involved in disease causation, studies of socioeconomic inequality, studies of response to clinical treatments and studies aiming to measure the impact of public health interventions. The

assessment of mediation can be the main aim of the study, whereas often the goal is to estimate the total effect, though exploratory mediation analyses are also conducted.
The traditional approach to mediation analysis consists of comparing two regression models, one with and one without conditioning on the mediator.[2] The exposure coefficient is then interpreted as a direct effect in the model adjusted for the mediator and as a total effect in the unadjusted model. In epidemiological studies, the proportion of the total effect explained by the mediator is typically obtained by the ratio of the unadjusted to the adjusted relative risks, and the percent excess risk explained by the mediator is obtained by a ratio where the numerator includes the difference between the unadjusted (total effect) and the adjusted (direct effect) relative risks, and the denominator includes the unadjusted excess risk (total effect).[3,4] For example, if a study found a total effect of low vs high socioeconomic status (SES) on lung cancer risk equal to a relative risk of 2.3 and, after adjustment for smoking, the relative risk decreased to 1.2, the percent excess risk of SES on lung cancer risk explained by the smoking would be 85% [(2.3-1.2)/(2.3-1)*100].
It is now recognized that the traditional approach to mediation analysis is prone to bias arising both from incorrect statistical analysis and suboptimal study design. There is already a large amount of literature on this issue, both from a theoretical and an applied point of view, and it continues to grow rapidly. New statistical methods have been developed, although some are not fully implemented, and appropriate methods for some situations simply do not yet exist. The traditional approach to mediation analysis is still frequently used, and findings from earlier epidemiological studies that used this approach should not be discarded. It is thus fundamental to understand when, and to what extent, bias hampers the possibility to use and interpret traditional mediation analyses. In this paper we will discuss, describe and provide examples of the three main sources of potential bias that may cause traditional approaches to mediation analyses to give flawed conclusions:(i) mediator-outcome confounding, (ii) exposure-mediator interaction and iii) mediator-outcome confounding affected by the exposure. Throughout the paper, if not otherwise specified, we will not consider issues of random variation, unmeasured exposure-outcome confounders or measurement errors. The paper is organized as follows: we will first discuss mediator-outcome confounding using the aforementioned conventional definition of direct effects (i.e. the direct effect is the effect of the exposure on the outcome in a model adjusted for the mediator); we will then introduce a formal definition of direct and indirect effects in a counterfactual framework and discuss exposure-mediator interaction; finally, we will briefly discuss situations in which mediator-outcome confounders are affected by the exposure.

Mediator-outcome confounding

It is well known that lack of exposure-outcome confounding is necessary to obtain a valid estimate of the total effect of a given exposure on a given outcome. In mediation analysis, lack of mediator-outcome confounding is also necessary. This issue has been discussed several times in the past 20 years, though it was overlooked in early epidemiological studies.[5–7] The direct acyclic graph (DAG) shown in Figure 1 clarifies the issue: according to the causal graph theory, conditioning on the mediator M induces a spurious association between the mediator-outcome confounder U and the exposure A, where U becomes a confounder of the exposure-outcome association and induces bias (Figure 1). This is an example of collider bias, which occurs frequently in epidemiological studies (e.g. selection bias[8]). A simple example of this situation in the context of mediation analysis would be given by a study designed to assess how much of the total effect of exposure to environmental noise on CHD is mediated by hypertension. All variables that affect both hypertension and CHD risk, such as body mass index, diet and smoking, act as mediator-outcome confounders; therefore they should be measured and considered in the analyses to validly estimate the direct effect of noise. For those unfamiliar with DAG language,[9] consider that M in Figure 1 is caused by A and U, both of which are sufficient causes of M. In this case, collider bias arises because in the stratum M = 1 (e.g. in the stratum of people with hypertension), if U were not present, A should be present in

order to have hypertension. Therefore, in this example, for a given level of M, A and U are inversely associated even if they are marginally independent.

