"Douter de tout ou tout croire, ce sont deux solutions également commodes, qui l'une et l'autre nous dispensent de réfléchir."

– Henri Poincaré

Supervisors

Prof. dr. Stijn Vansteelandt

Department of Applied Mathematics, Computer Science and Statistics
Faculty of Sciences

Prof. dr. Tom Loeys

Department of Data Analysis
Faculty of Psychology and Educational Sciences

Prof. dr. Beatrijs Moerkerke

Department of Data Analysis
Faculty of Psychology and Educational Sciences

Other members of the examination board

Prof. dr. Vanessa Didelez

Leibniz Institute for Prevention Research and Epidemiology
University of Bremen, Germany

Prof. dr. Theis Lange

Department of Public Health, Section of Biostatistics
University of Copenhagen, Denmark

Prof. dr. Yves Rosseel

Department of Data Analysis
Faculty of Psychology and Educational Sciences

Prof. dr. ir. Olivier Thas (chair)

Department of Mathematical Modelling, Statistics and Bio-informatics
Faculty of Bioscience Engineering

Dr. Karel Vermeulen

Department of Mathematical Modelling, Statistics and Bio-informatics
Faculty of Bioscience Engineering

Dean Prof. dr. Herwig Dejonghe

Department of Physics and Astronomy
Faculty of Sciences

Rector Prof. dr. Anne De Paepe

Department of Pediatrics and Medical Genetics
Faculty of Medicine and Health Sciences

FACULTY
OF SCIENCES

Flexible causal mediation analysis using natural effect models

Johan Steen

Dissertation submitted in fulfilment of the requirements for the degree of Doctor in Statistical Data Analysis

Academic year 2016–2017

Department of Applied Mathematics,
Computer Science and Statistics

GHENT
UNIVERSITY

This work was supported by
Research Foundation Flanders (G.0111.12)

Data analyzed to illustrate the methods in this thesis were kindly provided by

the Department of Experimental-Clinical and Health Psychology, Faculty of Psychology, Ghent University (Interdisciplinary Project for the Optimisation of Separation trajectories; IPOS).

the World Health Organizations European Centre for Environment and Health, Bonn office (Large Analysis and Review of European Housing and Health Status; LARES). The corresponding chapters in this thesis reflect the author's opinion and not necessarily the position of the WHO.

Contents

Dankwoord

Het heeft heel wat voeten in de aarde gehad. Dat is wel het minste wat je kunt zeggen. Ik denk dat heel wat mensen in mijn nabije omgeving dat inderdaad kunnen beamen. Nochtans, zonder hen was wat nu voor uw neus ligt niet geworden wat het geworden is. Een kleine bedanking is daarom wel op zijn plaats.

Niet één, niet twee, maar *drie* promotoren bleken nodig om bij mij het onderste uit de kan te halen. Bieke, ik heb daarom ook geen moment spijt gehad dat ik een tijdje geleden de nodige formulieren in orde gebracht heb om je ook officieel mijn derde promotor te mogen noemen. Bedankt voor de humoristische en relativerende noot!

Tom, al van tijdens je begeleiding bij mijn master thesis gaf je blijk mijn werk te appreciëren. De wederzijdse waardering maakte de samenwerking de voorbije jaren dan ook enkel aangenamer.

Stijn, van jou heb ik vooral geleerd dat ook moeilijke materie makkelijk verteerbaar wordt dankzij het aanbrengen van wat context en intuïtie. Bedankt vooral om telkens de rust te bewaren wanneer keer op keer moest blijken dat ik een afgesproken deadline niet zou halen. Je kalmte, eindeloze geduld en geloof in een goede afloop hebben ongetwijfeld bijgedragen tot waar ik nu sta. Ik geloof dat ik na deze 4,5 jaar misschien iets meer statisticus ben geworden, en jij iets meer psycholoog. Bedankt, Tom, Bieke en Stijn, voor de kansen die jullie me hebben gegeven. Ik heb veel van jullie geleerd de voorbije jaren!

I'ld also like to express my gratitude towards the members of the reading committee, Vanessa Didelez, Theis Lange, Yves Rosseel, Karel Vermeulen and Olivier Thas, for their careful reading of this thesis, their insightful,

critical and constructive comments and, most of all, for forcing me to take a step back and try to get a grip on the bigger picture.

Theis, thanks for detecting numerous bugs in medflex, for providing valuable input and the opportunity to contribute to the mediation workshop in Copenhagen. But most of all thanks for being such an enthusiastic advertiser of the package!

Yves, ik heb niet enkel een sterk vermoeden, maar ben er redelijk zeker van dat jij mijn interesse in statistiek hebt aangewakkerd.

Karel, jou wil ik ook minstens even hard bedanken voor de tijd als collega-student als voor al je inspanningen als jurylid van m'n proefschrift. Jouw deur stond altijd open op de momenten waarop ik dacht dat ik het onmogelijk kon maken om Stijn nóg maar eens lastig te vallen met technische vragen of verduidelijken. Machteld, ook jij bedankt om je deur telkens op een kiertje te laten. Samen met jullie en Joke waren Boston en New York trouwens een ongelooflijke ervaring!

Bedankt ook aan de vele andere fijne TWIST-collega's, in het bijzonder diegene waarmee ik het bureau gedeeld heb, en niet te vergeten de mensen van het secretariaat. Helaas zijn jullie, samen met de ex-collega's, te talrijk om allemaal bij naam te noemen. Het risico om iemand te vergeten is immers ook te groot. Nonetheless, I would like to especially thank Bashir, Hilmar, Bea, Diego, Oliver, Mushthofa, Gustavo, Holger, Xianming, Koen, Véronique, Vahe, Sjouke, Christophe, José, Camila and Paula, for the good times, lunches, dinners, trips and interesting intercultural discussions. I wish I had been less of a procrastinator, so that I would have had the time to thank each one of you individually.

I'ld also like to express my gratitude towards Ilya Shpitser for the many e-mail replies with technical clarifications on his work. They have surely kept me from getting cross eyed on all those cross-world counterfactual independencies. Many thanks to Kathleen Felix for granting me permission to use the Rube Goldberg cartoon as cover art for this thesis.

Bedankt ook aan de 'experimentele' klasgenootjes voor de vele reünies tussendoor. De boys, bedankt voor de nodige onzin en baldadigheden. De steeds zeldzamer wordende 'banquets' waren iedere keer een verademing. Ik vergeet ook niet snel de legendarische fietstochten (met of zonder

magische handoplegginge). Hopelijk kunnen we gauw die draad terug oppikken! Dank ook aan de bende van de Doornlaan voor de vele fijne etentjes en weekendjes, waar telkens weer naar uitgekeken wordt!

Ann, ook jou kan ik niet genoeg bedanken voor je steun en de altijd warme ontvangst in 'hotel Vorselaar' (zie ook Landuyt D., 2015). Ik hoop dat Marc ook stiekem fier zou geweest zijn. Bedankt, Evelien, voor je onuitputtelijk enthousiasme en je ondernemingszin, die er ongetwijfeld voor zullen zorgen dat de catering een waar succes wordt! Dries, een toffere en handigere schoonbroer kan een mens niet wensen. Wie weet leggen we ons ooit samen toe op Bayesian belief networks?

Ma, pa, hoe vervelend ik het ook vind om in clichés te vervallen, ik kan er niet omheen. Bedankt voor de kansen die jullie me hebben gegeven. Het waren (en zijn) er veel! Ik kan jullie, samen met David en Nathan, niet genoeg bedanken om me doorheen die moeilijke, donkere periode te sleuren. Dank ook aan de rest van de familie, in het bijzonder mijn grootmoeders, voor jullie zorg en onvoorwaardelijke steun.

Carmen, jouw geduld heb ik waarschijnlijk nog het meeste op de proef gesteld. Je immer positieve ingesteldheid heeft altijd zijn effect gehad, ook al had ze soms tijd nodig om onderhuids op me in te werken. Ik kan me geen beter lief inbeelden. Dankzij jou weet ik wat 'geluk' betekent. Dank je voor wat je doet. Maar vooral, voor wie je bent.

Johan Steen
Destelbergen, November 2016

Chapter 1

Introduction

The well-known mantra *association is not causation* has led to the widespread belief that one can only infer causal relations from randomized trials, as they are often considered the gold standard for causal inference.

For example, observational studies in the 1950s reporting associations between smoking and lung cancer have long been criticized for not providing decisive evidence on the supposed causal effect of smoking on lung cancer, because of the simple fact that smokers and non-smokers are different not only in their smoking behavior, but also in many other respects. Both the tobacco industry and some prominent statisticians strongly supported the hypothesis that this association could be explained by a genetic predisposition to both lung cancer and smoking. Although the impact of potential confounding factors, such as a genetic predisposition, is eliminated by design in randomized trials, these designs are often not feasible because of ethical concerns.

Over the last few decades, methodological advances in the causal inference literature have successfully demonstrated that appropriately analyzed data from observational studies may, nonetheless, shed light on causal enquiries. In particular, the *potential outcomes framework* (Splawa-Neyman et al., 1990; Rubin, 1974) has provided a formal language for clarifying and communicating sufficient conditions under which well-defined causal effects can be estimated from the data at hand.

This framework has proven especially useful for the analysis of data

from studies that aim to open the 'black box' of causality in order to deepen our understanding of the precise mechanisms behind established cause-effect relations, as witnessed by the widespread usage of *mediation analyses*. This statistical tool, which is the main topic of this thesis, aims to unravel different causal pathways by separating the component effect that acts through a given intermediate variable or so-called mediator – i.e. an *indirect* effect – from the remaining *direct* effect and by quantifying each of their respective contributions to the overall causal effect. The improved understanding into underlying processes that results from such analyses may not only be of pure scientific or etiologic interest, but may also inform policymakers as to which type of intervention or reform is most effective.

Below, we first list three empirical studies that focused on better understanding of the causal mechanisms behind the effect of a certain intervention or exposure. Each of these examples will be discussed and/or analyzed in more detail in later chapters of this thesis. Next, we briefly introduce the central notion of potential or counterfactual outcomes which naturally leads to formal yet intuitive definitions of the causal effects of interest and enables clearly articulating *causal assumptions* that are required for obtaining unbiased and valid estimates of these effects from observed data. We then provide some intuition into the main challenges in mediation analysis and give a short overview of the contributions of this thesis in terms of dealing with these challenges, followed by a more detailed outline of the subsequent chapters of this thesis.

1.1 Motivating examples

1.1.1 The Job Search Intervention Study (JOBS II)

The JOBS II field experiment (Vinokur et al., 1995), an often cited empirical mediation example, was designed to assess the effectiveness of a theory-driven job training intervention that aimed to both increase reemployment and reduce depressive symptoms in unemployed workers. 1,801 subjects were randomly assigned to either participate in several sessions of job search skills workshops that also focused on enhancing one's sense of

mastery or self-efficacy and inoculation against setbacks after losing one's job (treatment group) or receive a booklet with job search tips (control group).

Vinokur and Schul (1997) conducted a detailed analysis of potential mediating mechanisms after beneficial effects on both reemployment and mental health had been established in earlier analyses (Vinokur et al., 1995). One mediation question of interest was whether workshop participation leads to reduction in depressive symptoms (at two months follow-up) by increasing chances of getting reemployed (at two months follow-up).

1.1.2 The Interdisciplinary Project for the Optimization of Separation trajectories

The Interdisciplinary Project for the Optimization of Separation trajectories (Ghent University and Catholic University of Louvain, 2010) was a large-scale survey study which involved the recruitment of individuals who divorced between March 2008 and March 2009 in four major courts in Flanders. The main aim of this project was to improve the quality of life in families during and after the divorce by translating research findings into practical guidelines for separation specialists (such as lawyers, judges, psychologists, welfare workers...) and by promoting evidence-based policy.

A subsample of 385 individuals responded to a battery of questionnaires related to romantic relationship characteristics, such as adult attachment style, and break-up characteristics, such as break-up initiator status, experiencing negative affectivity and engaging in unwanted pursuit behaviors towards the ex-partner (De Smet et al., 2012). Respondents were asked to imagine their former partner as well as possible and to remember how they generally felt in their relationship *before* the breakup when completing the attachment style questionnaire. The mediation hypothesis of interest concerned the question whether and to what extent the level of emotional distress or negative affectivity experienced *during* the breakup mediates the effect of attachment style towards the ex-partner *before* the breakup exerts on the potential display of unwanted pursuit behaviors *after* the breakup (Loeys et al., 2013).

1.1.3 The Large Analysis and Review of European Housing and Health Status project

The last motivating example also concerns a survey study. The Large Analysis and Review of European Housing and Health Status (LARES) project conducted by the World Health Organization (Shenassa et al., 2007) collected survey data in the winter and spring of 2002/2003 from 5,882 adult respondents from 2,983 households in 8 European cities. Baseline measurements were available on both respondent characteristics (age, gender, marital status, education level, employment, smoking and environmental tobacco smoke at home) and household characteristics (ownership, size, tenure, crowding, ventilation, natural light, heating and city of residence).

One of the mediation questions of interest was whether and to what extent the effect of living in damp and moldy conditions on the risk of depression is mediated by respondent's perceived control over one's home.

1.2 Counterfactual outcomes

The counterfactual or potential outcomes framework appeals to human intuition, because it defines causal effects by comparing an outcome of interest in the population under different hypothetical scenarios or interventions. For instance, in this framework, the causal effect of smoking on lung cancer could be defined as the difference in lung cancer incidence if the entire population were to smoke versus no-one would smoke.

This 'what if' type of reasoning has been formalized by the use of so-called *counterfactual* or *potential outcomes*. For instance, when A denotes the exposure or treatment of interest and Y the outcome of interest, then $Y(a)$ denotes the value of the outcome that would have been observed had A – possibly contrary to the fact – been set to level a. This notation enables defining *total causal effects* as $E\{Y(a) - Y(a')\}$ where a and a' correspond to meaningful choices for active and reference (baseline) levels of treatment or exposure, respectively.[1] For expositional simplicity, we will restrict our

[1] This is essentially identical to the interventional contrast $E(Y|do(A = a)) - E(Y|do(A = a'))$ in terms of Pearl's *do*-operator.

current presentation to binary treatments ($a = 1$ and $a' = 0$), although definitions and results extend to multicategorical or continuous treatments. The population-average effect of smoking A – where $A = 1$ indicates smoking status – on lung cancer Y would thus be defined as $E\{Y(1) - Y(0)\}$.

1.3 Natural direct and indirect effects

Mediation analysis aims to decompose the average treatment or exposure effect, $E\{Y(1) - Y(0)\}$, into the components that respectively capture the treatment's *indirect* effect on the outcome along an intermediate variable of interest M, and the treatment's remaining *direct* effect via potential other mechanisms.

Robins and Greenland (1992) laid the foundations for such decomposition by introducing *nested counterfactuals* $Y(a, M(a'))$, which denote the value of the outcome that would have been observed had – possibly contrary to the fact – A been set to level a and M to $M(a')$, the value that would have been observed for the mediator had A been set to a'. Using such nested counterfactuals, one can now isolate and quantify part of the treatment effect that is transmitted through the mediator M by leaving treatment unchanged at $A = 1$, but changing the counterfactual intermediate outcome $M(1)$ to $M(0)$, the value it would have taken under no treatment, leading to the definition of the so-called *total indirect effect*

$$E\{Y(1, M(1)) - Y(1, M(0))\}.$$

Its complement, the *pure direct effect*

$$E\{Y(1, M(0)) - Y(0, M(0))\},$$

then captures the intuitive notion of blocking the treatment's effect on the mediator by keeping the latter fixed at whatever value it would have attained under no treatment.

For instance, in the motivating example in section 1.1.3, the aim is to

decompose the total exposure effect of mold on mental health, which compares the average risk of depression in the population if everyone were to be exposed to mold versus no-one were exposed. The total indirect effect then captures the average change in risk of depression in the population if everyone's perception of control were to be changed from what it would be under exposure to mold to what it would be under no exposure. The pure direct effect, on the other hand, captures the average change in risk of depression in the population if we were to change everyone's exposure status from being unexposed to being exposed, while leaving unchanged everyone's perceived control at the level that it would be under no exposure.

A primary appeal of these – and similar – effect estimands is that, as opposed to definitions in the linear structural equation modeling tradition, they are model-free: they combine to produce the total effect, irrespective of the scale of interest or the presence of interactions or nonlinearities, under the composition assumption that $Y(a, M(a)) = Y(a)$. For instance, although the above effects are expressed in terms of mean differences, the total effect risk ratio of a binary outcome could similarly be expressed as the product of the pure direct effect risk ratio and the total indirect effect risk ratio

$$\frac{P\{Y(1) = 1\}}{P\{Y(0) = 1\}} = \frac{P\{Y(1, M(0)) = 1\}}{P\{Y(0, M(0)) = 1\}} \frac{P\{Y(1, M(1)) = 1\}}{P\{Y(1, M(0)) = 1\}}.$$

The expectation of nested counterfactuals can be modelled using a so-called *natural effect model* (Lange et al., 2012, 2014; Loeys et al., 2013; Steen et al., 2016a,b; Vansteelandt et al., 2012a), e.g.

$$E\{Y(a, M(a'))\} = g^{-1}\{\beta_0 + \beta_1 a + \beta_2 a' + \beta_3 a a'\},$$

where $g(\cdot)$ is a known link function. If $g(\cdot)$ is the identity link, β_1 captures the pure direct effect and $\beta_2 + \beta_3$ captures the total indirect effect.[2] By differently apportioning the interaction term β_3, an alternative decomposition

[2]Similary, effect estimates on the risk and odds ratio scale can be obtained by choosing $g(\cdot)$ to represent the log and logit link function, respectively.

can be obtained in terms of the *total direct effect*

$$E\{Y(1, M(1)) - Y(0, M(1))\},$$

as captured by $\beta_1 + \beta_3$ and the *pure indirect effect*

$$E\{Y(0, M(1)) - Y(0, M(0))\},$$

as captured by β_2. In accordance with VanderWeele (2013), any of these two decompositions can thus be further refined leading to the same unique three-way decomposition into the pure direct effect β_1, the pure indirect effect β_2, and a mediated interactive effect β_3. Pearl (2001) later adopted the same definitions but named these parameters *natural* (rather than pure) direct and indirect effects to refer to the fact that pure direct effects, as opposed to *controlled* direct effects $E\{Y(1, m) - Y(0, m)\}$, allow for *natural* variation in the mediator. That is, pure direct effects reflect the effect of treatment upon fixing the mediator at values that would, for each individual, have *naturally* occurred under no treatment, rather than at some predetermined level m (uniformly across the population). In the remainder of this thesis, we will adopt Pearl's terminology of 'natural' effects to refer to any of the above instances.

1.4 Challenges in mediation analysis

1.4.1 Causal assumptions

Adopting this counterfactual notation naturally leads to framing causal inference as a missing data problem (Holland, 1986). That is, for each subject i, only one counterfactual outcome, i.e. $Y_i = Y_i(A_i) = Y_i(A_i, M_i(A_i))$, is observed. In order to infer causal effects from observational data, we will thus inevitably need to make some causal assumptions.

Although such assumptions will be discussed more formally and in more detail in the next chapters, an important difference between inferring a total causal effect (in point exposure studies) and, subsequently, learning about its component effects – such as natural direct and indirect effects – merits

attention here. While the former mainly requires that common causes of treatment or exposure and outcome are adjusted for by statistical methods or eliminated by experimental design, the latter, in addition, requires to adjust for common causes of mediator and outcome.

Moreover, additional complexities arise when such mediator-outcome confounders are themselves affected by treatment, because such variables are then simultaneously a confounder and a mediator on the causal pathways that we aim to disentangle. For this reason, causal assumptions generally get more complicated in mediation settings.

For instance, in the motivating example in section 1.1.3, the relation between perceived control over one's household and mental health may be confounded by many factors, such as age, education level, ventilation in the house, etc... Such potential common causes thus need to be taken into account in statistical analyses. However, some of these potential confounders, such as physical health, are likely also affected by exposure to mold (Kaufman, 2010).

As will be discussed in more detail later, the presence of such so-called *intermediate confounders* generally prevents us from obtaining valid estimates of natural direct and indirect effects with respect to the mediator of interest. Nonetheless, in cases with multiple sequential mediators, alternative decompositions of the total effect may still be obtained from the data at hand, in order to shed light on underlying causal mechanisms.

1.4.2 Modeling assumptions

It thus seems that answering mediation questions often, if not always, requires some form of statistical adjustment for confounders. In most applications, the set of confounders will be high-dimensional and will usually consist of a mix of discrete and continuous covariates. To deal with the curse of dimensionality, we will thus necessarily need to rely on some modeling assumptions, preferably as few as possible. A further challenge is that the risk of making incorrect modeling assumptions increases as more and more confounders and mediators enter the picture. Although this challenge is not unique to mediation analysis, semi-parametric approaches,

which allow to relax certain modeling assumptions, have only recently been adapted to this setting (Tchetgen Tchetgen and Shpitser, 2012, 2014; Zheng and van der Laan, 2012).

