Dealing with complex problems of confounding in mediation analysis

Stijn Vansteelandt

Abstract Mediation analysis is frequently utilized in diverse scientific fields such as psychology, sociology and epidemiology, to develop insight into the causal mechanism whereby an exposure affects an outcome. It concerns the study of indirect effects of that exposure that are mediated through a given intermediate variable or mediator, and/or the study of the remaining direct effect. Despite its popularity, the traditional approach to mediation analysis proceeds in a predominantly heuristic fashion, which can largely be ascribed to the lack of precise definitions of direct and indirect effect in the traditional mediation analysis literatures. Moreover, problems of confounding bias have been largely ignored.

James Robins, Sander Greenland and Judea Pearl laid the foundations for a rigorous approach towards mediation analysis, which is based on counterfactuals. They gave precise definitions of direct and indirect effect and elucidated the kind of data that must be collected in order to control for confounding bias. In addition, they provided generic ways to decompose a total effect into a direct and indirect effect that is not tied to a specific statistical model. In this presentation, after a brief review of some of these developments, I will concentrate on the - partly unsolved - methodological challenges that arise when confounders of the mediator-outcome association are affected by the exposure. In particular, I will present results on the the identification of (natural) direct and indirect effects in such settings, and on the estimation of (controlled) direct effects, thereby focussing on matched case-control studies and/or survival analysis.

Key words: causal inference, direct effect, G-estimation, indirect effect, intermediate confounding, mediation, time-varying confounding

Stijn Vansteelandt
Ghent University, Department of Applied Mathematics and Computer Science, Krijgslaan 281, S9, 9000 Gent, Belgium, e-mail: stijn.vansteelandt@UGent.be

1 Introduction

For many decades, scientists from diverse scientific fields - most notably, psychology, sociology and epidemiology - have been occupied with questions as to whether an exposure affects an outcome through pathways other than those involving a given mediator or intermediate variable. The answer to such questions is of interest because it brings insight into the mechanisms that explain the effect of exposure on outcome [12]. Mediation analyses are used for this purpose. They attempt to separate so-called 'indirect effects' from 'direct effects'. The former term is typically used in a loose sense to designate that part of an exposure effect which arises indirectly by affecting a (given) set of intermediate variables; the latter then refers to the remaining exposure effect.

In traditional mediation analysis, the direct effect is commonly connected with the residual association between outcome and exposure after adjusting for the mediator(s); the indirect effect is then obtained through a combination of the exposure's effect on the mediator and the mediator's effect on the outcome [1, 5]. For instance, when the associations between exposure A and mediator M and outcome Y can be modeled through linear regressions as

$$\begin{aligned} E(Y|A,M) &= \beta_0 + \beta_a A + \beta_m M \\ E(M|A) &= \alpha_0 + \alpha_a A, \end{aligned}$$

then β_a is commonly interpreted as a direct effect and $\beta_m \alpha_a$ as an indirect effect [1]. It is well known from the causal inference literature that these interpretations are often not justified as a result of confounding of the mediator-outcome association [9, 3]. Even when confounders L of this association have been measured, standard regression adjustment is not applicable when - as often - some of these confounders are themselves affected by the exposure, in which case we say that there is *intermediate* or *time-varying* confounding [4, 10, 13]. Furthermore, decomposition of a total effect into a direct and indirect effect becomes subtle when certain nonlinear associations exist between mediator and outcome [9, 7], e.g. when a logistic regression model for a dichotomous outcome is adopted [11].

Robins and Greenland [9] and Pearl [7] introduced model-free definitions of direct and indirect effect. Unlike the foregoing development due to Baron and Kenny [1], their formalism of so-called natural direct and indirect effects can therefore accommodate nonlinear associations between mediator and outcome. Natural direct and indirect effects are defined in terms of so-called composite or nested counterfactuals such as $Y(a, M(0))$, which denotes the counterfactual outcome that would have been observed if the exposure A were set to a and the mediator M to the value $M(0)$ that it would have taken at some reference exposure level 0. Because such composite counterfactuals are unobservable when $a \neq 0$, strong assumptions must be imposed for identification. The development of Robins and Greenland [9] precludes the existence of *moderation*, i.e. exposure effect modification by the mediator on the additive scale; it precludes such moderation even at the unit level. The development of Pearl [7] precludes the possibility of intermediate confounding of the mediator-outcome association. This places severe restrictions on the range of realistic applications that can be addressed. In fact, the prior absence of methodology to deal with intermediate confounding has been one of the difficulties with the causal inference literature on mediation.

This presentation will primarily focus on this problem of intermediate confounding in mediation analysis. First, I will consider the problem of estimating so-called controlled direct effects in the presence of exposure-induced confounding of the association between mediator and outcome. I will thereby focus on diverse settings like survival analysis and the analysis of matched case-control studies. Next, I will propose novel results on the identification of natural direct and indirect effect in the presence of intermediate confounding.