When the mediator is included as a covariate in a regression model to estimate the direct effect of an exposure on the outcome, adjusting for mediator-outcome confounders (U variable in Figure 1) is needed to avoid bias. If some, or all, of these confounders are unmeasured or unknown, estimate of the direct effect might be invalid. Therefore, it is always important to assess how the results obtained from any mediation analysis could be affected by the possible unmeasured/residual mediator-outcome confounding, the main question being whether this source of bias could explain away the estimated direct effect.[10]

The magnitude of the bias introduced by conditioning on a collider, both in a general setting and in the context of mediation analysis, is an issue that has been addressed by several authors.[11–14] Vanderweele provided simplified formulas to carry out, under specific assumptions, a quick sensitivity analysis for the estimate of the direct effect.[13] On the risk ratio scale, if γ is defined as the direct effect of the unmeasured binary confounder U on the outcome (for given levels of the exposure A and the mediator M), and $\pi_{a,m}$ and $\pi_{a^*,m}$ are the prevalences of the unmeasured confounder U among the two exposure levels *a* and *a** at a given level of the mediator M=*m*, under simplifying assumptions the bias in the direct effect estimate of *a* vs *a** would be obtained by: $B = [1 + (\gamma\text{-}1)\,\pi_{a,m}]/[1 + (\gamma\text{-}1)\,\pi_{a^*,m}]$. Assuming that the unmeasured confounder U is not itself affected by the exposure A, the bias-corrected direct effect estimate can be obtained by dividing the risks ratio adjusted for the mediator by the bias factor B obtained from different scenarios of values for the parameters γ, $\pi_{a,m}$ and $\pi_{a^*,m}$.

For example, in a recent study on the association between ethnicity (Maori women vs women of European origin) and late stage at diagnosis of cervical cancer in New Zealand, it was found that most of the total effect of Maori ethnicity on late stage at diagnosis (OR: 2.71) did not change much after adjustment for screening practices (direct effect OR: 2.39).[15] The study concluded that ethnicity-related differences in stage at diagnosis of cervical cancer in New Zealand could not be explained by ethnic-related differences in screening attendance. It is possible, however, that part of the estimated direct effect was due to bias introduced by unmeasured mediator-outcome confounders. However, in order to explain completely a direct effect estimate of 2.39 among, say, unscreened women with this source of bias, we would have to assume, for example, that the supposed mediator-outcome confounder was associated with the outcome with a relative risk (γ) equal to 4.0, had a prevalence of 65% among unscreened Maori women and a prevalence of 10% among unscreened women of European origin. This scenario seems unlikely to occur in real practice.

Weaker direct effects could however be entirely explained by bias due to mediator-outcome confounding. For example, there is a great deal of interest in understanding the role of SES inequalities in morbidity and mortality, and whether the effects of this variable remain after taking into account well known risk factors.[16] In these studies, the direct effect is often fairly small, as typically most—but not all—of the association between SES and the disease under study can be explained. Let us consider the hypothetical example described in the previous section: a study on lung cancer yields a total relative risk for low vs high SES of 2.3, which, after adjustment for smoking, decreases to 1.2. A mediator-outcome confounder (say family history of lung cancer, assuming that is not itself affected by socioeconomic status) with, for example, a relative risk (γ) for lung cancer of 2.5, a prevalence of 20% among non-smokers with low SES and a prevalence of 5% among non-smokers with high SES, could entirely explain a direct effect of 1.2 among non-smokers.

It is also of interest to consider the direction of the bias. According to the Vanderweele's formula, when, conditioned on the mediator, there is a positive association between the exposure and the unmeasured mediator-outcome confounder, which in turn has a positive direct effect on the outcome, the estimate of the direct effect of the exposure on the outcome is biased upwards (i.e. the unmeasured mediator-outcome confounder becomes a positive confounder of the exposure-outcome