1.5 Main contributions

In this thesis, we aim to contribute to the fast-growing field of mediation analysis by – at least partially – addressing each of the aforementioned challenges.

First, we give a detailed and up-to-date review of causal assumptions that permit to *identify* – i.e. obtain consistent estimates of – component or path-specific effects of interest from observed data. Recently, significant advances have been made towards a complete characterization of causal scenarios that permit non-parametric identification of natural (and more generally defined path-specific) effects, thus providing both sufficient and necessary conditions (Shpitser, 2013). However, to the best of our knowledge, a systematic comparison of this recent work on complete conditions and earlier work on sufficient conditions (Pearl, 2001) is currently lacking. We contribute to the field by providing such a detailed comparison. In doing so, we aim to offer the reader some deeper intuitive understanding of particular obstacles that may prevent us from making progress in our quest to learn about causal mechanisms. Such improved understanding of necessary causal assumptions – often encoded in graphical models – may 'aid [applied researchers] in planning of data collection and analysis, in communication of results, and in avoiding subtle pitfalls of confounder selection' (Greenland et al., 1999). Importantly, we further reflect upon the specific implications of the completeness of this recent result in terms of complementary identification strategies that rely on so-called mediating instruments. Moreover, we integrate these novel insights with earlier conceptual considerations on the controversial nature of certain key identifying assumptions (Robins and Richardson, 2010).

Second, we provide practical solutions for mediation analysis tailored to the needs of applied researchers. In doing so, we build on a recently proposed unified and flexible modeling framework for mediation analysis

(Lange et al., 2012, 2014; Loeys et al., 2013; Vansteelandt et al., 2012b) that, as compared to other modeling approaches, has the potential to both considerably simplify result reporting and hypothesis testing, and to enable straightforward implementations in standard statistical software. A main contribution of this thesis, in this respect, is the development of a user-friendly software package that implements two proposed semi-parametric estimators within this modeling framework (Steen et al., 2016b), each of which reduces modeling demands by allowing to refrain from modeling certain aspects of the observed data distribution. Importantly, this package handles a larger class of parametric models for mediator and outcome than alternative software applications for modern mediation analysis that rely on closed-form expressions (Valeri and VanderWeele, 2013), and is less computer-intensive as compared to implementations that rely on Monte Carlo integration (Imai et al., 2010a; Tingley et al., 2014a). The latter asset is, in part, due to the development and implementation of robust sandwich variance estimators, which permit to avoid reliance on bootstrap procedures. Finally, we further extend this *natural effect modeling* framework, along with semi-parametric estimators, to accommodate more complex mediation settings with multiple, causally ordered mediators (Steen et al., 2016a). In particular, we demonstrate that such an extension both enables a more comprehensive assessment of underlying mechanisms and their potential interactions, as compared to existing analytical approaches (VanderWeele and Vansteelandt, 2013), and reduces modeling demands – and thus risk of model misspecification bias – as compared to fully parametric approaches (e.g. Daniel et al., 2015). Moreover, it offers a more principled solution to cope with increasing complexity in the face of multiple mediators. In addition, we propose a sufficient criterion for identification of $(k+1)$-way decompositions in the presence of k sequential mediators. This criterion extends previous work, as it boils down to sequential application of an existing graphical identification criterion for adjustment for a common set of covariates (Shpitser et al., 2010; Shpitser and VanderWeele, 2011), leading to a standard and generally applicable identification result. Its simplicity can be considered to induce a trade-off between general applicability and reduced identification power.

1.6 Outline of this thesis

In the next two chapters, we mainly focus on causal assumptions.

In **chapter 2**, we first introduce the necessary theoretical background on *graphical causal models*, which are commonly used to visually encode and communicate the causal assumptions that serve to provide certain statistical parameters a causal interpretation. Moreover, we review some important algorithms that have mainly been developed within the field of artificial intelligence, but that can be widely applied in any field of empirical research that attempts to address causal queries. Their importance follows from the fact that, whereas often sufficient conditions are articulated, these algorithms enable to deduce conditions that are both sufficient and necessary for identifying total causal effects from available observed data, thus providing a (more) complete characterization of hypothetical causal scenarios that permit identification. As discussed in more detail in this chapter, this is of particular relevance for graphical causal models that considerably weaken certain causal assumptions by allowing for the presence of unobserved common causes.

In **chapter 3**, we provide intuition into the distinct and controversial nature of some of the identifying assumptions for mediation analysis. In particular, we revisit earlier assumptions for identifying natural direct and indirect effects (Pearl, 2001) in the light of recent developments (Shpitser, 2013) that build on the insights and algorithms discussed in chapter 2. Importantly, we point out that these recent developments also lead to novel insights that are in line with and help to frame some recent conceptual formulations that were inspired by the debate about the controversial nature of the targeted effects.

In the remaining chapters, we shift focus to flexible modeling and estimation of the causal effects of interest. In **chapter 4**, we discuss estimation of so-called natural effect models (Lange et al., 2012, 2014; Loeys et al., 2013; Vansteelandt et al., 2012b), which were recently introduced in the literature to offer a simple yet flexible alternative to other state-of-the-art modeling approaches that, from the perspective of an applied researcher, may either complicate obtaining interpretable results or hypothesis testing (Imai et al.,

2010a) or pose a barrier to routine application because of their relative complexity (Tchetgen Tchetgen and Shpitser, 2011; van der Laan and Petersen, 2008). In this chapter, we moreover give a detailed discussion of the features of medflex, our free, open-source software package for R, which implements two proposed semi-parametric estimators within this modeling framework.

More general methods for mediation analysis are provided in **chapter 5**, in which we extend the natural effect modeling framework to settings with multiple sequential mediators. Not only does such an extension offer feasible alternative decompositions in settings in which the mediator of interest is subject to intermediate confounding, it also enables parsimonious modeling, which may be advocated given the multitude of possible decompositions in the presence of an increasing number of mediators.

We conclude in **chapter 6** with some further reflections and challenges.

Individual contributions

The major parts of this dissertation are based on two accepted papers, one submitted handbook chapter and a software package. Although the aim is to present a coherent and well-structured overview of my research, inevitably, by merging these papers and chapters, which may not all have been presented in chronological order of writing, some repetition and loss of continuity may arise. In this subsection, a chronological overview is presented of the work in this thesis, along with a list of my individual contributions to each of the chapters, excluding the introduction and discussion (hence the switch in narrative voice).

Chapter 4 and **chapter 5** can be considered as a product of a close collaboration with Stijn Vansteelandt, Tom Loeys and Beatrijs Moerkerke. I have developed and documented the medflex package, which implements the methods in Lange et al. (2012) and Vansteelandt et al. (2012b) and is currently available[3] from CRAN: `https://cran.r-project.org/package=medflex`. In order to ensure both compatibility with future extensions of the package and optimal user experience, certain crucial choices had to be made, mainly with respect to the core structure of the package. These choices have greatly benefited from close consultation with S. Vansteelandt, T. Loeys and B. Moerkerke. Valuable input, especially concerning the

[3]Up-to-date development releases of the package are available from `https://github.com/jmpsteen/medflex/`.

`neWeight` function, has also been provided by Theis Lange. Occasional technical support in the developing stages of the package has been provided by Joris Meys. Several bugs have been reported by S. Vansteelandt, T. Loeys, T. Lange, as well as by users of the package (a more detailed list, along with some patches provided by users, can be found at `https://github.com/jmpsteen/medflex/issues`). S. Vansteelandt and Karel Vermeulen have provided guidance in constructing generic robust sandwich variance estimators for combinations of a wide class of parametric models (with canonical link functions).

Chapter 4 provides a detailed user guide for the package, using a dataset that has also been used in Loeys et al. (2013) as an illustrating example. The theoretical content of this chapter is largely based on Vansteelandt et al. (2012b), Lange et al. (2012) and Loeys et al. (2013) (a paper to which I have also contributed). I have taken the lead in writing this chapter, which is available as a vignette to the package, in a slightly modified version, and has been accepted for publication in *Journal of Statistical Software* (Steen et al., 2016b).

I have also taken the lead in writing **chapter 5**, although S. Vansteelandt has made major contributions in rewriting parts of this chapter in order to make it more accessible for an epidemiologic audience. The estimation procedure and graphical translation of identifying assumptions into a generalization of the adjustment criterion (Shpitser et al., 2010; Shpitser and VanderWeele, 2011) (in the technical appendix) were developed by myself, with guidance from S. Vansteelandt, T. Loeys and B. Moerkerke. In addition, I have implemented all R code in the technical appendix, and have conducted all data analyses. This chapter has been accepted for publication in *American Journal of Epidemiology* (Steen et al., 2016a).

The content of **chapter 2** is largely based on other introductory texts including Elwert (2013), Pearl (2000), Pearl et al. (2016), and Tian and Shpitser (2010).

Chapter 3 is based on a chapter that has recently been submitted for peer review and is to appear in M. Drton, S. Lauritzen, M. Maathuis, M. Wainwright (Eds.), *Handbook of Graphical Models.* CRC Press. The detailed comparison of identifying assumptions, novel insights and relation with Robins and Richardson (2010), as mentioned in section 1.5, are mainly individual contributions. Ilya Shpitser has helped a great deal in the shaping of this chapter by providing valuable clarifications regarding his paper in Cognitive Science (Shpitser, 2013). S. Vansteelandt has sigificantly contributed by improving the structure and clarity of earlier versions of this chapter.

Chapter 2

Inferring causal effects from observed data

Over the years, graphical models have proven to be an indispensable tool for visualizing and communicating causal assumptions within a given research context. Such models typically consist of a causal diagram or causal directed acyclic graph (DAG) $\mathcal{G}$ with nodes (or vertices) $V = \{V_1, ..., V_n\}$ representing random variables of interest and directed edges (or arrows) connecting these nodes.[1]

2.1 Encoding conditional independencies in a graph

These diagrams are used to visualize a set of assumed conditional independencies. More specifically, whereas arrows between variables encode probabilistic dependencies among those variables, the absence of an arrow translates into an assumption of conditional independence stating that each variable V_i is independent of its non-descendants conditional on its parents PA_i in the graph (i.e. the variables that have an arrow feeding directly into V_i). This *Markov assumption* allows linking the structure of the graph to the observed data on V. In particular, these conditional independence assumptions impose a set of restrictions on the joint probability distribution

[1] Typically, kinship terminology (i.e. 'parents', 'children', 'ancestors' and 'descendants') is used to describe the relationships between nodes implied by the arrows connecting them. By convention, we will denote V_i to be both an ancestor and a descendant of V_i.

of V, $P(V)$ so that it factorizes as a product of conditional distributions $P(V_i|PA_i)$ which only involve the parents PA_i for each V_i:

$$P(V) = \prod_i P(V_i|PA_i), \tag{2.1}$$

such that $P(V)$ satisfies the global Markov property relative to $\mathcal{G}$ (see next section).

Consider, for instance, the diagram in Figure 2.1A with $V = \{C, A, M, Y\}$. It follows from the Markov assumption relative to this diagram that M and C are conditionally independent given A

$$P(M|A, C) = P(M|A) \tag{2.2}$$

denoted, $M \perp\!\!\!\perp C|A$, and that Y and A are conditionally independent given $\{M, C\}$, i.e. $Y \perp\!\!\!\perp A|M, C$,

$$P(Y|A, M, C) = P(Y|M, C). \tag{2.3}$$

$P(V)$ thus factorizes as

$$P(C, A, M, Y) = P(Y|M, C)P(M|A)P(A|C)P(C).$$

2.1.1 *d*-separation

In this simple example, all conditional independencies encoded in the graph follow directly from the local Markov property. More generally, Pearl (1988)'s *d*-separation criterion provides a graphical rule that enables

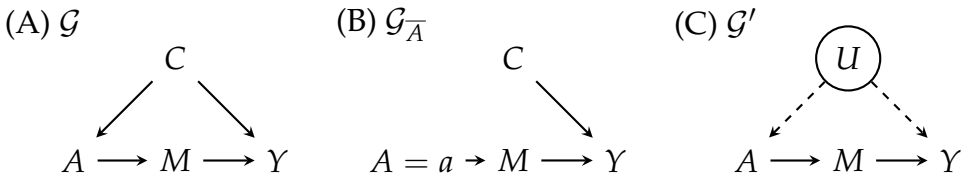


Figure 2.1: Original graph $\mathcal{G}$ (A), mutilated graph $\mathcal{G}_{\overline{A}}$ (B), and graph $\mathcal{G}'$ with C replaced by unobserved U (C).

summarizing all (conditional) independencies encoded in a given graph, irrespective of its complexity. To fully appreciate this rule, however, one needs to distinguish three elementary causal structures, which can be considered the building blocks of every causal DAG. Each of these structures corresponds to a different source of association between observed variables.

Confounding and – especially – causation are two potential sources of association that match relatively well with human intuition. Their causal structures correspond to chains $V_i \rightarrow V_j \rightarrow V_k$ and forks $V_i \leftarrow V_j \rightarrow V_k$, respectively. In both of these structures V_i and V_k are marginally dependent, but conditionally independent given V_j. That is, V_i and V_k are said to be *d-connected*. If these causal structures were viewed as an electric net (Shipley, 2002), in both cases, V_j could be considered an active switch that enables electricity to be transmitted between V_i and V_k along their connecting edges. The circuit can be broken by turning off the switch. Similarly, the path connecting V_i and V_k can be blocked upon conditioning on V_j, rendering V_i and V_k *d-separated*.

A third type of association, in contrast, arises when conditioning on a third variable. That is, if the structure is an inverted fork $V_i \rightarrow V_j \leftarrow V_k$, V_i and V_k are marginally independent, but they become dependent when conditioning on their common effect V_j. Nodes with converging edges, so-called *colliders*, such as V_j, act like inactive switches that do not transmit electricity, unless they are conditioned on. In this case, the blocked path between V_i and V_k is unblocked, rendering these formerly *d*-separated nodes *d*-connected. Conditioning on a collider may thus induce artificial or spurious associations. This seems to be at odds with human intuition (Burns and Wieth, 2004), as many would assume that conditioning on a third variable would, if anything, reduce or eliminate any dependence. A simple example may, however, help to elucidate this counterintuitive phenomenon (Pearl, 2000). Suppose the admission criteria for a graduate school are high grades and/or unusual musical talent and suppose one may assume these attributes to be uncorrelated in the general population. Learning that a random person has obtained high or low grades is thus uninformative as to whether this person has unusual musical talent (and vice versa). However, learning that a student of that school has obtained low grades tells us that

this student must be exceptionally gifted in music. Likewise, students that are not musically talented, are more likely to have obtained higher grades. These two causal attributes, which are uncorrelated (or marginally independent) in the general population, thus become dependent upon learning about their common consequence, i.e. that a student has gained admission. This phenomenon, which also occurs when conditioning on a descendant of a collider, has been termed *Berkson's paradox* in epidemiology and statistics (Berkson, 1946) or the *explaining away effect* in artificial intelligence (Kim and Pearl, 1983). Other commonly used terms are *collider(-stratification) bias* (Greenland, 2003) or *selection bias* (Hernán et al., 2004). The latter terms clarify, as in the above example, that this bias may not only occur because of, for instance, regression adjustment, but also by selective sampling from a specific subpopulation (i.e. stratification).

In contrast to the graphs associated with these three elementary structures, most graphs are of considerably higher complexity, containing both more nodes and more edges. In particular, two nodes possibly have multiple paths[2] connecting them, each of which may contain any combination of these structures and may hence be blocked or unblocked by a set of other nodes. Given these elementary structures, however, we may predict the dependencies encoded in a graph of any level of complexity, using the following graphical criterion.

Definition 2.1.1. *d-separation (Pearl, 2000) A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if*

(i) p contains a chain $V_i \to V_j \to V_k$ or a fork $V_i \leftarrow V_j \to V_k$ such that the middle node V_j is in Z, or

(ii) p contains an inverted fork $V_i \to V_j \leftarrow V_k$ such that the middle node V_j is not in Z and such that no descendant of V_j is in Z.

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

[2] A path is a sequence of distinct nodes where any two adjacent nodes in the sequence are connected by an edge (of any directionality).

While X and Y are said to be conditionally independent given Z if they are d-separated by Z, the converse does not necessarily hold. For instance, d-connected nodes may be independent if an exact cancellation of positive and negative effects occur. Because such exact cancellations are unlikely to occur, it is usually assumed that d-connected nodes are dependent, an assumption referred to as *faithfulness* (Spirtes et al., 1993).

2.1.2 Observational equivalence

Importantly, since conditional independencies encoded in a graph impose constraints on the probability distribution that governs the generated data, they can be tested from observed data on the variables in the graph. This enables us to partially test the validity of the causal model associated with a given graph, but also serves as the basis for causal discovery algorithms. However, the ability to falsify a given graphical model from observable data does usually not permit to distinguish between multiple graphs that are compatible with observed data.

For instance, the three graphs in Figure 2.2 encode the same set of conditional independencies, i.e. $X \perp\!\!\!\perp Y|W$ and $W \perp\!\!\!\perp Z|X,Y$. Because they share an identical set of testable implications, observational data does not carry any information to decide which of the three graphical models reflects the true underlying data generating mechanism. This example illustrates that conditional independencies usually do not allow us to infer directionality for all edges on a given graph. Nonetheless, we can infer some information about directionality, in each of the three graphs. Since we may learn from observed data that $X \perp\!\!\!\perp Y|W$ and that $X \not\perp\!\!\!\perp Y|W,Z$, we can infer that Z must be a collider and hence that the edges between Z and

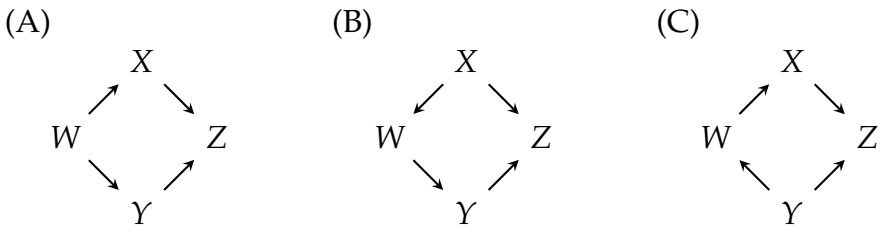


Figure 2.2: Three graphs that belong to the same Markov equivalence class.

X and between Z and Y must be pointing towards Z. Directionality may thus to some extent be inferred by discovering so-called v-structures (i.e. colliders whose parents are not adjacent).

Graphs that share a common skeleton (i.e. the same configuration of edges, irrespective of their direction) and common v-structures, such as the graphs in Figure 2.2, are said to be *observationally equivalent* or to belong to the same *Markov equivalence class* (Verma and Pearl, 1991). That is, because they share an identical set of conditional independencies, they are empirically indistinguishable. To assess the causal effect of, say W on Z, it is, however, crucial to distinguish between each of these graphs. Necessarily, to make progress, we will need to make certain assertions about directionality based on subject matter knowledge and/or expert judgment.

2.2 What makes a diagram a *causal* diagram

Since the notion of causation is often formalized by referring to hypothetical interventions, e.g. setting A to a, we ultimately wish to learn about some aspects of the joint distribution of the other observed variables $P(V \setminus A)$ (i.e. usually the mean of some outcome $Y \in V$) under such different interventions in the population. Our ability to do so rests on the assumption that the directed edges in a graph represent causal influences between the corresponding variables and that the graph can be conceived to reflect a modular system, in the sense that one can manipulate or change one part of the system without affecting the rest. More specifically, this *invariance property* states that each parent-child relation represents a stable and autonomous physical mechanism. The ideas of intervention and *modularity* match the intuitive notion of causation and conditions that enable turning purely correlational claims into causal ones. These are therefore considered to grant causal DAGs their causal interpretation.

Consider again, for example, the graph in Figure 2.1A. If we were to intervene locally on A, fixing it to a, we would only curtail A's natural tendency to vary in response to C (e.g. a potential confounder), without affecting the natural responses of the other variables. This action is often represented graphically by performing a kind of surgery on the original

graph $\mathcal{G}$, turning it into $\mathcal{G}_{\overline{A}}$, by removing all directed edges into A (as in Figure 2.1B), or mathematically, using Pearl (2000)'s *do*-operator, where $do(A = a)$ represents the action or intervention that fixes A to a. In order to learn about causal effects, we thus aim to compare joint interventional distributions $P(V \setminus A|do(A = a))$ – or interventional distributions of an outcome of interest $P(Y|do(A = a))$ – corresponding to different hypothetical interventions enforced uniformly over the population.