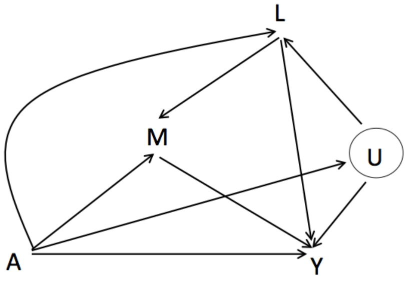

Fig. 1 Causal diagram with exposure A, mediator M, outcome Y, intermediate confounder L, and with U an unmeasured confounder of the L-Y relationship.

2 The problem of intermediate confounding in mediation analysis

The causal diagram of Figure 1 displays a setting with intermediate confounding. It visualizes prognostic factors L of the mediator (other than the exposure) that may also be associated with the outcome, and which thereby confound the association between mediator and outcome. This situation is representative of most empirical studies, including randomized experiments, because the fact that the exposure is randomly assigned does not prevent confounding of the mediator-outcome association. In the presence of such confounding, the residual association between outcome and exposure after adjusting for the mediator(s) (cfr. β_a in the above model) does not encode a direct exposure effect. This is technically seen because adjustment for a collider M (i.e. a node in which two edges converge) along the path $A \to M \leftarrow L \leftarrow U \to Y$ may render exposure A and outcome Y dependent along that path, and may thus induce a non-causal association [8, 3]. One of the major contributions of the causal inference literature has been to point this out and to make clear that specialized estimation techniques are often needed to be able to adjust for such confounders, as these may themselves be affected by the exposure (as illustrated in Figure 1). Indeed, additional regression adjustment for the confounder L once again amounts to adjustment for a collider L along the path $A \rightarrow L \leftarrow U \rightarrow Y$. It thereby renders A and Y dependent along that path, even in the absence of a direct effect.

3 Estimation of controlled direct effects in the presence of intermediate confounding

Let $Y(a,m)$ denoting the counterfactual outcome that would have been observed for given subject if the exposure were set to a and the mediator to m. Then a *controlled direct effect* [9, 7] refers to a contrast between two counterfactual outcomes for the same subject, corresponding to different exposure levels, but the same fixed mediator level. For instance, the controlled direct effect of exposure level a versus reference exposure level 0, controlling for M, can then be defined as the expected contrast

$$E\{Y(a,m)-Y(0,m)\}.$$

Likewise, the conditional controlled direct effect, given covariates C, of exposure level a versus reference exposure level 0, controlling for M, can then be defined as the expected contrast

$$E\{Y(a,m)-Y(0,m)|C\}.$$

Robins [8] showed that, under specific identification assumptions that we shall describe next, controlled direct effects can be identified in the presence of intermediate confounding. Specifically, provided that data have been recorded on all confounders of the exposure-outcome relationship, as well as all confounders of the mediator-outcome relationship, the conditional controlled direct effect can be identified using the so-called G-formula:

$$\begin{aligned} E\{Y(a,m)-Y(0,m)|C\} &= \int E(Y|A=a,M=m,L)f(L|A=a,C)dL \\ &\quad - \int E(Y|A=0,M=m,L)f(L|A=0,C)dL. \end{aligned}$$

It thus follows that parametric models for the outcome and intermediate confounders can be combined to result in an expression for the controlled direct effect. However, the G-formula does not admit a practical approach. It requires parametric models for the intermediate confounders, which can be problematic when the confounder is high-dimensional. Moreover, it can be computationally cumbersome as a result of the possibly high-dimensional integration which it involves. Finally, even simple models for the outcome and intermediate confounder may combine into intractable expressions for the controlled direct effect, which depend on the exposure level a and covariate C in a highly contrived way. This not only makes results impractical for reporting, but also makes interesting hypotheses difficult to test [8].

Various approaches have been developed to accommodate this, some of which we will review in this presentation.

One class of approaches involves weighting each subject's data by the reciprocal of the likelihood of the observed mediator, given exposure and confounders, and then regressing the outcome on exposure and mediator [8, 10]. Since the weighting corrects for confounding bias, the weighted regression analysis of the outcome can ignore confounders and therefore does not suffer the aforementioned problem of collider-stratification that was observed in Figure 1. However, a limitation of inverse probability weighting approaches is that they can perform poorly when some individuals get assigned large weights.

An alternative class of approaches avoids inverse probability weighting by using G-estimation strategies instead. These involve transforming the outcome in a way that removes the mediator's effect on the outcome and thereby the indirect effect. Next, the resulting transformed outcome is regressed on the exposure to obtain a measure of direct effect. This idea has been considered for additive and multiplicative models [8, 4, 13], for logistic regression models [14], for survival models [6], and for unmatched [13, 14] and matched [2] retrospective studies; see Vansteelandt [15] for a detailed review.