association after conditioning on the mediator). Although there are exceptions, conditioning on a variable (collider) that is affected by two other variables (parents) typically induces a negative association between the parents if they affect the collider in the same direction (either positive or negative), whereas the association is positive if the two parents affect the collider in opposite directions.[17,18] Thus, if an exposure positively affects the mediator, and the supposed mediator-outcome confounder is positively associated with both the outcome and the mediator, the direct effect for a given level of M is likely to be biased downwards. Returning to our hypothetical study on noise (exposure), hypertension (mediator) and CHD (outcome), many factors, such as smoking and body mass index, are likely to be positively associated with both hypertension and CHD risk. As noise is also expected to increase the risk of hypertension, all the associations involved are thus positive. In this situation, if a positive direct effect of noise on CHD is found, that effect is unlikely to be explained by the bias introduced from any unmeasured mediator-outcome confounders. Conversely, the magnitude of the positive direct effect is likely to be underestimated.

In the recent literature on mediation analysis, the so-called low birthweight paradox, i.e. the inverse association of maternal smoking on infant mortality that is typically observed in children with low birthweight (the mediator), has often been used as an example of bias introduced by unmeasured mediator-outcome confounding.[19] In this example, some confounders, such as birth defects, are positively associated with both low birthweight and infant mortality, and at the same time maternal smoking is positively associated with low birthweight. The resulting bias is thus downwards, corresponding to an apparent protective direct effect of maternal smoking on infant mortality among children with low birthweight. In sensitivity analyses, it has been shown that sensible assumptions regarding the magnitudes of the associations involved could explain away this apparent association.[20]

Obviously, the collider bias is not the only source of bias affecting mediation analysis although it is probably the most largely overlooked source in past mediation analyses. Misclassification of the mediator, for example, can also seriously bias conclusions. A trivial example would be a non-differential misclassification of a binary mediator so large as to obscure the presence of any indirect effect. As recently shown, the general rule is that a nondifferentially misclassified (binary) mediator overestimates the magnitude of the direct effect and underestimates the magnitude of the indirect effect.[21]

Exposure-mediator interaction

According to the traditional approach to mediation analysis, the direct effect is estimated by conditioning on the mediator M. In the hypothetical data reported in Table 1 there is a total risk difference for the exposure of 4.8%, which decreases to 2.3%, after adjustment for the mediator M, thus indicating the presence of a direct effect. If we estimate the effect of the exposure A in those without the mediator (M = 0), the risk difference for the event associated with the exposure is 2.0%. As the mediator in this example is a binary variable, there are two possible direct effects that can be estimated: the risk difference (2%) among those with M = 0, and the risk difference (18%) among those with M = 1. Although the risk difference is lower than the total effect in the stratum M = 0, it is much larger than the total effect in the stratum M = 1. In this example, the estimate of the direct effect depends on the value of the mediator.

Table 1
Hypothetical data on the risk of being a case associated with an exposure (A) and a mediator (M)

Exposure (A)	Mediator (M)	Risk	Cases	Non-cases	Total
0	0	1%	100	9900	10 000
1	0	3%	150	4850	5000
0	1	2%	10	490	500
1	1	20%	200	800	1000

The fact that the estimates of direct effect vary across different levels of the mediator implies that the exposure A and the mediator M interact in explaining the outcome. In Table 1, using unexposed subjects without the mediator as the reference, the observed effect of being exposed with the mediator (risk difference = 19%) is much larger than the linear combination of the two effects of being in the exposed group without the mediator (risk difference = 2%) and having the mediator without the exposure (risk difference = 1%).

If we are interested in estimating the effect of an exposure that is not explained by a mediator in the presence of exposure-mediator interaction, we need to introduce an alternative formal definition of direct effect,[5,6,22] which provides a population summary of the effects at different levels of the mediator. The same applies when the relationship between the exposure and the mediator is not linear, but here we will not discuss this case further.