2.3 The truncated factorization formula

Importantly, assuming modularity enables us to obtain the joint interventional distribution by applying the usual factorization to the manipulated graph $\mathcal{G}_{\overline{A}}$

$$P(V \setminus A|do(A = a)) = \prod_{i|V_i \notin A} P(V_i|PA_i)I(A = a), \tag{2.4}$$

since the factors $P(V_i|PA_i)$ corresponding to variables in A are either 1 (when $A = a$) or 0 (when $A \neq a$), while those corresponding to the other variables remain unaltered. It can be seen that the resulting *truncated factorization formula* (Pearl, 1995a) – which has been referred to earlier as the *g-computation formula* (Robins, 1986) and is implied by the *manipulation theorem* (Spirtes et al., 1993)) – in expression (2.4) simply omits (from expression (2.1)) the conditional distribution of the node A that we intervene on. The interventional distribution of some outcome of interest Y can then simply be obtained by summing[3] expression (2.4) over $V \setminus \{A, Y\}$

$$P(Y|do(A = a)) = \sum_{v \setminus \{a,y\}} \prod_{i|V_i \notin A} P(V_i|PA_i = pa_i), \tag{2.5}$$

where pa_i denotes the vector of value assignments to PA_i such that, if $A \in PA_i$, value assignment $PA_i = pa_i$ is consistent with $A = a$. Note that, in the absence of hidden variables, the modularity assumption implies

[3]Throughout, for continuous V_i, replace summations by integrals and probabilities by density functions.

$P(V_i|PA_i) = P(V_i|do(PA_i))$ for each V_i, such that the truncated factorization in expression (2.5) can be rewritten in terms of interventional distributions

$$P(Y|do(A = a)) = \sum_{v\setminus\{a,y\}} \prod_{i|V_i \notin A} P(V_i|do(PA_i = pa_i)). \tag{2.6}$$

2.3.1 An example

Suppose, for example, that the variables in Figure 2.1A, as in Pearl (2000), represent smoking A, amount of tar deposited in the lungs M, development of lung cancer Y and a certain genotype C that predisposes to both smoking and developing lung cancer. Application of the truncated factorization formula yields that, under the assumptions encoded in the graph in Figure 2.1A, the interventional distribution of Y under an intervention that would, irrespective of potential ethical objections, either ban, i.e. $do(A = 0)$, or enforce smoking, i.e. $do(A = 1)$ – or more generally, $do(A = a)$ – in the general population equals

$$P(Y|do(A = a)) = \sum_{c,m} P(Y|M = m, C = c)P(M = m|A = a)P(C = c).$$

Moreover, exploiting the conditional independencies (2.2) and (2.3) encoded in the graph, we can simplify this resulting expression as follows:

$$\begin{aligned}
&\sum_{c,m} P(Y|A = a, M = m, C = c)P(M = m|A = a, C = c)P(C = c) \\
&= \sum_{c,m} P(Y, M = m|A = a, C = c)P(C = c) \\
&= \sum_{c} P(Y|A = a, C = c)P(C = c).
\end{aligned} \tag{2.7}$$

This yields an expression commonly referred to as the *adjustment formula* or the *back-door formula* (Pearl, 1993).

2.4 The adjustment formula

The previous example illustrates that, in some cases, the identification result for $P(Y|do(A = a))$ obtained via the truncated factorization formula (in expression (2.5)) may be simplified to expression (2.7).

2.4.1 Conditional ignorability

This result can, in fact, be shown to naturally relate to a sufficient condition for identification of causal effects defined in the counterfactual outcomes framework, i.e. that of *conditional ignorability*. This assumption, denoted as a conditional independence statement involving counterfactual outcomes

$$Y(a) \perp\!\!\!\perp A|C, \qquad \text{for all } a \tag{2.8}$$

states that the counterfactual outcome $Y(a)$ that – possibly contrary to the fact – would have been observed under intervention that sets $A = a$, does not depend on the actual level A within strata of a set of covariates C. Assumption (2.8) has also been named the assumption of *no omitted confounders* or *no unmeasured confounding*, to capture the more intuitive notion that C constitutes a sufficient set to adjust for potential confounding of the relation between A and Y.

When combined with a *consistency assumption*, which states that $Y = Y(a)$ if $A = a$, conditional ignorability (2.8) allows the counterfactual distribution $P(Y(a))$ – which essentially corresponds to $P(Y|do(A = a))$ – to be expressed by the adjustment formula (2.7) as follows:

$$\begin{aligned} P(Y(a)) &= \sum_c P(Y(a)|C = c)P(C = c) \\ &= \sum_c P(Y(a)|A = a, C = c)P(C = c) \\ &= \sum_c P(Y|A = a, C = c)P(C = c). \end{aligned}$$

2.4.2 The adjustment criterion

Shpitser et al. (2010) provided a complete graphical criterion for identifi-

cation of $P(Y|do(A = a))$ by the adjustment formula (2.7); a criterion that, in other words, permits to find all possible adjustment sets C that satisfy conditional ignorability (2.8). This *adjustment criterion* has been shown to generalize and subsume Pearl (1995a)'s *back-door criterion*.[4]

In order to provide a more precise and formal definition of this criterion, especially in the case where A may be a joint or sequential intervention, as in the examples discussed below, we will need to introduce the following terminology.

Definition 2.4.1. *Proper causal path (Shpitser et al., 2010) Let X, Y be sets of nodes. A directed path from a node in $A \in X$ to a node in Y is called proper causal with respect to X if it does not intersect X except at A.*

More generally, a path from X to Y is called *proper* if only its first node is in X (Perković et al., 2015). For example, suppose $X = \{A, M\}$ in the graphs in Figure 2.3. In the graph in panel (A), there are two proper causal paths from X to Y, i.e. $A \rightarrow Y$ and $M \rightarrow Y$. Note that $A \rightarrow M \rightarrow Y$ is not proper causal with respect to X because it intersects X at M. In the graph in panel (B), there is an additional proper causal path from X to Y, i.e. $A \rightarrow L \rightarrow Y$.

Definition 2.4.2. *Adjustment criterion (Shpitser et al., 2010) Z satisfies the adjustment criterion relative to (X, Y) in the original graph $\mathcal{G}$ if*

(i) No element in Z is a descendant in $\mathcal{G}_{\overline{X}}$ of any $W \notin X$ which lies on a proper causal path from X to Y, and

(ii) All proper[5] non-causal paths in $\mathcal{G}$ from X to Y are blocked by Z.

The only non-causal path from $\{A, M\}$ to Y in the graph in Figure 2.3A is $M \leftarrow C \rightarrow Y$. This path can be blocked by C, which is not on a proper causal path from $\{A, M\}$ to Y, nor is it a descendant of a node on such a proper causal path. So C satisfies the adjustment criterion relative to

[4] For this reason, the back-door criterion is not further discussed.

[5] Shpitser et al. (2010)'s original formulation claimed that all non-causal paths in $\mathcal{G}$ from X to Y should be blocked by Z. However, in accordance with Perković et al. (2015), we provide a slight reformulation in which this is only required for all *proper* non-causal paths.

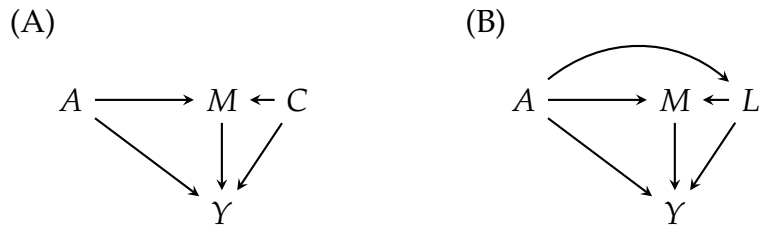


Figure 2.3: Two mediation graphs with different proper causal paths from $\{A, M\}$ to Y.

$(\{A, M\}, Y)$ in this graph, such that $P(Y|do(A = a, M = m))$ is identified by

$$P(Y|do(A = a, M = m)) = \sum_c P(Y|A = a, M = m, C = c)P(C = c).$$

Likewise, in the graph in Figure 2.3B, L blocks the only non-causal path from $\{A, M\}$ to Y, i.e. $M \leftarrow L \rightarrow Y$. However, L lies on the proper causal path $A \rightarrow L \rightarrow Y$ in $\mathcal{G}_{\overline{AM}}$ and thus does not satisfy the adjustment criterion relative to $(\{A, M\}, Y)$ in this graph. Nonetheless, $P(Y|do(A = a, M = m))$ can be computed from the observed data by expression (2.5), which yields

$$P(Y|do(A = a, M = m)) = \sum_l P(Y|A = a, M = m, L = l)P(L = l|A = a).$$

Intuitively, these examples illustrate that the first part of the adjustment criterion keeps us from adjusting for mediators, whereas the second part ensures that we adjust for common causes.

2.4.3 Flexible estimation strategies for the adjustment formula

Most often interest lies in comparing some mean outcome of interest under different hypothetical interventions in the population. That is, $E(Y|do(A = a))$ is the causal quantity of interest, rather than the interventional distribution $P(Y|do(A = a))$ *per se*. Estimating this quantity from observed data via direct application of the adjustment formula may be cumbersome, as it requires modeling $P(C = c)$. This can be challenging, especially when C contains continuous covariates and/or high-dimensional and data is sparse.

Below we show that there are two ways of rewriting the adjustment formula that give rise to estimators that may considerably reduce modeling demands in the sense that neither require modeling $P(C = c)$.

Inverse probability weighting

The first estimator arises from rewriting the adjustment formula as follows

$$\begin{aligned} E(Y|do(A = a)) &= \sum_{y,c} y \cdot P(Y = y|A = a, C = c)P(C = c) \\ &= \sum_{y,c} \frac{y \cdot P(Y = y, A = a, C = c)}{P(A = a|C = c)} \\ &= \sum_{y,c} \frac{y \cdot P(Y = y, C = c|A = a)P(A = a)}{P(A = a|C = c)} \\ &= E\left[\frac{YI(A = a)}{P(A = a|C)}\right]. \end{aligned}$$

The corresponding sample estimator

$$n^{-1} \sum_{i=1}^{n} \frac{Y_i I(A_i = a)}{\hat{P}(A_i = a|C_i)}$$

corresponds to a weighted mean outcome, where each individual exposed at level $A = a$ is weighted by the inverse of its propensity of being exposed at that exposure level given baseline covariates C, $\hat{P}(A = a|C)$. Inverse weighting can be thought of aiming to construct a pseudo-population in which confounding by C is eliminated (i.e. mimicking a randomized trial). This weighted-based estimator thus focuses solely on modeling the relation between A and C as it only requires a propensity score model for $P(A|C)$.

Imputation

The second estimator results from simply applying the law of iterated expectations, so that one can average over the empirical distribution of C in

the observed data, as follows:

$$\begin{aligned} E(Y|do(A=a)) &= \sum_c E(Y|A=a,C=c)P(C=c) \\ &= E\left[E(Y|A=a,C)|A=a\right]. \end{aligned}$$

The resulting expression gives rise to an imputation-based estimator

$$n^{-1}\sum_{i=1}^{n}\hat{E}(Y_i|A_i=a,C_i)$$

that requires imputing each individual's outcome under observed levels of the covariate set C but a (possibly) counterfactual exposure level a. $E(Y|do(A=a))$ can then be estimated by simply calculating the mean of these imputed outcomes. This estimator thus focuses on modeling the relation between Y and C within strata of A as it only requires an imputation model for the mean outcome $E(Y|A,C)$.

Marginal structural models

$E(Y|do(A=a))$ or $E(Y(a))$ can be parameterized using so-called *marginal structural models* (Robins, 1999; Robins et al., 2000). The parameters of such models correspond to interventional contrasts of interest. For instance, in the marginal structural model

$$E(Y(a)) = \beta_0 + \beta_1 a, \tag{2.9}$$

β_1 captures the average causal effect corresponding to a change in the exposure from $A=0$ to $A=a$, i.e. $E(Y(a)-Y(0))$.

Model (2.9) could be considered a special case of a wider class of generalized linear marginal structural models

$$E(Y(a)) = g^{-1}\{\beta^\top W(a)\} \tag{2.10}$$

with $W(a)$ a known vector with components that may depend on a. W may be specified so as to accommodate non-linearities in the case of a

continuous exposure. β is an unknown parameter vector and $g(\cdot)$ a known link function, the choice of which permits some flexibility as to the scale on which the causal effect of interest is desired to be expressed.

The marginal structural model framework provides a natural environment for implementing the aforementioned estimators. That is, marginal structural models are traditionally fitted by weighted regression models, in which the weights correspond to the inverse probability weights discussed in section 2.4.3 (Robins et al., 2000). Alternatively, one may regress imputed mean outcomes on the exposure (Snowden et al., 2011). The latter approach is, however, computationally more intensive, as it requires replicating the original data along multiple values of the exposure and imputing outcomes for each individual under each of these exposure levels.

In chapter 3, similar estimators will be developed for estimating natural direct and indirect effects in a mediation context. Similarly, marginal structural models will be generalized to parameterize mean nested counterfactuals $E(Y(a, M(a')))$. The motivation for these extensions follows from the fact that the adjustment criterion can be generalized to covariate sets that enable identifying natural direct and indirect effects by a generalized adjustment formula for mediation analysis (Shpitser and VanderWeele, 2011).

2.5 Identifiability in the presence of hidden variables

When all relevant variables are observed, all causal queries of the form $P(Y|do(A = a))$ can be computed from the observed joint distribution $P(V)$ via the truncated factorization formula (expression (2.4)). However, the assumption that all common causes of any two (or more) variables in the graph are also included in the graph, i.e. that of *causal sufficiency*, is often unrealistic because it dismisses the possibility of unmeasured confounding. Whenever we relax this assumption, the question of *identifiability* arises, i.e. whether $P(Y|do(A = a))$ can be expressed as a function of the joint distribution of observed variables $P(V)$.

2.5.1 A simple example: the front-door formula

Consider again the smoking example. Suppose the genetic predisposition for both smoking and developing lung cancer is unmeasured and named U, as in Figure 2.1C. The graphical model associated with this causal diagram can be considered a *semi-Markovian* model.[6] Often, semi-Markovian models are represented by *acyclic directed mixed graphs* (ADMGs) (Richardson, 2003), where the presence of an unobserved common cause of two nodes is indicated by bi-directed edges ($\leftrightarrow$). However, for the purpose of our presentation, we will explicitly represent hidden variables U by circled nodes and their direct effects on observed variables V by dashed edges.

Since U is unobserved, the adjustment criterion cannot be satisfied.[7] Likewise, the truncated factorization formula (expression (2.4)) yields

$$\begin{aligned} P(Y|do(A=a)) &= \sum_{u,m} P(M=m, Y, U=u|do(A=a)) \\ &= \sum_{u,m} P(Y|M=m, U=u)P(M=m|A=a)P(U=u), \end{aligned} \tag{2.11}$$

which involves U and thus cannot be evaluated. However, progress can be made upon noting that, when recovering the joint distribution $P(V)$ by summing over U, factors involving observed variables without unobserved parents, such as M, 'factor out' of the summation, as follows:

$$\begin{aligned} P(A, M, Y) &= \sum_{u} P(Y|M, U=u)P(M|A)P(A|U=u)P(U=u) \\ &= P(M|A)\sum_{u} P(Y|M, U=u)P(A|U=u)P(U=u). \end{aligned} \tag{2.12}$$

The joint distribution $P(A, M, Y)$ can thus be written as the product of

[6] A model whose corresponding graph only includes unobserved variables that have (i) no parents (i.e. is a root node) and (ii) exactly two observed children, is called a semi-Markovian model. Even though identification results and algorithms described below can be extended to more general Markovian models with arbitrary sets of unobserved variables upon obtaining a semi-Markovian projection of these models (Tian and Pearl, 2003), for ease of exposition, throughout this thesis, we will focus on semi-Markovian models.

[7] Also note that the graph in Figure 2.1C carries no more testable implications since all conditional independencies encoded in the graph involve U.

$P(M|A)$ and a factor that involves the confounded nodes A and Y.

2

A key observation is that, despite U being unobserved, the second factor can be expressed in terms of the observed data $V = \{A, M, Y\}$, as follows (from expression (2.12)):

$$\sum_u P(Y|M, U = u)P(A|U = u)P(U = u) \\ = P(A, M, Y)/P(M|A) = P(Y|A, M)P(A). \quad (2.13)$$

Moreover, because no factors in the summation (over U) depend on A, we can rewrite expression (2.11) as

$$\sum_m P(M = m|A = a) \sum_{a',u} P(Y|M = m, U = u)P(A = a'|U = u)P(U = u),$$

which, by expression (2.13) reduces to

$$\sum_m P(M = m|A = a) \sum_{a'} P(Y|A = a', M = m)P(A = a'), \quad (2.14)$$

an expression generally referred to as the *front-door* formula (Pearl, 1995a).

This example illustrates that, at least in some settings, we may still be able to identify $P(Y|do(A = a))$ from $P(V)$, despite the presence of unmeasured confounding. In fact, as will be elucidated in section 2.5.3, identification via the front-door formula can be considered to arise via sequential application of the adjustment formula, by which $P(M|do(A = a))$ is identified by $P(M|A = a)$ via adjustment for the empty set, while $P(Y|do(M = m))$ is identified by $\sum_a P(Y|A = a, M = m)P(A = a)$ via adjustment for A. A crucial assumption here, though, is that M intercepts all directed paths from A to Y, or in other words, that M mediates the entire effect of A on Y. If this *exclusion restriction* would not hold, we could not have written expression (2.11) as expression (2.14) and we would not have obtained identification.

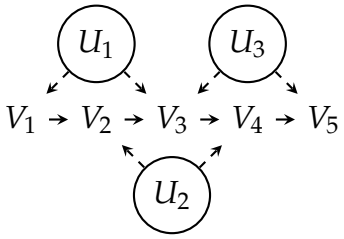


Figure 2.4: Graph for a semi-Markovian model with c-components $\{V_1, V_3, V_5\}$ and $\{V_2, V_4\}$.

2.5.2 C-component factorization

The factorization in expression (2.12) moreover illustrates that the set of observed variables V can be partitioned into j disjoint sets or components, according to whether they share common unobserved parents. These disjoint sets have been referred to as *confounded components* (abbreviated: *c-components*) (Tian and Pearl, 2002) or *districts* (Richardson, 2009). More generally, it is said that any two observed variables sharing a common unobserved parent belong to the same c-component S_j. The importance of c-components can be appreciated by the fact that their disjointness implies that the joint distribution of observed variables $P(V)$ can be factorized as a product of their corresponding *c-factors*.

For example, the joint distribution $P(V_1, V_2, V_3, V_4, V_5)$ in the graph in Figure 2.4, with c-components $S_1 = \{V_1, V_3, V_5\}$ and $S_2 = \{V_2, V_4\}$, factorizes as

$$\begin{aligned}
&\sum_{u_1,u_2,u_3} P(V_1|U_1 = u_1)P(V_2|V_1, U_2 = u_2)P(V_3|V_2, U_1 = u_1, U_3 = u_3) \\
&\qquad \times P(V_4|V_3, U_2 = u_2)P(V_5|V_4, U_3 = u_3) \\
&\qquad \times P(U_1 = u_1, U_2 = u_2, U_3 = u_3) \\
&= \sum_{u_1,u_3} P(V_1|U_1 = u_1)P(V_3|V_2, U_1 = u_1, U_3 = u_3)P(V_5|V_4, U_3 = u_3) \\
&\qquad \times P(U_1 = u_1, U_3 = u_3) \\
&\qquad \times \sum_{u_2} P(V_2|V_1, U_2 = u_2)P(V_4|V_3, U_2 = u_2)P(U_2 = u_2) \\
&= Q[S_1]Q[S_2],
\end{aligned}$$

where $Q[S_1]$ and $Q[S_2]$ are the corresponding c-factors of S_1 and S_2, respec-

tively.

More generally, in the presence of unobserved variables U, the probability distribution $P(V)$ of observed variables V in a semi-Markovian graph, has been shown to factorize as a mixture of products involving observed and unobserved variables:

$$P(V) = \sum_{u} \prod_{i} P(V_i | PA_i, U^i = u^i) P(U = u), \tag{2.15}$$

where PA_i and U^i stand for the observed and unobserved parents of V_i, respectively. Tian and Pearl (2002) pointed out that, because of disjointness of the c-components S_j, expression (2.15) can always be rewritten as a *c-component factorization* as follows

$$\prod_{j} Q[S_j] = \prod_{j} P(S_j | do(PA_{-}(S_j))),$$

where $PA_{-}(S_j) = PA(S_j) \setminus S_j$ denotes the set of observed parents of all nodes in S_j (excluding nodes in S_j itself) such that every c-factor $Q[S_j]$ can be interpreted as the interventional distribution of S_j under an intervention to all its observed parents (excluding those in S_j). Moreover, every $Q[S_j]$ can be expressed in terms of the observed data according to the following lemma.

Lemma 2.5.1. *(Tian and Pearl, 2002) Let a topological order over V be $V_1 < ... < V_n$ (where $V_i < V_j$ denotes that V_i precedes V_j), and let $V^{(i)} = \{V_1, ..., V_i\}$, $i = 1, ..., n$, and $V^{(0)} = \emptyset$. For any observed set C, let $\mathcal{G}_C$ denote the subgraph of $\mathcal{G}$ composed only of variables in C. Then*

(i) Each c-factor $Q[S_j]$, $j = 1, ..., k$, is identifiable and is given by

$$Q[S_j] = \prod_{i | V_i \in S_j} P(V_i | V^{(i-1)}).$$

(ii) Each factor $P(V_i | V^{(i-1)})$ can be expressed as

$$P(V_i | V^{(i-1)}) = P(V_i | PA(T_i) \setminus V_i),$$

where T_i is the c-component of $\mathcal{G}_{V^{(i)}}$ that contains V_i.