4 Identification results for natural direct and indirect effects in the presence of intermediate confounding

These developments on controlled direct effect have a number of limitations. First, the concept of controlling the mediator at level m uniformly in the population is often rather restrictive as it is often difficult to conceptualize a single level of the mediator that is realistic for all units in the population. Second, the difference between the total effect and a controlled direct effect cannot generally be interpreted as an indirect effect [9]. To overcome these limitations, alternative definitions have been proposed of so-called natural direct and indirect effect [9, 7]. These are more natural by allowing for variation between subjects in the level at which the mediator is controlled and, moreover, combine to the total effect regardless of the underlying data distribution. However, natural direct effects require stronger identification conditions than controlled direct effects. In particular, it remains unclear to date how natural direct and indirect effects can be identified in the presence of intermediate confounding, unless in the unrealistic case where the exposure and mediator do not interact (at the unit level) in the effect that they produce on the outcome.

Vansteelandt and VanderWeele [16] overcome this limitation by basing their development on the following definitions of *natural direct and indirect effects in the exposed*:

$$\begin{aligned} &E\{Y - Y(0,M)|A\} \\ &E\{Y(0,M) - Y(0,M(0))|A\}, \end{aligned}$$

respectively. The first expresses, within each exposure stratum, how much the outcome would change on average if the exposure were set to the reference level 0, but the mediator were held fixed at its *observed* level. The second evaluates how much the outcome would change on average if the exposure's effect acted only through modifying the mediator. These definitions enable decomposition of the total effect (in the exposed) into a direct and indirect effect (in the exposed), as follows

$$\begin{aligned} E\{Y-Y(0)|A\} &= E\{Y-Y(0,M(0))|A\} \\ &= E\{Y-Y(0,M)|A\}+E\{Y(0,M)-Y(0,M(0))|A\}. \end{aligned}$$

Vansteelandt and VanderWeele [16] show that natural direct and indirect effects on the exposed allow for effect decomposition under weaker identification conditions than population natural direct and indirect effects. When no confounders of the mediator-outcome association are affected by the exposure, identification is possible under essentially the same conditions as for controlled direct effects. Otherwise, identification is still possible with additional knowledge on a non-identifiable selection-bias function which measures the dependence of the mediator effect on the observed exposure within confounder levels, and which evaluates to zero in a large class of realistic data-generating mechanisms.

Vansteelandt and VanderWeele [16] furthermore argue that natural direct and indirect effects on the exposed are of intrinsic interest in various applications. They moreover show that these natural direct and indirect effects on the exposed coincide with the corresponding population natural direct and indirect effects when the exposure is randomly assigned. In such settings, their results are thus also of relevance for assessing population natural direct and indirect effects in the presence of exposure-induced mediator-outcome confounding, which existing methodology has not been able to address.

Acknowledgements The author acknowledges support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy) and is grateful to Carlo Berzuini (University of Cambridge), Torben Martinussen (University of Copenhagen) and Tyler VanderWeele (Harvard University), with whom parts of this work have been developed.

References

[1] R.M. Baron and D.A. Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, 51:1173–1182, 1986.

[2] C. Berzuini, S. Vansteelandt, L. Foco, R. Pastorino, and L. Bernardinelli. Direct genetic effects and their estimation from matched case-control data. Technical report, University of Cambridge, 2011.

[3] S.R. Cole and M.A. Hernán. Fallibility in estimating direct effects. *International Journal of Epidemiology*, 31:163–165, 2002.

[4] S. Goetgeluk, S. Vansteelandt, and E. Goetghebeur. Estimation of controlled direct effects. *Journal of the Royal Statistical Society, Series B*, 70:1049–1066, 2008.

[5] D.P. MacKinnon. *An Introduction to Statistical Mediation Analysis*. New York: Lawrence Erlbaum Associates, 2008.

[6] T. Martinussen, S. Vansteelandt, M. Gerster, and J.v.B. Hjelmborg. Estimation of direct effects for survival data using the aalen additive hazards model. *Journal of the Royal Statistical Society, Series B*, 73(5):773–788, 2011.

[7] J. Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, pages 411–420, San Francisco, 2001. Morgan Kaufmann.

[8] J.M. Robins. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In *Computation, causation, and discovery*, pages 349–405. AAAI Press, Menlo Park, CA, 1999.

[9] J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

[10] T. J. VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20:18–26, 2009.

[11] T. J. VanderWeele and S. Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2:457–468, 2009.

[12] T.J. VanderWeele. Mediation and mechanism. *European Journal of Epidemiology*, 24:217–224, 2009.

[13] S. Vansteelandt. Estimating direct effects in cohort and case-control studies. *Epidemiology*, 20:851–860, 2009.

[14] S. Vansteelandt. Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models. *Biometrika*, 97:921–934, 2010.

[15] S Vansteelandt. Estimation of direct and indirect effects. In C. Berzuini, P. Dawid, and L. Bernardinelli, editors, *Causal Inference: Statistical Perspectives and Applications*. Wiley and Sons, 2012.

[16] S. Vansteelandt and T.J. VanderWeele. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics, in press*, 2012.