The alternative definition uses a counterfactual framework to define natural direct effects and natural indirect effects that sum up to the total effect.[5,6] In a counterfactual framework, the individual causal effect of the exposure on the outcome is defined as the hypothetical contrast between the outcomes that would be observed in the same individual at the same time under the exposure and in the absence of the exposure (or in presence of two different levels of the exposure).[23,24] According to the counterfactual notation, Y_a is the potential outcome under exposure A = a and Y_{a^*} is the potential outcome under the exposure level A = a*, where $a \neq a^*$. Obviously, as these are potential outcomes under alternative exposure levels, it is not possible to observe both Y_a and Y_{a^*} in the same individual: only one of the two would be factual. The individual causal effect, defined as $Y_a - Y_{a^*}$, is unlikely to be the same for all individuals of a given population. We thus define the population causal effect as the average of the individual causal effects, i.e. $E(Y_a - Y_{a^*})$.

In the context of mediation analysis, $Y_{a,m}$ is the potential outcome under exposure level A = a and mediator level M = m. The natural direct effect is defined as $Y_{a,M(a^*)} - Y_{a^*,M(a^*)}$, i.e. the difference between the value of the counterfactual outcome if the individual were exposed to A = a and the value of the counterfactual outcome if the same individual were instead exposed to A = a*, with the mediator assuming whatever value it would have taken at the reference value of the exposure A = a* (Box 1). At the population level, the natural direct effect is $E(Y_{a,M(a^*)} - Y_{a^*,M(a^*)})$. The natural direct effect captures the effect of A on Y via pathways that do not involve M, although the value of the mediator M is allowed to vary among individuals according to all the determinants of M, with the exception of the exposure A. This illustrates one of the reasons why a counterfactual framework is used for the definition of natural direct effects, namely to conceptualize the hypothetical distribution of the mediator. The natural indirect effect can be defined as $Y_{a,M(a)} - Y_{a,M(a^*)}$, i.e. the contrast, having set the exposure to a fixed level A = a, between the value of the counterfactual outcome if the mediator assumed whatever value it would have taken at a level of the exposure A = a and the value of the counterfactual outcome if the mediator assumed whatever value it would have taken at a reference level of the exposure A = a* (Box 1). At the population level, the natural indirect effect is $E(Y_{a,M(a)} - Y_{a,M(a^*)})$. Intuitively, the natural indirect effect captures the effect of the exposure A on the outcome Y due to the effect of the exposure A on the mediator M. The total causal effect of A on Y can now be decomposed into the sum of the natural direct effect and the natural indirect effect, even in presence of exposure-mediator interaction. Note that slightly different ways to decompose the total effect into direct and indirect effects have been proposed.[5,25]

Box 1. Definitions of controlled direct effect, natural direct effect and natural indirect effect in the counterfactual framework

***Controlled direct effect*:** $Y_{a,m} - Y_{a^*,m}$

This effect is the contrast between the counterfactual outcome if the individual were exposed at A = a and the counterfactual outcome if the same individual were exposed at A = a*, with the mediator set to a fixed level M=*m*.

***Natural direct effect*:** $Y_{a,M(a^*)} - Y_{a^*,M(a^*)}$

This effect is the contrast between the counterfactual outcome if the individual were exposed at A = a and the counterfactual outcome if the same individual were exposed at A = a*, with the mediator assuming whatever value it would have taken at the reference value of the exposure A = a*.

***Natural indirect effect*:** $Y_{a,M(a)} - Y_{a,M(a^*)}$

This effect is the contrast, having set the exposure at level A = a, between the counterfactual outcome if the mediator assumed whatever value it would have taken at a value of the exposure A = a and the counterfactual outcome if the mediator assumed whatever value it would have taken at a reference value of the exposure A = a*.

In mediation analysis, a counterfactual framework is also used to define an additional meaningful effect, the controlled direct effect (Box 1). Controlled direct effect is defined as $Y_{a,m} - Y_{a^*,m,}$ i.e. a contrast between counterfactual outcomes with alternative exposure values, A = a and A = a*, if the mediator were set to a fixed value M = m. At the population level the controlled direct effect is $E(Y_{a,m} - Y_{a^*,m})$. As a consequence, there are as many controlled direct effects as there are levels of the mediator. A controlled direct effect thus corresponds to a situation in which a hypothetical intervention controls the mediator to a given value,[6,22] whereas a natural direct effect corresponds to a situation in which the natural relationship between the exposure and the mediator is maintained (i.e. we would intervene on the exposure but not directly on the mediator). In the example of a hypothetical study on noise, hypertension (the mediator) and risk of CHD, the controlled direct effect (for hypertension = 0) would be the effect of elimination of noise exposure when controlling hypertension to be absent, whereas for the natural direct effect hypertension would be set at the value that would have been observed in the absence of noise exposure.