For instance, the result for $Q[\{A, Y\}]$ in expression (2.13) arises from applying this lemma in the simple front-door example discussed in section 2.5.1.

2.5.3 A complete identification algorithm

The above insights allowed Tian and Pearl (2003) to develop a complete identification algorithm based on c-component factorization, referred to as Tian's algorithm or the **ID** algorithm, as depicted in Figure 2.5. In particular, they showed that the original problem of identifiability of $P(Y|do(A = a))$ can be reduced to smaller identifiability problems within a subgraph of $\mathcal{G}$ where certain non-essential nodes are systematically removed. More specifically, identifiability of $P(Y|do(A = a))$ depends on identifiability of the c-factors of districts D_i in the subgraph[8] $\mathcal{G}_D$, where D is the set of ancestors of Y (including Y) in the subgraph $\mathcal{G}_{V\setminus A}$. This dependence follows from the fact that $P(Y|do(A = a))$ can always be expressed as a sum over the product of the c-factors $Q[D_i]$:[9]

$$\begin{aligned} P(Y|do(A = a)) &= \sum_{d\setminus y} \prod_i Q[D_i] \\ &= \sum_{d\setminus y} \prod_i P(D_i|do(PA_{-}(D_i) = pa_{-}(D_i))). \end{aligned} \tag{2.16}$$

where each district D_i in the subgraph $\mathcal{G}_D$ (logically) constitutes a subset of a c-component S_j in the original graph $\mathcal{G}$, and, again, $pa_{-}(D_i)$ denotes the vector of value assignments to $PA_{-}(D_i)$ such that, if $A \in PA_{-}(D_i)$, value assignment $PA_{-}(D_i) = pa_{-}(D_i)$ is consistent with $A = a$. Consequently,

[8] Let $\mathcal{G}_C$, in contrast to the notation in Lemma 2.5.1, denote the subgraph of $\mathcal{G}$ composed only of nodes in C and hidden nodes with at least two children in C. In ADMGs, $\mathcal{G}_C$ can be denoted as the subgraph of $\mathcal{G}$ composed only of nodes in C and edges in $\mathcal{G}$ with both endpoints in C. Moreover, to increase clarity, we will henceforth refer to c-components or districts in the original graph $\mathcal{G}$ as *c-components* and in the subgraph $\mathcal{G}_D$ as *districts*.

[9] Note that expression (2.16) can be conceived as a generalization of the truncated factorization formula in expression (2.6) for graphs with latent variables. In the absence of latent variables, each district consists of a single node and expression (2.16) reduces to expression (2.6).

INPUT: two disjoint sets $A, Y \subset V$.
OUTPUT: the expression for $P(Y|do(A = a))$ or FAIL.

Phase 1:

1. Find the c-components of $\mathcal{G}$: $S_1, ..., S_k$. Compute each $Q[S_i]$ by Lemma 2.5.1.
2. Let D denote the ancestors of Y in $\mathcal{G}_{V \setminus A}$ and the c-components of $\mathcal{G}_D$ be $D_i, i = 1, ..., l$.

Phase 2:
For each set D_i such that $D_i \subseteq S_j$:
Compute $Q[D_i]$ from $Q[S_j]$ by calling **Identify**$(D_i, S_j, Q[S_j])$ in Figure 2.6.
If the function returns FAIL, then stop and output FAIL.

Phase 3:
Output $P(Y|do(A = a)) = \sum_{d \setminus y} \prod_i Q[D_i]$.

Figure 2.5: Algorithm **ID**(Y, A) (Tian and Pearl, 2003)

identification of each of the c-factors $Q[D_i]$ depends on whether it can be derived from its corresponding c-factor $Q[S_j]$ in the original graph $\mathcal{G}$, which can be determined by the **Identify** algorithm in Figure 2.6 (Tian and Pearl, 2003).

For instance, in the example in section 2.5.1, $\mathcal{G}_{V \setminus A}$ corresponds to the graph that only includes M and Y, both of which are ancestors of Y (or included in Y), such that $V \setminus A = D = \{M, Y\}$ and hence $\mathcal{G}_{V \setminus A} = G_D$. In addition, $\mathcal{G}_D$ contains two c-components, i.e. $D_1 = \{M\}$ and $D_2 = \{Y\}$, such that Tian's algorithm yields

$$\begin{aligned} P(Y|do(A = a)) &= \sum_m Q[\{Y\}]Q[\{M\}] \\ &= \sum_m P(Y|do(M = m))P(M = m|do(A = a)), \end{aligned} \tag{2.17}$$

an expression involving interventional distributions whose identification result can, in this case, be obtained via the adjustment criterion, leading to expression (2.14). Note that, more generally, $Q[D_2] = Q[\{Y\}]$ could have been computed from the c-factor $Q[\{A, Y\}]$ in the original graph $\mathcal{G}$ by the

INPUT: $C \subseteq T \subseteq V$, $Q = Q[T]$. $\mathcal{G}_T$ and $\mathcal{G}_C$ are both composed of one single c-component.
OUTPUT: the expression for $Q[C]$ in terms of Q or FAIL.

Let A denote the ancestors of C in $\mathcal{G}_T$.

- If $A = C$, output $Q[C] = \sum_{t \setminus c} Q$.
- If $A = T$, output FAIL.
- If $C \subset A \subset T$

  1. Assume that in $\mathcal{G}_A$, C is contained in a c-component T'.
  2. Compute $Q[T']$ from $Q[A] = \sum_{t \setminus a} Q$ by Lemma 2.5.2.
  3. Output **Identify**$(C, T', Q[T'])$.

Figure 2.6: Algorithm **Identify**(C, T, Q) (Tian and Pearl, 2003)

Identify algorithm in Figure 2.6. Since $Q[\{A, Y\}] = P(Y|A, M)P(A)$ by Lemma 2.5.1, $Q[\{Y\}] = P(Y|do(M = m))$ can indeed be shown to equal $\sum_a P(Y|A = a, M)P(A = a)$. That is, because of the exclusion restriction discussed in section 2.5.1, A is not an ancestor of Y in $\mathcal{G}_{\{A,Y\}}$, so that $Q[\{Y\}]$ is identified and can simply be obtained by summing $Q[\{A, Y\}]$ over A.

As opposed to this simple front-door example, in many settings, $\mathcal{G}_D \neq \mathcal{G}_{V \setminus A}$ and, moreover, determining whether each $Q[D_i]$ is computable from a corresponding $Q[S_j]$ is less straightforward as it often leads to recursive applications of the **Identify** algorithm, as illustrated in the next, somewhat more involved example.

2.5.4 A somewhat more involved example

Consider the graph $\mathcal{G}$ in Figure 2.7, which was discussed in Pearl (2014). It is easily shown that $P(Y|do(A = a))$ cannot be identified by covariate adjustment.[10] That is, the non-causal path $A \leftarrow U_3 \rightarrow C_3 \rightarrow Y$ can only be

[10]The steps below can easily be followed using DAGitty, a browser-based environment for creating, editing, and analyzing causal models (Textor et al., 2011). Graph $\mathcal{G}$ in Figure 2.7 can be loaded from this url: `http://dagitty.net/mMdmQxs`

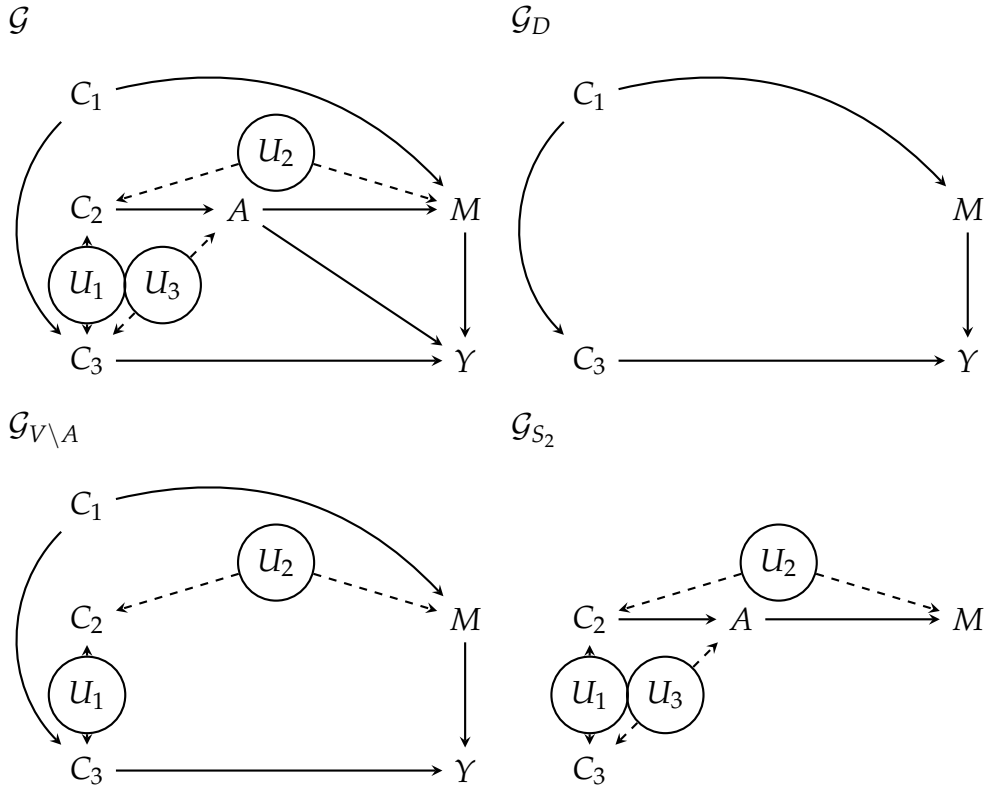


Figure 2.7: A somewhat more involved graph $\mathcal{G}$ and three of its subgraphs required for application of the **ID** algorithm.

blocked by C_3. Adjusting for C_3 also blocks the non-causal path $A \leftarrow C_2 \leftarrow U_1 \rightarrow C_3 \rightarrow Y$. However, since C_3 is a collider, adjusting for it would open a spurious pathway, i.e. $A \leftarrow U_3 \rightarrow \boxed{C_3} \leftarrow C_1 \rightarrow M \rightarrow Y$.[11] This spurious pathway can again be blocked upon adjusting for C_1. This leaves us with one remaining non-causal path, i.e. $A \leftarrow C_2 \leftarrow U_2 \rightarrow M \rightarrow Y$, which can only be blocked by C_2. However, since C_2 is also a collider, adjusting for it opens yet another spurious pathway that passes collider C_3, which is already adjusted for, i.e. $A \leftarrow U_3 \rightarrow \boxed{C_3} \leftarrow U_1 \rightarrow \boxed{C_2} \leftarrow U_2 \rightarrow M \rightarrow Y$. The only way to block this spurious pathway would be to adjust for M. However, this would imply blocking a proper causal path that we are interested in. Indeed, the adjustment criterion dictates that no element in the adjustment set lies on a proper causal path from A to Y.

[11] Adjusted variables are represented by a boxed-in node.

Nevertheless, non-parametric identification of $P(Y|do(A = a))$ can be obtained using the Tian's **ID** algorithm (Figure 2.5).[12] Intuitively, this may be appreciated by the fact that progress can be made by relying on exclusion restrictions encoded in the graph in Figure 2.7. Indeed, below, we illustrate that, just as the exclusion restrictions that A does not affect Y other than through M and that C does not affect M other than through A in the simple front-door example in section 2.5.1 enabled us to make progress, reliance on similar exclusion restrictions may aid in obtaining identification of $P(Y|do(A = a))$ in the graph in Figure 2.7.

The districts in the original graph $\mathcal{G}$ are $S_1 = \{C_1\}$, $S_2 = \{C_2, C_3, A, M\}$ and $S_3 = \{Y\}$. Their corresponding c-factors can be obtained by applying Lemma 2.5.1. Because S_1 and S_3 are singletons, their c-factors have a unique expression, which can easily be obtained by Lemma 2.5.1 as $Q[S_1] = P(C_1)$ and $Q[S_3] = P(Y|A, M, C_3)$, respectively. The second district, S_2, on the other hand, consists of multiple nodes, which, moreover, are subject to multiple possible topological orders. That is, within S_2, there is no order restriction with respect to C_3. In other words, whereas C_2 strictly precedes A, which, in turn, strictly precedes M, i.e. $C_2 < A < M$, the location of C_3 within the topological ordering is unconstrained: it may precede or succeed any of the other nodes in S_2. This observation may be exploited later on when we need to sum out $Q[S_2]$ over certain variables in order to obtain c-factors of some of the districts in the subgraph $\mathcal{G}_D$. In particular, we will need to cleverly choose two specific topological orderings in order to make progress. First, according to the ordering $C_1 < C_3 < C_2 < A < M < Y$, $Q[S_2]$ can be expressed as

$$P(C_3|C_1)P(C_2|C_1, C_3)P(A|C_1, C_2, C_3)P(M|A, C_1, C_2, C_3) \tag{2.18}$$

by Lemma 2.5.1, whereas the ordering $C_1 < C_2 < A < M < C_3 < Y$ enables

[12] Identification results from both the **ID** algorithm and the **IDC** algorithm, discussed in the next section, can be obtained using the R package causaleffect (Tikka, 2016). The added value of this package follows from the fact that applying these algorithms 'by hand' can be tedious, as illustrated in this and the next section.

us to express $Q[S_2]$ as

$$P(C_2)P(A|C_2)P(M|A,C_1,C_2)P(C_3|A,M,C_1,C_2). \tag{2.19}$$

In the subgraph $\mathcal{G}_{V\setminus A}$ (Figure 2.7), C_2 is no longer an ancestor of Y, such that $D = \{C_1, C_3, M, Y\}$. The resulting subgraph $\mathcal{G}_D$ (Figure 2.7) has four districts, i.e. $D_1 = \{C_1\}$, $D_2 = \{C_3\}$, $D_3 = \{M\} \subset S_2$ and $D_4 = \{Y\}$, such that $P(Y|do(A = a))$ can be expressed as

$$\begin{aligned}&\sum_{c_1,c_3,m} Q[\{C_1\}]Q[\{C_3\}]Q[\{M\}]Q[\{Y\}]\\ &= \sum_{c_1,c_3,m} P(C_1 = c_1)P(C_3 = c_3|do(C_1 = c_1))P(M = m|do(A = a, C_1 = c_1))\\ &\qquad \times P(Y|do(A = a, M = m, C_3 = c_3)).\end{aligned} \tag{2.20}$$

Since $D_1 = S_1$ and $D_4 = S_3$, their corresponding c-factors will also be identical, i.e. $Q[D_1] = Q[S_1] = P(C_1)$ and $Q[D_4] = Q[S_3] = P(Y|A, M, C_3)$.

Obtaining the c-factors of D_2 and D_3 – both of which are subsets of S_2 – will, however, be somewhat more involved, as this involves application of the **Identify** algorithm (Figure 2.6). Since C_3 only has itself as an ancestor in the subgraph $\mathcal{G}_{S_2}$ (Figure 2.7), obtaining $Q[D_2]$ is relatively simple, as further instructions are then indicated by the first bullet in Figure 2.6. That is, **Identify**$(D_2, S_2, Q[S_2])$ yields $Q[D_2] = \sum_{c_2,a,m} Q[S_2]$, which, by expression (2.18) reduces to

$$\begin{aligned}&\sum_{c_2,a,m} P(C_3|C_1)P(C_2 = c_2|C_1, C_3)P(A = a|C_1, C_2 = c_2, C_3)\\ &\qquad \times P(M = m|A = a, C_1, C_2 = c_2, C_3) = P(C_3|C_1).\end{aligned} \tag{2.21}$$

Note that the careful choice of letting C_3 precede the other nodes in S_2 in the topological ordering $C_1 < C_3 < C_2 < A < M < Y$, indeed leads to an expression for $Q[S_2]$ which can easily be summed over the other variables in S_2.

Obtaining $Q[D_3]$, on the other hand, is quite tedious, because it involves recursive applications of the **Identify** algorithm. To see why, note that the set of ancestors of M in $\mathcal{G}_{S_2}$ corresponds to $\{C_2, A, M\}$, which, in turn, is a subset of S_2, thus leading us to the third bullet in Figure 2.6. Furthermore,

in the subgraph $\mathcal{G}_{\{C_2,A,M\}}$, M is contained in the district $\{C_2, M\}$. Before we can compute $Q[\{C_2, M\}]$ we first need to obtain $Q[\{C_2, A, M\}] = \sum_{c_3} Q[S_2]$. By expression (2.19), the latter can simply be expressed as

$$\begin{aligned}\sum_{c_3} & P(C_2)P(A|C_2)P(M|A, C_1, C_2)P(C_3 = c_3|A, M, C_1, C_2) \\ & = P(C_2)P(A|C_2)P(M|A, C_1, C_2). \end{aligned} \tag{2.22}$$

However, obtaining $Q[\{C_2, M\}]$ from $Q[\{C_2, A, M\}]$, now requires application of Lemma 2.5.2, which is a more complex variant of Lemma 2.5.1.

Lemma 2.5.2. *(Tian and Pearl, 2003) Let $H \subseteq V$, and assume that H is partitioned into c-components $H_1, ..., H_l$ in the subgraph $\mathcal{G}_H$. Then we have*

(i) $Q[H]$ decomposes as

$$Q[H] = \prod_i Q[H_i].$$

(ii) Each $Q[H_i]$ is computable from $Q[H]$. Let k be the number of variables in H, and let a topological order of the variables in H be $V_{m_1} < ... < V_{m_k}$ in $\mathcal{G}_H$. Let $H^{(i)} = \{V_{m_1}, ..., V_{m_i}\}$ be the set of variables in H ordered before V_{m_i} (including V_{m_i}), $i = 1, ..., k$, and $H^{(0)} = \emptyset$. Then each $Q[H_j]$, $j = 1, ..., l$ is given by

$$Q[H_j] = \prod_{i|V_{m_i} \in H_j} \frac{Q[H^{(i)}]}{Q[H^{(i-1)}]},$$

where each $Q[H^{(i)}]$, $i = 1, ..., k$, is given by

$$Q[H^{(i)}] = \sum_{h \setminus h^{(i)}} Q[H].$$

Applying this lemma, we get $H = \{C_2, A, M\}$ with $C_2 < A < M$ and only admissible topological order. We then get

$$Q[\{C_2, M\}] = \frac{Q[\{C_2, A, M\}]Q[\{C_2\}]}{Q[C_2, A]}$$

$$
\begin{aligned}
&= \frac{Q[\{C_2, A, M\}] \sum_{a,m} Q[\{C_2, A, M\}]}{\sum_m Q[\{C_2, A, M\}]} \\
&= \frac{P(C_2)P(A|C_2)P(M|A, C_1, C_2)P(C_2)}{P(C_2)P(A|C_2)} \\
&= P(C_2)P(M|A, C_1, C_2). \qquad (2.23)
\end{aligned}
$$

As a last step, we still need to obtain $Q[D_3] = Q[\{M\}]$ from $Q[\{C_2, M\}]$ by invoking **Identify**$(M, \{C_2, M\}, Q[\{C_2, M\}])$. Since M does not have any ancestors (except itself) in the subgraph $\mathcal{G}_{\{C_2,M\}}$, we get

$$
Q[\{M\}] = \sum_{c_2} Q[\{C_2, M\}] = \sum_{c_2} P(C_2 = c_2)P(M|A, C_1, C_2 = c_2). \qquad (2.24)
$$

It follows that, since every $Q[D_i]$ is identifiable, $P(Y|do(A = a))$ is also identifiable. Its identification result can be obtained by putting all pieces together and substituting every $Q[D_i]$ in expression (2.20) by its corresponding functional of the observed data. We hence obtain:

$$
\begin{aligned}
P(Y|do(A = a)) = &\sum_{c_1,c_3,m} P(C_1 = c_1)P(C_3 = c_3|C_1 = c_1) \sum_{c_2} P(C_2 = c_2) \\
&\times P(M = m|A = a, C_1 = c_1, C_2 = c_2) \\
&\times P(Y|A = a, M = m, C_3 = c_3) \\
= &\sum_{c_1,c_2,c_3,m} P(Y|A = a, M = m, C_3 = c_3) \\
&\times P(M = m|A = a, C_1 = c_1, C_2 = c_2) \\
&\times P(C_1 = c_1)P(C_2 = c_2)P(C_3 = c_3|C_1 = c_1). \qquad (2.25)
\end{aligned}
$$

2.5.5 Conditional causal effects

An extension of the **ID** algorithm for identifying conditional causal effects in subsets of the population defined by strata of a covariate set C – i.e. causal queries of the form $P(Y|do(A = a), C)$ – was later developed by Shpitser and Pearl (2006a) and is referred to as the **IDC** algorithm, as shown in Figure 2.8. The logic behind this algorithm is to re-express $P(Y|do(A = a), C)$ in terms of unconditional interventional distributions, such that further identification can be obtained using the **ID** algorithm. This

INPUT: disjoint sets $A, Y, Z \subset V$.
OUTPUT: Expression for $P(Y|do(A = a), Z)$ in terms of P or FAIL.

1. If there exists a variable $W \in Z$ such that $Y \perp\!\!\!\perp W|A, Z \setminus \{W\}$ in the subgraph $\mathcal{G}_{\overline{A}\underline{W}}$, return $\mathbf{IDC}(Y, A \cup \{W\}, Z \setminus \{W\})$.

2. Else let $P' = \mathbf{ID}(Y \cup Z, A)$ and return $\frac{P'}{\sum_y P'}$.

Figure 2.8: Algorithm $\mathbf{IDC}(Y, A, Z)$ (Shpitser and Pearl, 2006a)

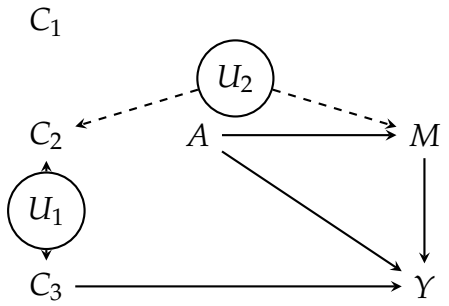
can be achieved by relying on rule 2 of Pearl (1995a)'s *do*-calculus, which allows interventions $do(W = w)$ and observations $W = w$ to be exchanged if the conditional independence $Y \perp\!\!\!\perp W|A, Z \setminus \{W\}$ holds in the subgraph $\mathcal{G}_{\overline{A}\underline{W}}$, obtained by removing from the original graph $\mathcal{G}$ all edges pointing to nodes in A and all edges emanating from nodes in W. The idea is to iteratively apply this rule to find a unique maximal set $W \subseteq C$ that enables expressing $P(Y|do(A = a), C)$ as $P(Y|do(A = a, W), C \setminus W)$. If $W = C$, $P(Y|do(A = a), C)$ then simply equals $P(Y|do(A = a, C))$. However, often one may not get rid of all conditioning variables, that is $W \subset C$. In this case, $P(Y|do(A = a), C)$ equals

$$\frac{P(Y, C \setminus W|do(A = a, W))}{P(C \setminus W|do(A = a, W))} = \frac{P(Y, C \setminus W|do(A = a, W))}{\sum_y P(Y = y, C \setminus W|do(A = a, W))}.$$

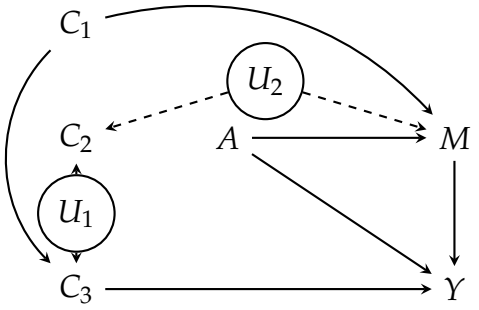
such that its identification ultimately depends on identification of the unconditional joint interventional distribution $P(Y, C \setminus W|do(A = a, W))$, which can be assessed by the **ID** algorithm. Suppose that interest lies in the effect of A on Y conditional on $\{C_1, C_2, C_3\}$, i.e. $P(Y|do(A = a), C_1, C_2, C_3)$, in graph $\mathcal{G}$ in Figure 2.7 (or Figure 2.10).[13] It follows from the subgraphs

[13]It is worth noting here that if, contrary to the fact, $P(Y|do(A = a))$ were identified by the adjustment formula upon adjusting for $\{C_1, C_2, C_3\}$ (refer to section 2.5.4), this would have necessarily implied that $P(Y|do(A = a), C_1, C_2, C_3)$ were likewise identified. This can be seen upon noting that identification of both interventional distributions can be obtained under conditional ignorability $Y \perp\!\!\!\perp A|C_1, C_2, C_3$, which would be implied if $\{C_1, C_2, C_3\}$ would satisfy the adjustment criterion with respect to (A, Y). As opposed to more general identification strategies, such as the **ID** algorithm, identification by the adjustment criterion can thus be conceived as being agnostic as to whether the interventional distribution is

$(Y \perp\!\!\!\perp C_1 | A, C_2, C_3)_{\mathcal{G}_{\overline{A}\underline{C_1}}}$



$(Y \not\perp\!\!\!\perp C_2 | A, C_1, C_3)_{\mathcal{G}_{\overline{A}\underline{C_2}}}$



$(Y \not\perp\!\!\!\perp C_3 | A, C_1, C_2)_{\mathcal{G}_{\overline{A}\underline{C_3}}}$

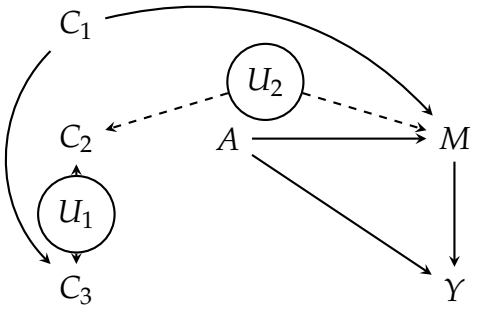


Figure 2.9: Different subgraphs of $\mathcal{G}$ Figure 2.7 that aid in finding a maximal set $W \subseteq C$ through recursive applications of the first step of the **IDC** algorithm.

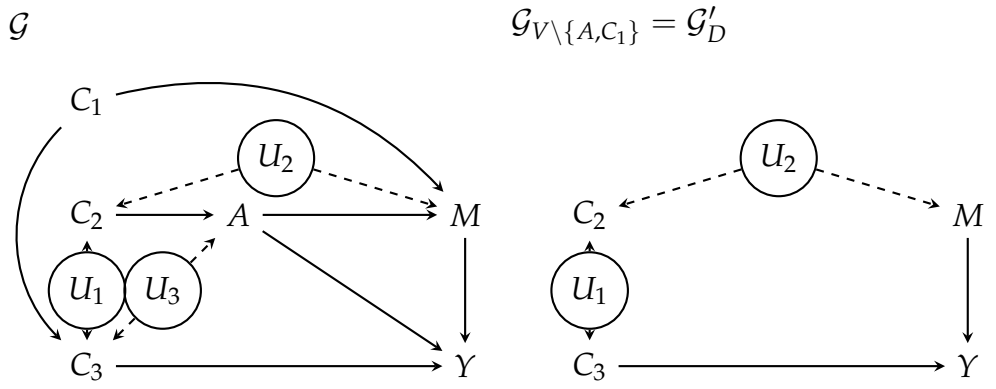


Figure 2.10: A somewhat more involved graph $\mathcal{G}$ and a subgraph required for application of the second step of the **IDC** algorithm.

$\mathcal{G}_{\overline{A}\underline{C_1}}$, $\mathcal{G}_{\overline{A}\underline{C_2}}$ and $\mathcal{G}_{\overline{A}\underline{C_3}}$ in Figure 2.9 that the unique maximal set $W \in C = \{C_1, C_2, C_3\}$ such that $P(Y|do(A = a), C) = P(Y|do(A = a, W), C \setminus W)$ contains only C_1. That is, the first iteration of the **IDC** algorithm (Figure 2.8) first picks $W = C_1$, then reinvokes the algorithm as **IDC**$(Y, \{A, C_1\}, \{C_2, C_3\})$ which assesses whether $Y \perp\!\!\!\perp C_2|A, C_1, C_3$ in the subgraph $\mathcal{G}_{\overline{A,C_1}\underline{C_2}}$ or $Y \perp\!\!\!\perp C_3|A, C_1, C_2$ in the subgraph $\mathcal{G}_{\overline{A,C_1}\underline{C_3}}$. However, note that, since no edges are entering C_1, $\mathcal{G}_{\overline{A,C_1}\underline{C_2}}$ and $\mathcal{G}_{\overline{A,C_1}\underline{C_3}}$ correspond to $\mathcal{G}_{\overline{A}\underline{C_2}}$ and $\mathcal{G}_{\overline{A}\underline{C_3}}$ in Figure 2.9, respectively. Since we already have that the above conditional independencies do not hold in the latter subgraphs, we conclude that C_1 is the unique maximal set such that $P(Y|do(A = a), C) = P(Y|do(A = a, W), C \setminus W)$. Consequently, we have

$$P(Y|do(A = a), C_1, C_2, C_3) = \frac{P(Y, C_2, C_3|do(A = a, C_1))}{\sum_y P(Y = y, C_2, C_3|do(A = a, C_1))},$$

such that identification of $P(Y|do(A = a), C_1, C_2, C_3)$ depends on identification of the unconditional joint interventional distribution $P(Y, C_2, C_3|do(A = a, C_1))$, which can be obtained by the **ID** algorithm as follows.

All variables in the subgraph $\mathcal{G}_{V\setminus\{A,C_1\}}$ (Figure 2.10), are ancestors of $\{Y, C_2, C_3\}$, such that $\mathcal{G}_{V\setminus\{A,C_1\}} = \mathcal{G}_{D'}$.[14] This subgraph $\mathcal{G}_{D'}$ contains two

conditional on (a subset of) covariates in the sufficient adjustment set.

[14] In order to avoid confusion, we will denote the set of ancestors of $\{Y, C_2, C_3\}$ in $\mathcal{G}_{V\setminus\{A,C_1\}}$ by D' and the districts in $\mathcal{G}_{D'}$ by D'_i.

districts, i.e. $D_1' = \{C_2, C_3, M\}$ and $D_2' = \{Y\}$, such that $P(Y, C_2, C_3|do(A = a, C_1 = c_1))$ can be expressed as

$$\sum_{c_2, c_3, m} Q[\{C_2, C_3, M\}]Q[\{Y\}].$$

It was shown in section 2.5.4 that $Q[\{Y\}] = P(Y|A, M, C_3)$, so we only need to obtain $Q[D_1']$ from its corresponding district in $\mathcal{G}$, i.e. $Q[S_2]$. For this purpose, we need to invoke **Identify**$(D_1', S_2, Q[S_2])$. However, because the set of ancestors of D_1' in the subgraph $\mathcal{G}_{S_2}$ (Figure 2.7) coincides with S_2, $Q[D_1']$ is not identifiable. Because identification of $Q[D_1']$ fails, identification of $P(Y, C_2, C_3|do(A = a, C_1))$ also fails, which ultimately leads to the conclusion that the conditional effect $P(Y|do(A = a), C_1, C_2, C_3)$ is not identifiable from observable data.

However, it can easily be shown that, in contrast, e.g. $P(Y|do(A = a), C_1, C_3)$ is identifiable. This can mainly be appreciated upon noting that by avoiding to condition on collider C_2, C_3 may – in addition to C_1 – also be included in the unique maximal subset W such that $P(Y|do(A = a), C \setminus C_2)$ = $P(Y|do(A = a, W), (C \setminus C_2) \setminus W)$. As a result, $P(Y|do(A = a), C_1, C_3)$ equals $P(Y|do(A = a, C_1, C_3))$, which can be shown to be identified via the **ID** algorithm.

Chapter 3

Identifying natural and path-specific effects from observed data

This chapter is an adapted version of a handbook chapter submitted for peer review in M. Drton, S. Lauritzen, M. Maathuis, M. Wainwright (Eds.), *Handbook of Graphical Models.* CRC Press.

In this chapter, we will study non-parametric identification of natural direct and indirect effects, and of path-specific effects (Avin et al., 2005) in general. In particular, we revisit earlier identifying assumptions (Pearl, 2001) in the light of a recently proposed graphical identification criterion for path-specific effects (Shpitser, 2013) that extends previous work on complete conditions for non-parametric identification of total treatment effects (Huang and Valtorta, 2006; Shpitser and Pearl, 2006a,b, 2008a; Tian and Pearl, 2002, 2003) – as discussed in chapter 2 – to allow for effect decomposition. Through various worked-out examples, we aim to provide insight into the use of this graphical criterion, as well as into the nature of the assumptions on which mediation analysis relies. Before concluding this chapter by extending notions of natural direct and indirect effects to more generally defined path-specific effects, we highlight that Shpitser

(2013)'s graphical criterion leads to novel insights that may contribute to a more comprehensive understanding of recent conceptual developments and formulations inspired by the debate about the distinct and controversial nature of both definitions and required assumptions of targeted path-specific effects (Robins and Richardson, 2010).

3.1 Cross-world counterfactuals...

Despite their formal and intuitive appeal, non-parametric identification of natural effects is subtle and a source of much controversy. The reason is that the usual consistency assumptions alone – namely that $M(a)$ and $Y(a)$ equal M and Y, respectively, when $A = a$, and that $Y(a, m)$ equals Y when $A = a$ and $M = m$ – do not suffice to link all counterfactual data to observed data. In particular, nested counterfactual outcomes $Y(a, M(a'))$ are unobservable when $a \neq a'$. Data, whether experimental or observational, thus never carry information about the distribution of these counterfactuals as they imply a union of two incompatible states a and a' that may only seem to coexist 'across multiple worlds'. Mediation analyses based on natural effects are thus bound to rely on assumptions that cannot be empirically verified or guaranteed by the study design. Even randomised cross-over trials, where one would first manipulate A to a' to observe $M(a')$, and then manipulate A to a and M to $M(a')$ to finally observe $Y(a, M(a'))$, would require strong assumptions of no period effect and no carry-over effects at the individual level (Imai et al., 2013; Josephy et al., 2015; Robins and Greenland, 1992).

3.2 ... require cross-world assumptions

To develop intuition into non-parametric identification of natural effects – and, by extension, path-specific effects – we will work through a number of simple, but typical examples.

Consider the basic mediation setting depicted in the causal diagram in Figure 3.1. Identification of natural effects in this setting can be obtained if we recover the distribution of nested counterfactuals $P(Y(a, M(a')) = y)$. This requires summing the joint counterfactual distribution $P(Y(a, m) =$

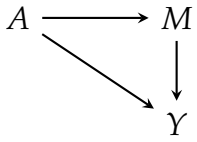


Figure 3.1: A simple, typically over-simplistic, mediation graph.

$y, M(a') = m)$ over m. When $a \neq a'$, the observed data carry no information about the dependence of $Y(a, m)$ and $M(a')$. This articulates why natural effects cannot, in general, be identified from experimental data without further, untestable assumptions.

One such assumption is that of cross-world independence,

$$Y(a, m) \perp\!\!\!\perp M(a'), \tag{i}$$

which Pearl (2001) claimed to be key to 'experimental' identification of natural effects. Under this assumption, we can factorize $P(Y(a, m) = y, M(a') = m)$ as a product of interventional distributions[1] – each of which is identified from observed data under the assumptions encoded in the causal diagram in Figure 3.1 – as follows

$$\begin{aligned} P(Y(a, M(a')) = y) &= \sum_m P(Y(a, m) = y, M(a') = m) \\ &= \sum_m P(Y(a, m) = y) P(M(a') = m) \\ &= \sum_m P(Y = y | A = a, M = m) P(M = m | A = a'). \end{aligned}$$

3.2.1 Non-parametric structural equation models

Cross-world independence (i) is satisfied under the non-parametric structural equation model (NPSEM) associated with the causal diagram in Fig-

[1] In this chapter, we will use counterfactual notation instead of Pearl's *do*-notation (since cross-world counterfactuals simply cannot be expressed using *do*-notation). However, we will refer to counterfactual distributions as interventional distributions, whenever appropriate.

ure 3.1. This model is defined by the following structural equations:

$$\begin{aligned} A &= f_A(\epsilon_A) \\ M &= f_M(A, \epsilon_M) \\ Y &= f_Y(A, M, \epsilon_Y) \end{aligned}$$

where f_A, f_M and f_Y are unknown deterministic functions and ϵ_A, ϵ_M and ϵ_Y are mutually independent random error terms (representing unobserved background variables).

The invariance of these structural equations enables to deduce the counterfactual dependencies that they encode. For example, under the joint intervention $do(A = a, M = m)$ and the single intervention $do(A = a')$, the structural equations can be written respectively as

$$\begin{aligned} A &= a & \qquad A &= a' \\ M(a) &= m & \qquad M(a') &= f_M(a', \epsilon_M) \\ Y(a, m) &= f_Y(a, m, \epsilon_Y) & \qquad Y(a') &= f_Y(a', M(a'), \epsilon_Y) \end{aligned}$$

Under this representation, the only variation in the so-called one-step ahead counterfactuals $V_i(pa_i) = f_{V_i}(pa_i, \epsilon_i)$ – where V_i could be any variable on the causal diagram and pa_i refers to its parents – is due to the mutually independent error terms. It thus follows that all such one-step ahead counterfactuals are also mutually independent, irrespective of the value pa_i to which V_i's parents are set. As a result, independence of the error terms $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ in the above structural equations not only translates into $Y(a, m) \perp\!\!\!\perp M(a)$ but also into cross-world independence (i). This may sound reassuring, but also signals the restrictiveness of the NPSEM (e.g. Robins and Richardson, 2010).

3.2.2 Unmeasured mediator-outcome confounding

However, the assumption of independent error terms $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ encoded in the NPSEM representation of Figure 3.1 requires all common causes of mediator and outcome to be represented on the graph. It is therefore likely violated, even if treatment is randomised. In this section, we will therefore

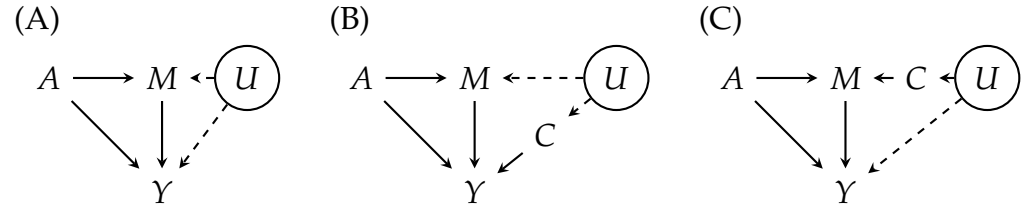


Figure 3.2: A more realistic mediation graph with unmeasured mediator-outcome confounding (A) along with two scenarios where a covariate set C may deconfound the mediator-outcome relation (B,C).

relax assumptions by adding a hidden node U, which induces unmeasured confounding of the mediator-outcome relation, as in Figure 3.2A.

The NPSEM associated with this causal diagram can thus be considered a semi-Markovian NPSEM, a broader class of graphical models which are often represented by acyclic directed mixed graphs (ADMGs) that employ bi-directed edges to indicate (potential) unmeasured confounding (Richardson, 2003). However, as in chapter 2, for the purpose of our presentation, we will explicitly represent hidden variables U by circled nodes and their direct effects on observed variables V by dashed edges.

Not surprisingly, by treatment randomization, $P(Y(a) = y)$ is still identified by $P(Y = y|A = a)$ under Figure 3.2A. Formally, since $U \perp\!\!\!\perp A$, the truncated factorization formula yields

$$\begin{aligned} P(Y(a) = y) &= \sum_{u,m} P(Y = y|A = a, M = m, U = u) \\ &\qquad \times P(M = m|A = a, U = u)P(U = u), \qquad (3.1) \\ &= \sum_{m} P(Y = y|A = a, M = m)P(M = m|A = a) \qquad (3.2) \\ &= P(Y = y|A = a). \end{aligned}$$

Unfortunately, a similar marginalisation over the distribution of U does not permit to identify $P(Y(a, M(a')) = y)$, even under conditional cross-world independence $Y(a, m) \perp\!\!\!\perp M(a')|U$. Indeed, we obtain

$$\begin{aligned} P(Y(a, M(a')) = y) = \sum_{u,m} &P(Y(a, m) = y|U = u) \\ &\times P(M(a') = m|U = u)P(U = u) \end{aligned}$$

$$= \sum_{u,m} P(Y = y|A = a, M = m, U = u) \times P(M = m|A = a', U = u)P(U = u), \tag{3.3}$$

an expression that closely resembles expression (3.1), but that cannot be further simplified as a functional of observed variables (such as expression (3.2)) because of the conflicting treatment assignments in its first two factors.

3.2.3 Identification by the mediation formula

Issues of non-identifiability of $P(Y(a, M(a')) = y)$ can, however, be remedied when one has available a measured set of prognostic covariates C for mediator and/or outcome that renders the mediator-outcome relationship unconfounded given treatment assignment. This is because the existence of such a set C, as, for instance, in the causal diagrams of Figures 3.2B and 3.2C, no longer necessitates stratifying on U to establish cross-world independence.

For example, in Figure 3.2B, conditioning on C suffices, since

$$\begin{aligned} Y(a, m) &= f_Y(a, m, C, \epsilon_Y) \\ M(a') &= f_M(a', U, \epsilon_M), \end{aligned}$$

such that we obtain cross-world independence within strata of C, i.e.

$$Y(a, m) \perp\!\!\!\perp M(a')|C. \tag{ii}$$

This then implies the same functional as expression (3.3) but with unobserved U replaced by the observed set C

$$P(Y(a, M(a')) = y) = \sum_{c,m} P(Y(a, m) = y|C = c) \times P(M(a') = m|C = c)P(C = c) \tag{3.4}$$

$$= \sum_{c,m} P(Y = y|A = a, M = m, C = c) \times P(M = m|A = a', C = c)P(C = c). \tag{3.5}$$

This functional is commonly referred to as Pearl (2001)'s *mediation formula*.