In the absence of interaction between the exposure and the mediator, controlled direct effect and natural direct effect are equivalent. The intuitive explanation for this equivalence is that if the direct effect of the exposure is constant for the different levels of the mediator, setting the mediator to a fixed value (controlled direct effect) or considering the value that the mediator would have taken at the reference level of the exposure (natural direct effect) gives the same estimate (i.e. a weighted average between constant values gives the same result irrespectively of the weights).

When there is interaction between the exposure and the mediator, the natural direct effect and the natural indirect effect still sum up to the total effect and they represent a sort of interpretable population average over the levels of the mediator. Note that, if exposure-mediator interaction exists, the estimate of the total effect associated with the exposure in a given population depends on the population prevalence of the mediator. The same applies to the natural effects: when exposure-mediator interaction is present, natural effects can be estimated and interpreted, but their estimates are population-specific. We would like to propose the same example discussed by Judea Pearl to illustrate the use and interpretation of natural direct effects.[6] Pearl considered a situation where a drug could induce headache as a side effect, and, at the same time, could interact with aspirin taken to treat the drug-induced headache on its effects on the outcome. In this situation, the drug is the exposure and the aspirin is the mediator. Obviously, aspirin may be taken in the population for reasons other than the drug-induced headache. Now, imagine that the producer of the drug manages to eliminate headache as a side effect, and would like to know what the effect of the drug will be in the population, knowing that use of the drug will no longer be a cause of aspirin intake. The natural direct effect is the key quantity that answers this question, but its estimate depends on the aspirin use in absence of the exposure in that population. To further explore this concept, let us assume now that the drug does not work when taken without aspirin. If people in the population take aspirin for reasons other than the drug-induced headache, the drug would still have a natural direct effect, whereas if people in the population only take aspirin for the drug-induced headache there would be no natural direct effect after this headache was eliminated. Conversely, controlled direct effect, when the aspirin intake is set to be 0, would be the same in the two populations.

How can we estimate these effects? Under specific assumptions, controlled and natural direct effects can be estimated using standard regression models. Assuming no unmeasured mediator-outcome confounding and no mediator-outcome confounding affected by the exposure, the controlled direct effect can be estimated by conditioning the analysis on the mediator. Assuming also no unmeasured exposure-mediator confounding, the natural direct effect can be estimated as a weighted average of the controlled direct effects, with weights for each level of the mediator given by the probability that the mediator would have taken that value if the exposure were set at its reference level. Going back to the hypothetical data reported in Table 1, the estimate of the natural direct effect can be non-parametrically obtained by averaging the two controlled direct effects of 2% and 18%, using the frequency of the mediator among the unexposed subjects as the weighting function. [26] Table 1 shows a 4.76% probability of M being present (M = 1) among the unexposed subjects, and the natural direct effect can be obtained by the following: [2%*(1- 4.76%) + 18% *(4.76%)] = 2.8%. More complex methods (see Discussion), based on parametric assumptions, are used when simpler non-parametric estimates are not feasible.