To appreciate the importance of adjustment for prognostic factors C, reconsider the Job Search Intervention Study (JOBS II) (Vinokur et al., 1995), an often cited empirical mediation example that was introduced in section 1.1.1 of chapter 1. Recall that the JOBS II field experiment was designed to assess the effectiveness of a theory-driven job training intervention that aimed to both increase reemployment and reduce depressive symptoms in unemployed workers. One mediation question of interest was whether workshop participation led to reduction in depressive symptoms (at two months follow-up) by increasing chances of getting reemployed (at two months follow-up). Randomization of the intervention in itself did not suffice to eliminate potential confounding between re-employment M and the outcome. It is therefore essential to adjust for pretreatment level of depression, a strong prognostic factor of the outcome of interest and most likely also related to re-employment. Measurements on a range of other baseline covariates, including demographics, previous occupation and financial strain, were also collected and adjusted for in order to strengthen the validity of cross-world assumption (ii).

3.2.4 Treatment-induced mediator-outcome confounding

The previous example may have led the reader to erroneously conclude that, given treatment randomization, adjustment for a measured covariate set C that deconfounds the mediator-outcome relation within treatment arms, is all that is required to establish cross-world independence (ii) under NPSEMs, thus enabling identification of $P(Y(a, M(a')) = y)$.

An important additional requirement, however, is that no element of C is affected by treatment. Intuitively, such adjustment would block the pathway from A to M via C, which makes up part of the natural indirect effect of interest. More importantly, and more formally, a lack of identification can be understood as follows.

According to the NPSEM associated with the causal diagram in Figure 3.3A, $Y(a, l, m)$, $M(a', l')$ and $\{L(a), L(a')\}$ are mutually independent, such that, by the generalized consistency assumption (Pearl, 2000) – which

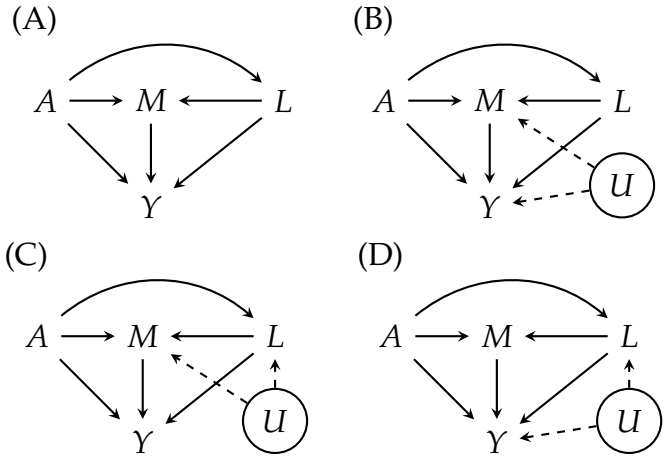


Figure 3.3: Problematic mediation graphs with treatment-induced confounding.

states that for each a and a', $Y(a, L(a), m) = Y(a, m)$ and $M(a', L(a')) = M(a')$ with probability 1 – $Y(a, m) \perp\!\!\!\perp M(a')$ holds conditional on $\{L(a) = l, L(a') = l'\}$. This can also be seen upon noting that

$$Y(a, m) = f_Y(a, m, L(a), \epsilon_Y)$$
$$M(a') = f_M(a', L(a'), \epsilon_M),$$

which allows to express $P(Y(a, M(a')) = y)$ as

$$\begin{aligned}
&\sum_{l,l',m} P(Y(a,m) = y | L(a) = l, L(a') = l') P(M(a') = m | L(a) = l, L(a') = l') \\
&\qquad \times P(L(a) = l, L(a') = l') \\
&= \sum_{l,l',m} P(Y = y | A = a, L = l, M = m) P(M = m | A = a', L = l') \\
&\qquad \times P(L(a) = l, L(a') = l').
\end{aligned}$$

However, this expression cannot be further reduced to a functional of the observed data as it requires the joint counterfactual distribution $P(L(a) = l, L(a') = l')$. Since this distribution again involves conflicting treatment assignments, strong untestable restrictions (beyond those encoded in the NPSEM representation of Figure 3.3A) would be needed to enable identification.

In the JOBS II study, all available covariates were measured prior to

randomization. It may thus be safely assumed that none of them was affected by the intervention. However, as will be discussed later, other mediators of the intervention's effect on mental health, such as an altered sense of self-efficacy, may well have affected re-employment and thus manifest themselves as mediator-outcome confounders that are affected by the intervention. In that case, cross-world independence (ii) would be violated.

3.2.5 Pearl's graphical criteria for cross-world independence

Pearl (2001) devised two graphical criteria for assessing cross-world independence (ii) under a NPSEM, the logic for which can be understood from the previous two examples in sections 3.2.2 and 3.2.4.

The first criterion requires the availability of an adjustment set C that is sufficient, along with treatment A, to adjust for confounding of the association between mediator and outcome. Such covariate set C should block all back-door paths between mediator and outcome (except those traversing A) in the sense that

$$(Y \perp\!\!\!\perp M|C)_{\mathcal{G}_{\underline{AM}}} \tag{iii}$$

that is, C d-separates Y from M in $\mathcal{G}_{\underline{AM}}$, the subgraph formed from the original graph $\mathcal{G}$ by deleting all arrows emanating from A and M.

The second criterion requires that

$$\text{no element of } C \text{ is affected by treatment.} \tag{iv}$$

We will henceforth refer to this criterion as 'no treatment-induced confounding' or 'no intermediate confounding'.

3.3 Avoiding recantation...

The key problem in the examples in sections 3.2.2 and 3.2.4, which respectively violate (iii) and (iv), is that the occurrence of conflicting treatment assignments in certain factors prevents further identification. It can be shown that this problem arises whenever the conflict is situated in the re-

spective (interventional) distributions of variables belonging to the same *confounded component* (abbreviated: *c-component*) (Tian and Pearl, 2002) or *district* (Richardson, 2009). As discussed in more detail in section 2.5.2 of chapter 2, any two observed variables that share a common unobserved parent belong to the same c-component or district.

For instance, in Figure 3.2A, identification of $P(Y(a, M(a')) = y)$ (see expression (3.3)) requires evaluating M when A is set to a' and Y when A is set to a. However, since M and Y share a common unmeasured cause U and thus belong the same district $\{M, Y\}$, the truncated factorization formula yields univariate factors for M and Y that both require conditioning on U. However, since U is unobserved, we cannot generally marginalise over it in expression (3.3), unless $a = a'$, as in expression (3.1). Likewise, $P(Y(a, M(a')) = y)$ is not identifiable in Figure 3.3A because the conflict arises within a single node L, which is itself a district.

In the next sections we provide a more formal treatment of the perhaps rather intuitive notion of such conflicting treatment assignments.

3.3.1 From recanting witnesses...

In the terminology of Avin et al. (2005), L in Figure 3.3A would be called a *recanting witness* for the following reason. For identification of the natural indirect effect it would need to retract an earlier statement, which allows treatment to transmit its entire effect on the mediator in order *not to block* the path from A to M via L, in favour of a new statement that keeps treatment from transmitting its effect on the outcome other than through the mediator, so as to *block* the path from A to Y via L. The recanting witness criterion (Avin et al., 2005) formalizes this requirement of having no such witnesses on the paths of interest in order to identify the targeted path-specific effect – such as the natural direct or indirect effect – transmitted along those paths.

3.3.2 ... to recanting districts

Shpitser (2013) recently developed a complete graphical criterion for identifying path-specific effects under NPSEMs by generalizing the recanting witness criterion to also account for settings with sequential treatments and

unmeasured confounding. Informally, this criterion – which is formalized by theorem 3.4.1 below – requires there to be no 'conflict of interest' between members of a common district within a particular subgraph of $\mathcal{G}$. Districts in which such conflicts exist, are said to be *recanting* with respect to a path-specific effect of interest, as formally defined as follows.[2]

Definition 3.3.1. *Recanting district (Shpitser, 2013). Let $\mathcal{G}$ be an ADMG, V the set of observed nodes in $\mathcal{G}$, and π a subset of directed paths from $A \in V$ to $Y \in V$ in $\mathcal{G}$. Let D be the set of ancestors of Y (including Y) in the subgraph[3] $\mathcal{G}_{V \setminus A}$. Then a district D_i in the subgraph $\mathcal{G}_D$ is called a recanting district for the π-specific effect of A on Y – i.e. the path-specific effect along all paths in π – if there exist nodes $Z_j, Z_k \in D_i$ (possibly $Z_j = Z_k$) such that there is a directed path $A \to Z_j \to ... \to Y$ in π, and a directed path $A \to Z_k \to ... \to Y$ not in π.*

Although this definition applies to generally defined path-specific effects which consist of arbitrary bundles of causal or directed paths π, for now, we will solely focus on natural effects with respect to a single (possibly vector-valued) mediator of interest M.

3.3.3 Some examples

The natural direct effect in the causal diagram $\mathcal{G}$ in Figure 3.2A consists only of the directed path $A \to Y$, such that $\pi = \{A \to Y\}$. The set of observed variables in $\mathcal{G}$ corresponds to $V = \{A, M, Y\}$. The set of ancestors of Y in $\mathcal{G}_{V \setminus A}$ (including Y) corresponds to $D = \{M, Y\}$. Note that D, in this case, is itself a district, because of unoberved mediator-outcome confounding by U. Moreover, it is recanting, since there exist nodes $M \in D$ and $Y \in D$, such that the directed path $A \to Y$ is in π and the directed path $A \to M \to Y$ is not in π. It can easily be shown that, by symmetry, D is also a recanting

[2] In Shpitser (2013)'s original paper, a more general definition was provided in the sense that A and Y could be sets of nodes in $\mathcal{G}$. For the purpose of this chapter, and for ease of exposition, we provide a slightly simplified version of this definition which is restricted to singletons A and Y.

[3] Let $\mathcal{G}_C$ be a subgraph of $\mathcal{G}$ composed only of nodes in C and edges in $\mathcal{G}$ with both endpoints in C. Similarly, if $\mathcal{G}$ corresponds to a DAG with hidden nodes, such as displayed throughout this chapter, then $\mathcal{G}_C$ corresponds to a subgraph of $\mathcal{G}$ composed only of nodes in C, hidden nodes with at least two children in C and edges in $\mathcal{G}$ with both endpoints in C.

district for the natural indirect effect, which consists of the directed path $A \rightarrow M \rightarrow Y$. In other words, since natural direct and indirect effects are each other's complement – i.e. particular instances of each effect combine to produce the total treatment effect – D is recanting with respect to both the natural direct and the natural indirect effect.

Similarly, if π is chosen to represent the natural indirect effect in the causal diagram $\mathcal{G}$ in Figure 3.3A, then $\pi = \{A \rightarrow M \rightarrow Y, A \rightarrow L \rightarrow M \rightarrow Y\}$. The set of observed variables in $\mathcal{G}$ now corresponds to $V = \{A, L, M, Y\}$. The set of ancestors of Y in $\mathcal{G}_{V \setminus A}$ (including Y) corresponds to $D = \{L, M, Y\}$, with districts $D_1 = \{L\}$, $D_2 = \{M\}$ and $D_3 = \{Y\}$. As already intuitively motivated, it is easily shown that D_1 is a recanting district with respect to π, since there exists a node $L \in D_1$, such that the directed path $A \rightarrow L \rightarrow M \rightarrow Y$ is in π and the directed path $A \rightarrow L \rightarrow Y$ is not in π. Again, by symmetry, D_1 is also a recanting district with respect to the natural direct effect in $\mathcal{G}$.

3.4 ...yields interventional identification

Having provided a formal definition of recanting districts, we are now ready to discuss a graphical criterion that enables transporting cross-world quantities – used to define path-specific effects – into a strictly interventional framework under NPSEMs.

3.4.1 The recanting district criterion

Theorem 3.4.1. *Recanting district criterion (Shpitser, 2013). Let $\mathcal{G}$ be an ADMG, V the set of observed nodes in $\mathcal{G}$, and π a subset of directed paths from $A \in V$ to $Y \in V$ in $\mathcal{G}$. Then the π-specific effect of A on Y is expressible as a functional of interventional distributions if and only if there does not exist a recanting district for this effect.*

From this perspective, the need for an observed covariate set C, that is sufficient to adjust for confounding of the mediator-outcome relation (within strata of A), serves to establish that mediator and outcome belong to separate districts so that no conflict arises (provided that no member

of C is affected by treatment). For instance, in Figure 3.2A, a sufficient adjustment set C enables to pull apart the district $\{M, Y\}$ and resolve the conflict in order to ensure the validity of cross-world assumption (ii) that permits factorizing $P(Y(a,m) = y, M(a') = m|C = c)$ as $P(Y(a,m) = y|C = c)P(M(a') = m|C = c)$.

Importantly, the central notion of recantation thus groups Pearl (2001)'s graphical criteria (iii) and (iv) for establishing cross-world independence (ii) under NPSEMs by offering a framework that allows their respective violations to be interpreted as distinct instances of essentially the same problem. As will be discussed in section 3.2.5, the implications of this graphical criterion reach beyond those provided earlier by Pearl (2001).

3.4.2 Interventional identification 1.0

Since cross-world independence (ii) thus enables expressing the cross-world counterfactual distribution $P(Y(a, M(a')) = y)$ in terms of interventional distributions, as in expression (3.4), Pearl (2001) complemented (ii) with a second and third condition for identification. In particular, $P(Y(a, M(a')) = y)$ can be estimated from observed data if, in addition to (ii),

$$P(M(a') = m|C = c) \text{ is identifiable by some means, and} \tag{v}$$
$$P(Y(a,m) = y|C = c) \text{ is identifiable by some means.} \tag{vi}$$

In accordance with Pearl (2014), we explicitly add 'identifiable by some means', since these last two conditions have often been interpreted too strictly in the literature in terms of identifiability by means of adjustment for C. Specifically, (v) has typically been replaced by

$$M(a') \perp\!\!\!\perp A|C \tag{v'}$$

and (vi) by requiring that

$$Y(a,m) \perp\!\!\!\perp A|C, \tag{vi'}$$
$$Y(a,m) \perp\!\!\!\perp M|A = a, C \tag{vi''}$$

both hold, implying once more Pearl's well-known mediation formula (expression (3.5)).

The adjustment criterion for natural effects

A complete graphical criterion for identification of $P(Y(a, M(a')) = y)$ under NPSEMs by the mediation formula was developed by Shpitser and VanderWeele (2011). They termed expression (3.5) the *adjustment formula for natural direct and indirect effects* in order to emphasize the restrictiveness of the identification strategy. This criterion generalizes the adjustment criterion (Shpitser et al., 2010), a complete graphical criterion for identification of total treatment effects $P(Y(a) = y)$ by the adjustment formula $\sum_c P(Y = y|A = a, C = c)P(C = c)$, as discussed in section 2.4.2 of chapter 2.

Specifically, in order for $P(Y(a, M(a')) = y)$ to be identified by expression (3.5) under NPSEMs, this generalized adjustment criterion demands that both $P(M(a) = m)$ and $P(Y(a, m) = y)$ are identifiable by means of adjustment for a common set of measured baseline confounders C. That is, $P(M(a) = m)$ is identified by $\sum_c P(M = m|A = a, C = c)P(C = c)$ and $P(Y(a, m) = y)$ by $\sum_c P(Y = y|A = a, M = m, C = c)P(C = c)$, implying that $P(M(a) = m|C = c)$ and $P(Y(a, m) = y|C = c)$ in conditions (iii) and (iv) are readily identified as $P(M = m|A = a, C = c)$ and $P(Y = y|A = a, M = m, C = c)$, respectively, without needing additional auxiliary covariates for identification.

Intuitively, the adjustment criterion for natural effects can be thought of aiming to establish both cross-world independence (ii) and conditions (v) and (vi) solely by means of adjustment for a common measured covariate set C. First, it demands no unmeasured mediator-outcome confounding, as in Figure 3.2A, which would violate cross-world independence (ii) and, moreover, hamper identification of $P(Y(a, m) = y)$ by means of covariate adjustment. Second, it demands the absence of treatment-induced mediator-outcome confounders, such as L in Figure 3.3A, since the presence of such intermediate confounders would both violate cross-world independence (ii) and hinder the availability of a common set C that enables identification

of both $P(M(a) = m)$ and $P(Y(a, m) = y)$ by means of adjustment for C.[4] Establishing cross-world independence (ii) and conditions (iii) and (iv) by means of the generalized adjustment criterion thus appear to go hand in hand.

Semi-parametric estimators

In section 2.4.3 of chapter 2, we illustrated that rewriting the adjustment formula for treatment effects (in point treatment studies) leads to two semi-parametric estimators. Similar estimators for natural direct and indirect effects have been developed based on the generalized adjustment formula for mediation analyis (or Pearl's mediation formula) (expression (3.5)).

Ratio-of-mediator-probability-weighting estimator One such estimator (Hong, 2010; Lange et al., 2012) requires a working model for the mediator distribution $P(M|A, C)$ in order to calculate weights that are based on the ratio of mediator probabilities (under different treatment assignments). It can be seen to arise by rewriting expression (3.5) as follows

$$
\begin{aligned}
&E\{Y(a, M(a'))\} \\
&= \sum_{c,m} E(Y|A = a, M = m, C = c)P(M = m|A = a', C = c)P(C = c) \\
&= \sum_{y,c,m} y \cdot P(Y = y|A = a, M = m, C = c) \\
&\qquad \times P(M = m|A = a', C = c)\frac{P(C = c, A = a)}{P(A = a|C = c)} \\
&= \sum_{y,c,m} y \cdot P(Y = y, M = m|A = a, C = c) \\
&\qquad \times \frac{P(M = m|A = a', C = c)}{P(M = m|A = a, C = c)} \frac{P(C = c|A = a)P(A = a)}{P(A = a|C = c)} \\
&= \sum_{y,c,m} y \cdot P(Y = y, M = m, C = c|A = a)\frac{P(A = a)}{P(A = a|C = c)}
\end{aligned}
$$

[4] This can be seen upon noting that identification of $P(M(a) = m)$ by covariate adjustment insists L not to be included in C since doing so would amount to adjusting away part of the effect of interest. On the other hand, even though identification of $P(Y(a, m) = y)$ cannot be obtained by the adjustment criterion, a more general identifying functional for $P(Y(a, m) = y)$ can be shown to require some form of adjustment for L.

$$\times \frac{P(M=m|A=a',C=c)}{P(M=m|A=a,C=c)}$$
$$= E\left[\frac{YI(A=a)}{P(A=a|C)}\frac{P(M|A=a',C)}{P(M|A=a,C)}\right].$$

Just as for the inverse probability weighted estimator, discussed in section 2.4.3, one needs to (additionally) weight by the inverse of the probability of being assigned to treatment arm a, to account for the possibly selective nature of subjects with treatment assignment $A = a$. One thus additionally needs to fit a propensity score model $P(A|C)$.

Imputation estimator Another estimator (Albert, 2012; Vansteelandt et al., 2012b) requires an imputation model for the mean outcome given treatment, mediator and covariate set C, to impute counterfactual outcomes under a (possibly) counterfactual treatment assignment $A = a$. It can be seen to arise by rewriting expression (3.5) as follows

$$\begin{aligned}
&E\{Y(a,M(a'))\}\\
&= \sum_{c,m} E(Y|A=a,M=m,C=c)P(M=m|A=a',C=c)P(C=c)\\
&= \sum_{c,m} E(Y|A=a,M=m,C=c)\frac{P(M=m,C=c,A=a')}{P(A=a'|C=c)}\\
&= \sum_{c,m} E(Y|A=a,M=m,C=c)\frac{P(M=m,C=c|A=a')P(A=a')}{P(A=a'|C=c)}\\
&= E\left[\frac{I(A=a')}{P(A=a'|C)}E(Y|A=a,M,C)\right].
\end{aligned}$$

Similarly, the resulting estimator additionally requires to weight by the inverse of the probability of being assigned to treatment level a', to account for the possibly selective nature of subjects with treatment assignment $A = a'$ and hence also requires fitting a propensity score model $P(A|C)$.[5]

Implementation of these estimators (as well as stratum-specific analogs)

[5]Technically, such a propensity score model could be avoided by a two-stage imputation approach. This can be seen upon noting that expression (3.5) can also be rewritten as $E\{E[E(Y|A=a,M,C)|A=a',C]|A=a'\}$.

will be discussed in more detail in chapter 4. The unbiasedness of these estimators logically depends on whether the used covariate set C satisfies the adjustment criterion relative to both (A, M) and $(\{A, M\}, Y)$.

Increased identification power in observational studies

At least from a theoretical point of view, it can be argued that the adjustment criterion seriously limits the ability to identify $P(Y(a, M(a')) = y)$. Indeed, as recently indicated by Pearl (2014), the interventional distributions in (v) and (vi) can be identified under a much wider range of research settings by Shpitser and Pearl (2006a)'s **IDC** algorithm for conditional treatment effects, which may involve

(a) additional adjustment for separate (but possibly overlapping) covariate sets or

(b) *mediating instruments*[6] that enable application of the front-door estimator (as discussed in section 2.5.1 in chapter 2).

Pearl (2014) referred to the first identification strategy (a) as *piecemeal deconfounding*, because it can be regarded as a compromise between identification by the adjustment criterion (which requires identification by adjustment for a common set of covariates) and more general identification strategies such as (b). More specifically, this 'divide and conquer' strategy requires finding a set C that both satisfies (ii) and enables identification of $P(M(a') = m|C = c)$ and $P(Y(a, m) = y|C = c)$ by means of adjusting for sufficient adjustment sets C_m and C_y, respectively.[7]

The increased identification power of these additional strategies is, however, only relevant for observational studies (e.g. Imai et al., 2014), as both (v') and (vi') are satisfied by design if treatment is randomized, whereas (vi'') follows from combining either of these two assumptions

[6] This terminology was used in Pearl (2014) to refer to strong intermediate variables that fully mediate certain effects whose identification can thus be obtained by the front-door criterion.

[7] This thus implies that $P(M(a') = m|C = c)$ is identified by $\sum_{c_m} P(M = m|A = a', C = c, C_m = c_m)P(C_m = c_m)$ and $P(Y(a, m) = y|C = c)$ by $\sum_{c_y} P(Y = y|A = a, M = m, C = c, C_y = c_y)P(C_y = c_y)$.

with (ii). This indeed implies that, in experimental studies in which treatment is randomized, only cross-world independence (ii) is required.

3.4.3 Interventional identification 2.0

We now show how the insights provided by the central notion of recantation can be incorporated within recent work on complete identification algorithms for interventional distributions (Huang and Valtorta, 2006; Shpitser and Pearl, 2006a,b, 2008a; Tian and Pearl, 2003) so as to arrive at identification strategies with complete identification power. Specifically, Shpitser (2013) recently proved that any π-specific effect of A on Y is identifiable under NPSEMs if, and only if the following two conditions hold:

there is no recanting district for the π-specific effect of A on Y (vii)

$P(Y(a) = y)$ is identifiable by some means. (viii)

Since π may refer to a subset of paths that constitute either a natural direct or a natural indirect effect, this result also provides a complete identification criterion for $P(Y(a, M(a')) = y)$.

Below we demonstrate that this result opens avenues towards novel strategies for identifying natural effects, mainly by resorting to alternative cross-world assumptions that may substitute for cross-world independence (ii). Before doing this, we give a more detailed review of Shpitser (2013)'s main results, followed by some examples.

From cross-world to interventional quantities

In the beginning of this section, we already mentioned that the recanting district criterion enables translating cross-world quantities – used to define path-specific effects – into interventional quantities. It does so by demanding there to be no recanting district for the π-specific effect of interest, such that conflicting treatment assignments only occur in different districts. Specifically, if condition (vii) holds, by Theorem 3.4.1, π-specific effects on some outcome Y can be expressed as a functional of interventional

distributions[8]

$$\sum_{d \setminus y} \prod_{i} P(D_i | do(PA\_(D_i) = pa^i\_(D_i))), \tag{3.6}$$

where $PA\_(D_i) = PA(D_i) \setminus D_i$ denotes the set of parents of all nodes in district D_i (excluding nodes in D_i itself). Here, the product runs across all districts D_i in the subgraph $\mathcal{G}_D$, and the summation is made over all possible realisations of the nodes in these districts, except for the outcome. Note that this expression closely matches Tian and Pearl (2003)'s identifying functional for interventional distributions $P(Y(a) = y)$ (expression (2.16) in chapter 2), which builds on c-component factorization and can be considered a generalization of the truncated factorization formula to DAGs with hidden variables. In fact, expression (3.6) differs from expression (2.16) only to the extent that treatment assignments are allowed to be different across districts. This is indicated by the superscript i in $pa^i\_(D_i)$, which denotes the vector of value assignments to $PA\_(D_i)$. For instance, $pa^i\_(D_i)$ includes the treatment assignment a or a' depending on whether D_i includes children of treatment A that transmit part of the natural direct or indirect effect, respectively. More generally, if interest lies in identifying a certain π-specific effect, then $pa^i\_(D_i)$ includes e.g. the active treatment assignment a (or baseline assignment a') depending on whether (or not) D_i includes children of treatment A that transmit part of that π-specific effect, respectively.

From interventional to observed quantities

Further identification of π-specific effects in terms of observational distributions thus depends on whether each factor $P(D_i | do(PA\_(D_i) = pa^i\_(D_i)))$ in expression (3.6) is identifiable from the observed data. However, because expression (3.6) only differs from Tian and Pearl (2003)'s identifying functional in that the former allows for conflicting treatment assignments between districts, identifiability of $P(Y(a) = y)$ by Tian's **ID** algorithm logically implies each factor in expression (3.6) to be expressible in terms of

[8]For notational simplicity, we choose to display expression (3.6) in terms of interventional rather than counterfactual notation.

observational distributions. This is made explicit in condition (viii).

Some examples

To illustrate the above results, we give a number of examples in which the recanting district criterion mainly serves to establish cross-world independence (ii); natural effects are therefore also identified under Pearl's identifying conditions, as discussed in section 3.4.2. In the next section, we develop a novel complementary identification strategy that is inspired by the logic of the recanting district criterion and circumvents cross-world independence (ii).

A simple example In the subgraph $\mathcal{G}_D$ of causal diagram $\mathcal{G}$ in Figure 3.2B, there are two districts in $D = \{C, M, Y\}$, i.e. $D_1 = \{M, C\}$ and $D_2 = \{Y\}$, such that, by Theorem 3.4.1, $P(Y(a, M(a')) = y)$ is expressible as

$$\sum_{c,m} P(Y = y|do(A = a, M = m, C = c))P(M = m, C = c|do(A = a')).$$

Since we know that $P(Y(a) = y)$ is identifiable, each of the factors in this expression must also be identifiable. Indeed, by Tian's **ID** algorithm, we find that this expression can be written as a functional of the observed data, i.e. expression (3.5).

If C were unmeasured, such as in Figure 3.2A, the identifying functional for $P(Y(a) = y)$ would be

$$\sum_{m} P(Y = y, M = m|do(A = a)),$$

an expression that necessarily groups together M and Y into a joint interventional distribution, because they belong to a common district $\{M, Y\}$. This grouping into joint distributions once more illustrates the obstacle to enable conflicting treatment assignments for nodes in a common district, as was also exemplified by the inability to marginalise over U in expression (3.3).

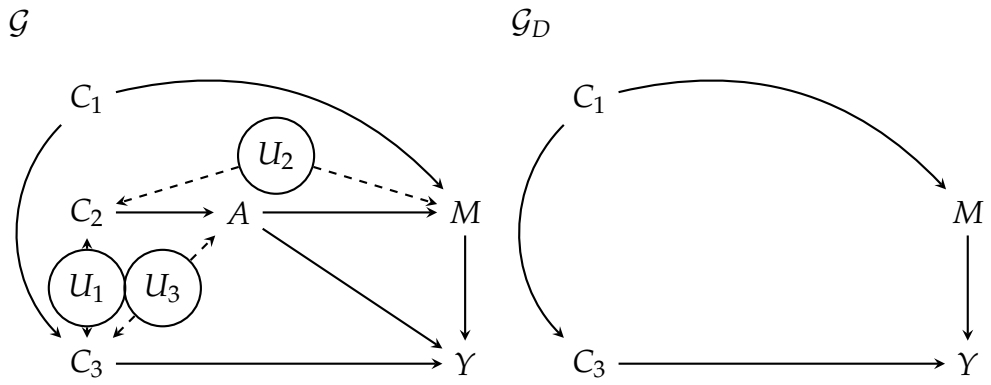


Figure 3.4: The somewhat more involved graph $\mathcal{G}$ from chapter 2 along with its subgraph $\mathcal{G}_D$.

A somewhat more involved example Consider again the causal diagram $\mathcal{G}$ from Figure 2.7, which corresponds to Figure 5F in Pearl (2014) and is reproduced here in Figure 3.4, for convenience, along with the subgraph of interest $\mathcal{G}_D$.

In section 2.5.4 of chapter 2, we had demonstrated that $P(Y(a) = y)$ is not identifiable in $\mathcal{G}$ via the adjustment criterion. Similarly, no common set of baseline covariates C can be found such that the adjustment criterion is satisfied relative to both (A, M) and $(\{A, M\}, Y)$. That is, any subset of $\{C_1, C_2, C_3\}$ that includes C_2 but not C_3 can be shown to satisfy the adjustment criterion relative to (A, M). This can be seen by noting that, in any case, we must adjust for C_2. Since C_3 is a collider, adjusting for it opens spurious pathways $A \leftarrow U_3 \rightarrow \boxed{C_3} \leftarrow C_1 \rightarrow M$ and $A \leftarrow U_3 \rightarrow \boxed{C_3} \leftarrow U_1 \rightarrow \boxed{C_2} \leftarrow U_2 \rightarrow M$ that cannot be blocked by additionally adjusting for C_1. However, the adjustment criterion relative to $(\{A, M\}, Y)$ insists that C_3 be included in the adjustment set, because the spurious path $A \leftarrow U_3 \rightarrow C_3 \rightarrow Y$ can only be blocked by C_3. The adjustment criterion for natural effects (Shpitser and VanderWeele, 2011) thus tells us that, logically, since $P(Y(a) = y)$ is not identifiable via the adjustment criterion, neither is $P(Y(a, M(a')) = y)$.

Nonetheless, since there is no recanting district for the natural direct or indirect effect in $\mathcal{G}_D$, the identifying functional for $P(Y(a, M(a')) = y)$ can

be expressed as

$$\sum_{c_1,c_3,m} P(Y|do(A=a,M=m,C_3=c_3))P(M=m|do(A=a',C_1=c_1)) \times P(C_3=c_3|do(C_1=c_1))P(C_1=c_1). \quad (3.7)$$

Moreover, since – as illustrated in section 2.5.4 of chapter 2 – $P(Y(a)=y)$ is identifiable by Tian's **ID** algorithm, the above expression is likewise identified from the observed data, namely as:

$$\sum_{c_1,c_2,c_3,m} P(Y|A=a,M=m,C_3=c_3)P(M=m|A=a',C_1=c_1,C_2=c_2) \times P(C_1=c_1)P(C_2=c_2)P(C_3=c_3|C_1=c_1). \quad (3.8)$$

Note that, in contrast to the previous example, this functional cannot be reduced to an expression of the form of the adjustment formula.

Nonetheless, it can easily be verified that Pearl's 'piecemeal deconfounding' approach would have yielded the same identification result. However, it would have required searching the space of candidate covariate sets C that not only satisfied cross-world indendence (ii) (i.e. $\{C_1\}$, $\{C_3\}$ or $\{C_1, C_3\}$) but also conditions (v) and (vi) (i.e. only $\{C_1\}$). Shpitser's identification approach is not only (more) complete, but arguably also more insightful as it clarifies that identification of $P(Y(a)=y)$ is the main difficulty in identifying the natural (in)direct effect, which cannot be achieved solely by covariate adjustment.

3.4.4 Stratum-specific natural effects

Before going on to discuss further implications of Shpitser (2013)'s result, one comment merits attention at this point.

Even though the adjustment criterion for natural effects can be shown to serve identification of both population-averaged and stratum-specific natural direct and indirect effects – that is, both $P(Y(a,M(a'))=y)$ and $P(Y(a,M(a'))=y|C^*)$, with $C^* \subseteq C$ – more general identification of stratum-specific natural effects cannot be obtained under (vii) and (viii), as these conditions are sufficient for identifying $P(Y(a,M(a'))=y)$, but not necessarily for identifying $P(Y(a,M(a'))=y|C^*)$.

Whereas complete conditions for identifying stratum-specific natural effects have yet to be articulated, it is clear, at this point, that corresponding identification algorithms or criteria will possibly be slightly more involved than for population-averaged natural effects.[9] This can be appreciated by referring to the increased complexity of the **IDC** algorithm for conditional effects (Figure 2.8 in chapter 2) relative to the **ID** algorithm for marginal effects (Figure 2.5 in chapter 2).

3.5 Complementary identification strategies

The completeness of Shpitser (2013)'s identification approach reveals that Pearl (2001)'s identifying conditions (v) and (vi) may not be necessary, since identifiability of $P(Y(a) = y)$ – i.e. condition (viii) – suffices. Interestingly, the completeness of conditions (vii) and (viii) also highlights that, in some rare cases, cross-world independence (ii) – despite being a sufficient condition for experimental identification (Pearl, 2001) – may not be required either.

3.5.1 Interchanging cross-world assumptions

For example, cross-world independence (ii) is violated in the causal diagram in Figure 3.5A because there is no adjustment set C such that graphical criterion (iii) holds. Nonetheless, $\{M, Y\}$ is not a recanting district because of the assumed absence of a direct path from A to M. $P(Y(a, M(a')) = y)$ can thus be identified as

$$
\begin{aligned}
&P(Y(a, M(L(a'))) = y) \\
&= \sum_{l,m} P(Y(a,m) = y, M(l) = m, L(a') = l) \\
&= \sum_{u,l,m} P(Y(a,m) = y|U = u)P(M(l) = m|U = u)P(L(a') = l)P(U = u) \\
&= \sum_{u,l,m} P(Y = y|A = a, M = m, U = u)P(M = m|L = l, U = u) \\
&\qquad\qquad \times P(L = l|A = a')P(U = u)
\end{aligned}
$$

[9] Thanks to Ilya Shpitser for clearly pointing this out.

$$= \sum_{u,l,m} P(Y = y|A = a, L = l, M = m, U = u)P(M = m|A = a, L = l, U = u)$$
$$\times P(L = l|A = a')P(U = u|A = a, L = l)$$
$$= \sum_{l} P(Y = y|A = a, L = l)P(L = l|A = a').$$

The same result can be obtained more elegantly via expression (3.6):

$$\sum_{l,m} P(Y = y, M = m|do(A = a, L = l))P(L = l|do(A = a'))$$
$$= \sum_{l,m} P(Y = y|A = a, L = l, M = m)P(M = m|A = a, L = l)P(L = l|A = a')$$
$$= \sum_{l} P(Y = y|A = a, L = l)P(L = l|A = a').$$

This novel result can be explained by the fact that the recanting district criterion does not serve to establish identifiability via cross-world independence (ii), but, instead, via an alternative cross-world independence assumption encoded in the associated NPSEM, i.e. $Y(a, m) \perp\!\!\!\perp L(a')$. Indeed, the above derivations illustrate that the mediating instrument L achieves to prevent the conflict between treatment assignments a and a' from taking place *within* the district $\{M, Y\}$ by diverting treatment state a' to itself, thereby fulfilling its mediating role, literally and figuratively. A crucial insight here is that when L is assumed to mediate the entire treatment effect on the mediator M, then the latter is no longer a child of A, and hence, cannot receive any input from A that may conflict with any input to other children of A in the same district.

The above result should, on second thought, not come as a big surprise: since the effect of treatment on the mediator is (assumed to be) entirely

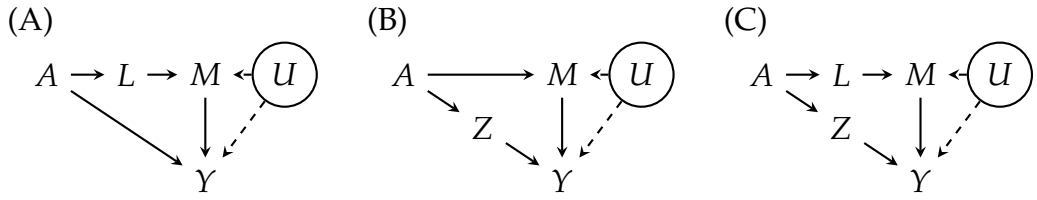


Figure 3.5: Mediation graphs with mediating instruments L for the $A \to M$ path (A), Z for the $A \to Y$ path (B), and a combination of both (C).

mediated by L, and, in addition, L only affects the outcome via M (i.e. L is not an intermediate confounder), L can simply substitute for M. Note that a mediating instrument on the path between treatment and outcome, such as Z in Figure 3.5B, would similarly allow to make progress upon substituting (ii) by cross-world independence $Z(a) \perp\!\!\!\perp M(a')$.

3.5.2 Two types of auxiliary variables

The above examples illustrate that when the treatment effect is identifiable – i.e. (viii) holds – further identification of $P(Y(a, M(a')) = y)$ – by (vii) – can be achieved under NPSEMs with the help of two types of auxiliary variables. Each type has its own distinct strategy of preventing recantation.[10]

The first type – such as C in Figures 3.2B and 3.2C – aims to prevent conflicting treatment assignments within districts by separating nodes of a district that is very likely to recant because of potential mediator-outcome confounding – such as $\{M, Y\}$ in Figure 3.2A – into different districts. Adjustment for this type of covariates specifically aims to strengthen assumption (ii).

The second type – such as L or Z in Figures 3.5A, 3.5B and 3.5C – avoids conflicts in district $\{M, Y\}$, not by separating its nodes, but instead hosting one potential 'troublemaker' in its own district. Such mediating instruments therefore do not aspire to establish assumption (ii), but instead target identification by means of alternative cross-world assumptions that may substitute for assumption (ii). This result is important since Pearl (2014) pointed out that, although mediating instruments can be used to identify $P(M(a) = m|C = c)$ and/or $P(Y(a, m) = y|C = c)$ in order to satisfy conditions (v) and (vi), they cannot aid in (avoiding recantation by) establishing cross-world independence (ii) as this can only be achieved by means of covariate adjustment. However, the recanting district criterion leads to the novel insight that recantation can be avoided in alternative ways.

[10] Note that this classification is analogous to the one often used for auxiliary variables that aid identification of treatment effects, where identification can be achieved via two main strategies: using either the back-door criterion (i.e. standard adjustment for covariates) or the front-door criterion (i.e. sequential adjustment by means of a mediating instrument).

More specifically, we have demonstrated the extended utility of mediating instruments as auxiliary variables that may help to avoid recantation, thereby establishing cross-world independencies that may substitute for cross-world independence (ii).

In section 3.7, we further discuss and illustrate complete identification strategies for generally defined path-specific effects. In the next two sections, we first aim to provide a bridge between the insights provided by this complementary identification strategy on the one hand, and certain recent conceptual developments on the other hand.

3.6 From mediating instruments to conceptual clarity

Contrary to the long-held belief that identification of $P(Y(a, M(a')) = y)$ hinges on the assumption that no mediator-outcome confounding is left unadjusted, mediating instruments arm us with additional identification power in the presence of such unmeasured confounding. This is interesting as it provides researchers with different identification strategies, each relative to a specific set of assumptions. One may use this as a basis for a sensitivity analysis, or adopt the strategy that corresponds with the most plausible assumptions given a certain research context. However, some caution is warranted as mediating instruments can be difficult to justify.

First, the assumption that L or Z is a mediating instrument involves strong and often unrealistic no-direct-effect assumptions (often referred to as *exclusion restrictions*).[11] For instance, for L in Figure 3.5A to be a mediating instrument, it would need to mediate the entire effect of A on M. Despite being a strong assumption, it is partially testable when there is no unmeasured $L - M$ or $A - M$ confounding, as in Figure 3.5A, for then M must be conditionally independent of A, given L. Likewise, for Z in Figure 3.5B to be a mediating instrument for the $A \rightarrow Y$ path, it would need to mediate the entire direct treatment effect on the outcome that is not mediated by M. However, the requirement that Z and M together mediate the entire treatment effect (which implies $Y \perp\!\!\!\perp A|Z, M$) is untestable in the

[11] Needless to say, L or Z may also constitute a set of covariates - rather than being singletons - which satisfy the stated conditions.

presence of unmeasured $M - Y$ confounding.

Second, even if one is willing to assume that L or Z are mediating instruments, the requirement of no unmeasured confounding simply shifts from the mediator-outcome relation to both the instrument-mediator and instrument-outcome relations. This can be seen upon noting that either type of unmeasured confounding results in the instrument being 'absorbed' into the district $\{M, Y\}$, so that it expands to $\{L, M, Y\}$ in Figure 3.5A or to $\{Z, M, Y\}$ in Figure 3.5B. Both of these would be recanting, thereby violating assumption (iv). As for mediator-outcome confounding, neither types of unmeasured confounding can be avoided by treatment randomization.

Third, mediating instruments do not resolve the previously considered identification problems in the presence of intermediate confounding. In fact, the (testable) exclusion restriction that Z does not affect M, in Figure 3.5B, can also be thought of as a constraint that prevents Z from turning into a recanting witness or intermediate confounder. For the same reason the (untestable) exclusion restriction that L does not affect Y (other than through M) is key to identification in Figure 3.5A.

Even though the practical use of mediating instruments, as an alternative route to identification of natural effects, may be debatable, undoubtedly, their added value is more immediate on a conceptual level. In particular, such instruments might help to clearly frame some recent conceptual development that aims to cast mediation analysis into a more strict interventionist paradigm, void of untestable cross-world assumptions (Robins and Richardson, 2010). Before going on to discuss this development in more detail, we briefly sketch some difficulties that may arise when interpreting natural (but also controlled) effects, at least from a interventional point of view.