As we have shown in this section, the presence of exposure-mediator interaction may introduce large problems in mediation analysis and in its interpretation, and therefore should be considered whenever interpreting the results of traditional analyses. Let us consider an additional example of a study that aims to understand to what extent differences in mortality by SES among cancer patients are explained by stage at diagnosis. If we assume there is no interaction between SES and stage at diagnosis, it implies that SES inequalities in mortality are the same irrespective of the stage at diagnosis (even if, for example, low SES is associated with later stage at diagnosis), whereas presence of an interaction would imply that the stage at diagnosis may increase or decrease the effect of SES on mortality. We assumed a simplified scenario in which, after cancer is diagnosed, SES has an impact on mortality only through the type and quality of treatment received by the patients. Then, to understand whether standard analyses designed to estimate controlled direct effects can provide interpretable estimates in terms of mechanisms, one of the key questions that needs to be posed is whether an interaction exists between SES and stage at diagnosis on mortality. For example, for some cancer types SES inequalities regarding treatment may occur in patients diagnosed at an early stage, whereas at very advanced stages where effective treatments are lacking, SES inequalities disappear. For other cancer types, SES inequalities might be more constant across stages.

Mediator-outcome confounding affected by the exposure

Finally, we introduce the third potential source of bias. As mentioned previously in the section on mediator-outcome confounding, it is necessary to adjust for mediator-outcome confounding in standard regression models to avoid collider bias. However, there are exceptions in which adjustment for such confounders in standard regression models still produces flawed estimates. Figure 2 depicts a scenario where the mediator-outcome confounder L is now affected by the exposure (A). In this scenario, L, also referred to as intermediate confounder, [27] is both a mediator-outcome confounder and a variable that lies on the direct path from the exposure A to the disease Y (Figure 2a).

Intermediate confounding is probably not rare in mediation analysis. Let us consider a hypothetical study aiming to assess to what extent the effect of smoking on CHD is mediated by atherosclerosis.[28] A number of variables, including blood pressure, affect both atherosclerosis and the risk of CHD, and are also affected by smoking (Figure 2b). Adjustment for blood pressure in traditional regression models would bias the estimate of the direct effect by blocking the effect of smoking on CHD acting through blood pressure, but not atherosclerosis (i.e. the path smoking → blood pressure → CHD). This would induce an attenuation of the direct effect and a consequent overestimate of the indirect effect. On the other hand, adjustment for blood pressure is necessary to

prevent collider bias (that is inherently introduced by adjusting for the mediator atherosclerosis). As discussed in the section on mediator-outcome confounding, if we assume that: (i) smoking and blood pressure both positively affect atherosclerosis; (ii) smoking positively affects blood pressure; and (iii) blood pressure positively affects the risk of CHD, adjustment for atherosclerosis would likely bias the direct effect of smoking on CHD downwards, although the determination of the direction and magnitude of bias may be difficult in complex DAGs.[29] Interestingly, in this scenario the bias goes in the same direction whether adjusting or not adjusting for blood pressure, implying that it is not possible to conduct both analyses and conclude that the unbiased estimate lies somewhere in the middle.

The causal structure depicted in Figure 2 has been discussed in depth, first in scenarios of time-dependent exposures and confounders, and then in the framework of mediation analyses.[30] Statistical approaches, such as inverse probability weighting[30,31] and g-computation,[32] which are both based on the counterfactual framework, are generally able to adjust for the confounding effect of L without blocking the corresponding direct path from the exposure A to the outcome Y, and to estimate controlled direct effects, as well as, under stronger assumptions, natural direct and indirect effects.[5,22,27,33] Briefly, these methods model the expected potential outcome under exposure A = a and the mediator M = m, $E(Y_{a,m})$: the inverse probability weighting by regressing the outcome on the exposure and the mediator and by controlling for potential confounders by re-weighting the population instead of introducing them in the regression model; the g-computation by an extension of the standardization using Monte Carlo simulations.[34]

To assess the amount of bias that traditional analyses could introduce in the presence of intermediate confounding, the strengths of the associations between the exposure and the mediator-outcome confounder L and between L and the outcome (in our example it would be between smoking and blood pressure and between blood pressure and CHD) should be evaluated. If the presence of any of these two associations is more an issue of theoretical discussion rather than a real threat to the analysis, more advanced methods to deal with intermediate confounding will produce estimates similar to standard methods. On the contrary, if, as in our example, both associations are likely to play an important role, traditional analyses will not provide the correct answers.

Discussion

Research on methods for mediation analysis is a fast growing field in epidemiology; its development is related to the need to better understand mechanisms, and follows with somewhat surprising delay earlier discussions on black box epidemiology,[35] conceptual frameworks[36] and molecular epidemiology.[37] Standard or traditional approaches to mediation analysis can produce flawed conclusions and their main limitations have been addressed at length in the methodological literature.

Although the investigation of statistical methods for mediation analysis is not in the scope of this paper, we should emphasize that new non-parametric and parametric approaches, based on counterfactual framework, are now available to address some of the problems we describe herein, including the Mediation formula, inverse probability weighting and g-formula.[5,26,27,30,33,34] These methods are reaching now a wide spread and are entering the epidemiological literature and textbooks, though they are still underused in applied epidemiology. It should be emphasized that their implementation may be complex, and that they are subject to strong assumptions that need to be met in order to obtain valid and interpretable estimates.[38] Furthermore, there are epidemiological scenarios for which valid methods are not yet available and, for other scenarios, new approaches have either only recently been suggested, or more options exist but their performance has not been fully compared.[27,39,40]

In this paper, we reviewed some of the most basic problems that can arise in mediation analysis, the concepts and the methods that have been developed to tackle them, and provided some examples. The rapid development in this field is characterized by levels of formalism and conceptualization that may be somewhat difficult for applied epidemiologists to integrate. This is probably the main

reason why the new methods are being introduced rather slowly in epidemiological research. Indeed, one of the recent focuses of research in mediation analysis has been the development of simplified or unified approaches that could be adopted by a broader group of users.[26,41] We predict that the use of new and more correct approaches to mediation analyses in common epidemiological studies will increase rapidly in the next years.

Funding
Lorenzo Richiardi was partially funded by the Compagnia SanPaolo Foundation and the Italian Association for Cancer Research. Rino Bellocco was partially funded by the Italian Ministry of University and Research (MIUR), PRIN 2009 X8YCBN
Conflict of interest: None declared.

KEY MESSAGES

- Mediation analysis is common in epidemiology; it aims to disentangle the effect of an exposure on an outcome explained (indirect effect) or unexplained (direct effect) by a given set of mediators.
- Traditional approaches to estimate the direct effect, based on simply adjusting for the mediator in a standard regression setting, may produce invalid results.
- Potential sources of bias include unmeasured mediator-outcome confounding, interaction between exposure and mediator, and presence of intermediate confounding.
- The validity and interpretation of mediation analysis is enhanced by using the counterfactual framework to conceptualize the controlled direct effect, the natural direct effect and the natural indirect effect of the exposure on the outcome.
- Research on methods for mediation analysis is a fast growing field in epidemiology and biostatistics.

Acknowledgements
We are indebted to Arvid Sjolander and Milena Maule for very useful comments on an early version of the manuscript

References

1 Porta M (ed.). A Dictionary of Epidemiology . 5th edn. Oxford: Oxford University Press, 2008.
2 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.
3 Kaufman JS, Maclehose RF, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to iden- tify biologic mediation. Epidemiol Perspect Innov 2004; 1: 4.
4 Szklo M, Nieto J. Epidemiology: Beyond the Basic . Sudbury, MA: Jones and Bartlett, 2004.
5 Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992; 3: 143–55.
6 Pearl J. Direct and Indirect Effects . Seventeenth Conference of Uncertainy in Artificial Intelligence, 2001. San Francisco, CA: Morgan Kaufmann, 2001.
7 Cole SR, Hernan MA. Fallibility in estimating direct ef- fects. Int J Epidemiol 2002; 31: 163–65.
8 Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15: 615–25.
9 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999; 10: 37–48.
10 Blakely T. Commentary: Estimating direct and indirect effects—fallible in theory, but in the real world? Int J Epidemiol 2002; 31: 166–67.
11 Greenland S. Quantifying biases in causal models: clas- sical confounding vs collider-stratification bias. Epidemiology 2003; 14: 300–06.
12 Hafeman DM. Confounding of indirect effects: a sensitiv- ity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. Am J Epidemiol 2011; 174: 710–17.
13 VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology 2010; 21: 540–51.
14 Pizzi C, De Stavola B, Merletti F et al . Sample selection and validity of exposure-disease association estimates in cohort studies. J Epidemiol Community Health 2012; 65: 407–11.
15 Brewer N, Zugna D, Daniel R, Borman B, Pearce N, Richiardi L. Which factors account for the ethnic inequalities in stage at diagnosis and cervical cancer survival in New Zealand? Cancer Epidemiol 2012; 36: e251–57.
16 Hafeman DM, Schwartz S. Opening the Black Box: a mo- tivation for the assessment of mediation. Int J Epidemiol 2009; 38: 838–45.
17 Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes JM, Kaufman JS (eds). Methods in Social Epidemiology . San Francisco, CA: Jossey-Bass, 2006.
18 VanderWeele TJ, Robins JM. Directed acyclic graphs, suf- ficient causes, and the properties of conditioning on a common effect. Am J Epidemiol 2007; 166: 1096–104.
19 Hernandez-Diaz S, Schisterman EF, Hernan MA. The birth weight ‘‘paradox’’ uncovered? Am J Epidemiol 2006; 164: 1115–20.
20 VanderWeele TJ, Mumford SL, Schisterman EF. Conditioning on intermediates in perinatal epidemiology. Epidemiology 2012; 23: 1–9.
21 Ogburn EL, Vanderweele TJ. Analytic results on the bias due to nondifferential misclassification of a binary medi- ator. Am J Epidemiol 2012; 176: 555–61. 22 Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. Epidemiology 2006; 17: 276–84.
23 Sjolander A. The language of potential outcomes. In: Sons JW (ed.). Causality: Statistical Perspectives and Applications . Chichester: Wiley, 2012.

24 Rubin DB. Estimating causal effects of treatment in ran- domized and nonrandomized studies. J Educ Psych 1974; 56: 688–701.
25 VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. Epidemiology 2013; 24: 224–32.
26 Pearl J. The causal mediation formula—a guide to the assessment of pathways and mechanisms. Prev Sci 2012; 13: 426–36.
27 Vansteelandt S. Estimating direct effects in cohort and case-control studies. Epidemiology 2009; 20: 851–60.
28 Ambrose JA, Barua RS. The pathophysiology of cigarette smoking and cardiovascular disease: an update. J Am Coll Cardiol 2004; 43: 1731–37.
29 VanderWeele TJ, Hernan MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured con- founding bias. Epidemiology 2008; 19: 720–28.
30 Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000; 11: 550–60.
31 Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000; 11: 561–70.
32 Robins JM. A new approach to causal inference in mor- tality studies with sustained exposure periods—applica- tion to control of the healthy worker survivor effect. Mathe Model 1986; 7: 1393–512.
33 VanderWeele TJ. Marginal structural models for the esti- mation of direct and indirect effects. Epidemiology 2009; 20: 18–26.
34 Daniel RM, De Stavola B. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. STATA J 2011; 11: 479–517.
35 Susser M, Susser E. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemi- ology. Am J Public Health 1996; 86: 674–77.
36 Victora CG, Huttly SR, Fuchs SC, Olinto MT. The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. Int J Epidemiol 1997; 26: 224–27.
37 Perera FP, Weinstein IB. Molecular epidemiology and car- cinogen-DNA adduct detection: new approaches to stu- dies of human cancer causation. J Chronic Dis 1982; 35: 581–600.
38 Pearl J. Interpretable Conditions for Identifying Direct and Indirect Effects . Technical Report R.389. Department of Computer Science, University of California, 2012.
39 VanderWeele TJ. Causal mediation analysis with survival data. Epidemiology 2011; 22: 582–85.
40 Lange T, Hansen JV. Direct and indirect effects in a sur- vival context. Epidemiology 2011; 22: 575–81.
41 Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. Am J Epidemiol 2012; 176: 190-95.