3.6.1 In search of operational definitions

When it comes to the interpretation of natural direct effects, critics adhering to the slogan 'no causation without manipulation' have repeatedly emphasised the operational question of *how* exactly one may go about blocking the treatment's effect on the mediator, in order to recover $M(0)$ in treated

subjects, without affecting the direct path from treatment to outcome (e.g. Didelez et al., 2006). Inevitably, any answer to this question invokes a mediating instrument, such as L in Figure 3.5A, that can be intervened on in order to prevent treatment from exerting its effect on the mediator.

In order to avoid interpretational ambiguities or strong, untestable cross-world assumptions, critics have therefore proposed to, instead, study 'manipulable' causal quantities, such as controlled direct effects (e.g. Naimi et al., 2014a), which express what the treatment effect would have been if the mediator were kept fixed at some predetermined level m uniformly in the population. As opposed to natural effects' descriptive formulation, their prescriptive interpretation has been claimed to be more directly informative as to potential policy implications of certain interventions. However, apart from the fact that, in many settings, the mediator may often be difficult to manipulate, when aiming to decompose the treatment effect into a mediated and unmediated component, general attempts to define the controlled direct effect's counterpart, a so-called 'controlled indirect effect', have stumbled upon similar operational difficulties. Whereas some would simply define such an effect as the difference between treatment effect and the controlled direct effect, such a definition is not entirely satisfactory, especially in the presence of treatment-mediator interactions, in which case, their interpretation may be highly ambiguous. Without the introduction of a mediating instrument, such as Z in Figure 3.5B, it is indeed difficult to conceive of an intervention that would block only the direct path from treatment to outcome (also see VanderWeele, 2011b).

3.6.2 Deterministic expanded graphs

It thus seems that mediating instruments provide some sort of necessary extension to the original causal diagram[12] that allows for interventionist – some may say, empirically meaningful – interpretations of natural effects. Moreover, they are key for a clear operational definition of controlled

[12]Note that the expanded graphs with mediating instruments, depicted in Figure 3.5, can be marginalized over L and/or Z to result in the original causal diagram in Figure 3.2A, only if the above exclusion restrictions pertaining to L and/or Z hold, such that neither L nor Z is a common cause of any two variables on the original graph in Figure 3.2A.

indirect effects. The conceptual notion of an expanded graph with two mediating instruments, as depicted in Figure 3.5C, corresponds very closely to what has been described in Robins and Richardson (2010).

In this – possibly unrealistic – setting, Shpitser (2013)'s result tells us that identification of $P(Y(a, M(a')) = y)$ may be obtained – provided (viii) holds and L and Z are truly mediating instruments – if L and Z are in separate districts. If not, their district would be recanting and identification would fail. The associated cross-world assumption – i.e. $Z(a) \perp\!\!\!\perp L(a')$ – indeed formalizes the need for no unmeasured confounding between the two instruments. However, this can never be guaranteed unless, as postulated by Robins and Richardson (2010), both L and Z are deterministic functions of (a randomized) treatment A. In that case, both $Z(a)$ and $L(a')$ are constant, and hence trivially independent.[13]

Ironically, this required determinism seems to leave us incapable of pulling apart the causal pathways that we meant to assess in the first place, which brings us back to square one. However, progress can be made if one can conceive of separate interventions on L and Z that would enable to break their perfect correlation. From this perspective, the deterministic characterization of an expanded graph such as Figure 3.5C, gives rise to a specific type of experimental design that requires one to think of L and Z as inherent, but distinct, properties of the treatment, which may be intervened on separately, but when combined, fully capture all of its active ingredients; see section 3.6.3 for an example. The feasibility of such designs thus primarily mirrors the extent to which different active components of treatment or exposure can be conceived of being manipulated in isolation (Didelez, 2013b).[14] Moreover, when combined with the aforementioned exclusion

[13] In addition, as in Robins and Richardson (2010), independence of $Z(a)$ and $L(a')$ can be shown to lead to cross-world assumption (ii).

[14] One may (justly) claim that, in theory, only one of the mediating instruments would need to be a deterministic function of treatment. However, it seems difficult to conceptually conceive of any scenario where only one of the mediating instruments would be fully determined by treatment. Try, for instance, to imagine manipulating A, while L is fully determined by A. Nonetheless, one needs to be able to manipulate L, which captures an inherent aspect of A, while leaving untouched all other aspects of A. This seems to necessarily imply that all other inherent aspects of A need to be captured by another deterministic variable, which would then naturally lead to Z.

restrictions, such designs thus entail separate manipulations of L and Z, which capture distinct but exhaustive features of the treatment to which, respectively, solely M and Y are (directly) responsive. Importantly, this characterization enables to interpret natural effects as specific interventional contrasts.

Consider, for instance, the causal diagram in Figure 3.2A and its expansion in Figure 3.5C, where Z and L are deterministic functions of treatment which can be conceived as two complementary components that fully characterise treatment such that $A = \{L, Z\}$, $a = \{l_a, z_a\}$ and $a' = \{l_{a'}, z_{a'}\}$. Then identification of $P(Y(a, M(a')) = y)$ under the NPSEM associated with the causal diagram in Figure 3.5C is tantamount to identification of the interventional distribution $P(Y(z_a, l_{a'}) = y)$ since

$$\begin{aligned} P(Y(a, M(a')) = y) &= \sum_{l,z,m} P(Y = y, M = m | do(L = l, Z = z)) \\ &\qquad \times P(L = l | do(A = a')) P(Z = z | do(A = a)) \\ &= \sum_{l,z,m} P(Y = y | L = l, Z = z, M = m) P(M = m | L = l) \\ &\qquad \times P(L = l | A = a') P(Z = z | A = a) \qquad (3.9) \\ &= \sum_{l,z} P(Y = y | L = l, Z = z) I(L = l_{a'}) I(Z = z_a) \\ &= P(Y = y | L = l_{a'}, Z = z_a) \\ &= P(Y(z_a, l_{a'}) = y). \end{aligned}$$

The first equality holds by expression (3.6), the second by Tian's **ID** algorithm and the third by the conditional independence $M \perp\!\!\!\perp Z | L$ and determinism.

The mere feasibility of such an experimental design and the plausibility of the two aforementioned exclusion restrictions thus provide the necessary context for interpreting natural effects as manipulable and hence – as critics may claim – policy-relevant parameters (Robins and Richardson, 2010). In other words, instead of actually conducting such an experiment, one could estimate natural effects based on available (experimental or observational) data, provided (vii) and (viii) hold, while the construction of a scientifically plausible story – encoded in a deterministic extended graph – then serves

to license their interventional interpretation. Moreover, the problem of identification of natural effects from available data on observable variables represented in a deterministic extended graph is thus effectively reduced to one of identification of the effect of a joint intervention on $\{L, Z\}$, to which the usual calculus for joint treatment effects (Tian and Pearl, 2003) is applicable, which typically avoids reliance on cross-world assumptions.[15]

3.6.3 Examples

Some existing designs, such as double-blind placebo-controlled trials, were in fact devised in the spirit of Robins and Richardson (2010)'s deterministic extended graphs (Didelez, 2013b). Such trials indeed aim to isolate part of the effect of the drug A that may be attributed to active chemical components Z, and is not mediated by the patient's or doctor's expectations about the effectiveness of the drug M. In such designs it is often reasonable to assume that expectations are solely affected by the knowledge of (possibly[16]) being treated L and that the active component itself does not affect expectations. The natural direct effect of the drug, not mediated by

[15]This perspective leads to the conjecture that, given identifiability of $P(Y(a) = y)$, the recanting district criterion serves to assess whether the identifying functional of the joint interventional distribution $P(Y(z_a, l_{a'}) = y)$ that corresponds to $P(Y(a, M(a')) = y)$ under a deterministic extended graph $\mathcal{G}'$ is expressible in terms of the observed variables in the original graph $\mathcal{G}$ (which excludes deterministic variables such as L and Z). For instance, without actually conducting the aforementioned experiment that involves a joint intervention on L and Z, the distribution of $P(Y(z_a, l_{a'}) = y)$ is not identified by $P(Y = y|L = l_{a'}, Z = z_a)$ if $a \neq a'$, since then $P(L = l_{a'}, Z = z_a) = 0$ in the observed data. Nonetheless, in the absence of unmeasured $M - Y$ confounding under the deterministic extended graph in Figure 3.5C, $P(Y(z_a, l_{a'}) = y)$ is still identifiable: in that case expression (3.9) can be shown to reduce to the mediation formula since the conditional independence $Y \perp\!\!\!\perp L|A, M$ holds. Such reduction is not possible in the presence of unmeasured $M - Y$ confounding since then L and Y are dependent conditional on $\{A, M\}$ because of collider stratification. Similarly, it can be shown that $P(Y(z_a, l_{a'}) = y)$ is identifiable if one has measurements of a mediating instrument for the directed path $L \to M$ or $Z \to Y$.

[16]For ethical reasons, patients need to be informed about the possibility of being assigned to the (placebo) control arm. So, technically speaking, the actual expectation M may not perfectly correspond to $M(1)$, the expectation that would be invoked when (possibly) being 'tricked' into thinking actually being treated. Nonetheless, L is, by design, controlled at the same level for every patient, for instance, by administering similar looking pills with or without the active component, so that the patient nor doctor cannot possibly find out whether one is on active treatment or not.

expectations, could therefore be interpreted as the interventional contrast comparing effectiveness between treatment and (placebo) control arm.[17]

Unfortunately, success is not always guaranteed. Side effects in the treatment arm may for instance affect the expectation of being on active treatment, thereby violating the assumptions of Figure 3.5C. To accommodate for known side effects, active placebos have been designed that mimic side effects of the active treatment (e.g. Didelez, 2013a; Lok, 2016), illustrating that the ability to increase the credibility of required exclusion restrictions may often be highly dependent on the creativity of the researcher (Robins and Richardson, 2010).[18]

In other contexts, experimental designs in the spirit of Robins and Richardson (2010)'s deterministic extended graphs are more difficult to conceive. For instance, even though the JOBS II study (Vinokur and Schul, 1997) involved a job-search skills workshop that targeted specific component processes grounded in psychological theory, it may still be hard to imagine similar interventions or workshops that isolate the distinct triggering elements of separate targeted processes, let alone, to conceive of distinct elements that exclusively affect either re-employment or mental health. Any attempt to endow natural direct and indirect effects with an interventionist interpretation would thus necessarily rely on strong theoretical assertions about the active components of the job training intervention.

3.7 Path-specific effects

In section 3.4, we pointed out that the recanting district criterion not only serves to identify natural effects, but any path-specific effect of a treatment or exposure on an outcome of interest. The utility of this criterion may thus be particularly appealing in settings with multiple causally ordered mediators, as it sheds light on which alternative or more fine-grained decompositions of the total treatment effect can be obtained non-parametrically.

[17] Note that experimental designs reflecting an expanded graph, do not require any measurements on the mediator.

[18] In a strict sense, active placebo designs also violate the required exclusion restrictions. Nonetheless, they enable to arrive at a measure of a direct effect that more closely resembles the natural direct effect of primary interest (Didelez, 2013a).

3.7.1 Alternative decompositions in the presence of multiple mediators or intermediate confounding

Consider again the causal diagram in Figure 3.3A. In the previous sections, it was indicated that, because of the presence of an intermediate confounder (or recanting witness) L, the natural effects with respect to mediator M are not (non-parametrically) identifiable under the NPSEM associated with Figure 3.3A.

When L would be the primary mediator of interest, on the other hand, non-parametric identification of $P(Y(a, L(a')) = y)$ is trivial, since, in Figure 3.3A there is no recanting district for the natural direct and indirect effect with respect to L. Specifically, the natural indirect effect with respect to L

$$\begin{aligned} &E\{Y(1, L(1)) - Y(1, L(0))\} \\ &= E\{Y(1, L(1), M(1, L(1))) - Y(1, L(0), M(1, L(0)))\} \end{aligned} \tag{3.10}$$

consists of the directed paths in $\pi = \{A \to L \to Y, A \to L \to M \to Y\}$. Since it is never the case that a path in π and a path not in π traverses children of A that are in the same district, it can indeed be concluded that there is no recanting district for the π-specific effect of interest, i.e. in this case the natural direct effect with respect to L. Again, by symmetry, there is no recanting district for the natural indirect effect with respect to L. Hence, since both (vii) and (viii) hold, $P(Y(a, L(a')) = y)$ is identifiable.

Unlike the natural indirect effect with respect to L, the natural direct effect

$$\begin{aligned} &E\{Y(1, L(0)) - Y(0, L(0))\} \\ &= E\{Y(1, L(0), M(1, L(0))) - Y(0, L(0), M(0, L(0)))\}, \end{aligned}$$

which consists of directed paths in $\pi = \{A \to Y, A \to M \to Y\}$, can be further decomposed into more fine-grained components or path-specific effects, i.e. into a direct effect not mediated by either L nor M

$$E\{Y(1, L(0), M(0, L(0))) - Y(0, L(0), M(0, L(0)))\}, \tag{3.11}$$

as captured by the directed path $A \to Y$, and part of the effect mediated by M that bypasses L

$$E\{Y(1, L(0), M(1, L(0))) - Y(1, L(0), M(0, L(0)))\}, \tag{3.12}$$

sometimes referred to as the *partial* (Huber, 2014) or *semi-natural* (Pearl, 2014) indirect effect with respect to M, as captured by the path $A \to M \to Y$. Identifiability of each of these component effects can easily be verfied via the recanting district criterion.

3.7.2 Coarser decompositions in the presence of unobserved confounding

Whereas in Figure 3.3A, a three-way decomposition of the total treatment effect could be obtained, unmeasured confounding in Figures 3.3B, 3.3C and 3.3D only enables to obtain a two-way decomposition, leading to coarser decompositions of the total treatment of A on Y. Again, the recanting district criterion provides insight into this.

For instance, under the NPSEM associated with the causal diagram in Figure 3.3B, the natural effects with respect to L are still identifiable, because, by the same logic applied in the previous paragraph, the district $\{M, Y\}$ is not recanting relative to these effects. However, the natural direct effect with respect to L cannot be further decomposed, as in Figure 3.3A, because $\{M, Y\}$ is recanting relative to the 'direct' effect captured by the path $A \to Y$ and, by symmetry, also relative to the partial indirect effect with respect to M. Under the NPSEM representation of the causal diagram in Figure 3.3C, on the other hand, we can no longer identify the natural effects with respect to L nor the partial indirect effect with respect to M, as $\{L, M\}$ is a recanting district for each of these effects. However, we may still identify the natural indirect effect with respect to $\{L, M\}$ as a joint mediator

$$\begin{aligned} &E\{Y(1, L(1), M(1)) - Y(1, L(0), M(0))\} \\ &= E\{Y(1, L(1), M(1, L(1))) - Y(1, L(0), M(0, L(0)))\}, \end{aligned}$$

which consists of paths in $\pi = \{A \to L \to Y, A \to M \to Y, A \to L \to M \to$

$Y\}$ and can be considered a combination of the natural indirect effect with respect to L and the partial indirect effect with respect to M. By symmetry, we can also identify the natural direct effect with respect to $\{L, M\}$, which corresponds to the directed path $A \to Y$.

Finally, in Figure 3.3D, the partial indirect effect with respect to M is identifiable. However, its complement path-specific effect cannot be further decomposed into the natural indirect effect with respect to L and the direct effect along the directed path $A \to Y$, because the district $\{L, Y\}$ is recanting with respect to both of these more fine-grained path-specific effects.

For instance, in section 3.2.4, we mentioned that the natural indirect effect of job-search skills workshop participation A on the presence of depressive symptoms Y mediated by re-employment M is not identifiable if $M - Y$ confounders, such as increased self-efficacy L, are also affected by the job-search intervention. Interestingly, we may still identify e.g. the partial indirect effect mediated by re-employment but not by increased self-efficacy

$$E\{Y(1, L(1), M(1, L(1))) - Y(1, L(1), M(0, L(1)))\}, \tag{3.13}$$

even in the presence of unobserved confounding between sense of self-efficacy and mental health, as depicted in Figure 3.3D (also see Miles et al., 2014).

Note that the availability of mediating instruments may aid to recover all three aforementioned component effects in the presence of unmeasured confounding in Figures 3.3B, 3.3C and 3.3D. Suppose, for instance, that interest lies in identification of the partial indirect effect with respect to M in Figures 3.3B and 3.3C. If one is willing to make certain strong and possibly untestable exclusion restrictions, progress can be made with the help of a mediating instrument on the direct path $A \to M$ or the paths $A \to Y$ or $A \to L$ in Figures 3.3B and 3.3C, respectively.

In chapter 5, we cast estimation of alternative and fine-grained decompositions into a more general modeling approach for mediation analysis that enables dealing with multiple sequential mediators, such as L and M in the causal diagram in Figure 3.3A. Moreover, in the same chapter, we further

extend the adjustment criterion to accommodate effect decomposition in such settings with multiple sequential mediators. Similar to single mediator settings in chapter 4, we will, however, restrict our focus to identification by this generalized adjustment criterion for mediation analysis, since it leads to a standard form of identification result that can be seen as a generalization of Pearl's mediation formula. This, in turn, enables constructing generally applicable semi-parametric estimators, such as those discussed in section 3.4.2.[19] In chapter 6, we further discuss the feasibility to incorporate results obtained by more general identification strategies in this modeling framework.

3.7.3 Costs of fine-grained decompositions: assumptions

An appealing side effect of adopting the recanting district criterion is that it makes explicit formulations of cross-world independence assumptions redundant. One should not forget, though, that non-parametric identification of generally defined path-specific effects relies on such untestable assumptions, just as for natural effects. Their precise nature can be shown to vary depending on the path-specific effect of interest that one aims to identify (Shpitser, 2013). Moreover, the number of such assumptions on which identification relies, increases with the number of component effects (e.g., if one aims to obtain fine-grained decompositions into more than two component effects). This proliferation of untestable cross-world assumptions also makes it harder to completely avoid such assumptions via particular, feasible experimental designs which enable interventionist interpretations of generally defined path-specific effects, as discussed for natural effects in section 3.6.2. A detailed discussion of deterministic mediating instruments for generally defined path-specific effects is, however, beyond the scope of this chapter (although see Robins and Richardson, 2010).

[19] An additional rationale for focusing on identification by generalizations of the adjustment formula, beyond its standard form results, is that, as indicated in section 3.4.4, the adjustment criterion also serves to identify stratum-specific natural effects.

3.7.4 Costs of fine-grained decompositions: interpretation

Fine-grained decompositions into more than two path-specific effects along multiple causally ordered mediators also bring about additional conceptual or interpretational challenges. For instance, in chapter 5, we indicate that the aforementioned three-way decomposition into path-specific effects (in Figure 3.3A) can be parameterized by a natural effect model for the mean of recursively nested counterfactual outcomes

$$E\{Y(a, L(a'), M(a'', L(a')))\} = g^{-1}\{\gamma_0 + \gamma_1 a + \gamma_2 a' + \gamma_3 a'' + \gamma_4 a a' + \gamma_5 a a'' + \gamma_6 a' a'' + \gamma_7 a a' a''\}.$$

This model highlights that, in total, six possible three-way decompositions can be obtained by differently apportioning the interaction parameters γ_4 to γ_7. These decompositions involve four distinct instances of each of the path-specific effects of interest. For instance, if $g(\cdot)$ corresponds to the identity link function, the path-specific effects of the particular three-way decomposition discussed in section 3.7.1, as defined by expressions (3.10), (3.11) and (3.12), are captured by $\gamma_2 + \gamma_4 + \gamma_6 + \gamma_7$, γ_1, and $\gamma_3 + \gamma_5$, respectively. Depending on the levels at which a, a' and a'' are set, other results might be obtained. For instance, the partial indirect effect with respect to M can be defined as (3.13), instead of (3.12), which is captured by $\gamma_3 + \gamma_5 + \gamma_6 + \gamma_7$.

In general, in the presence of k causally ordered mediators, maximally $k + 1$ fine-grained path-specific effects are (possibly) non-parametrically identifiable,[20] each of which can be defined in 2^k different ways. This multitude of definitions gives rise to $(k + 1)!$ different ways in which path-specific effects of interest can be combined to produce the total treatment effect (also see Daniel et al., 2015). Differences in interpretation between distinct instances of a path-specific effect may, however, be subtle and often a substantive motivation may be lacking to prefer one specific decomposition

[20] Specifically, in chapter 5, we target identification of the most fine-grained $(k + 1)$-way decomposition characterized in terms of k path-specific effects identifiable by the recanting witness criterion (Avin et al., 2005). In the absence of unmeasured confounding, we may indeed obtain $(k + 1)$-way decompositions (since the recanting district criterion reduces to the recanting witness criterion in that case). In semi-Markovian NPSEMs, on the other hand, we may need to settle with coarser decompositions (see section 3.7.2).

over another. The absence of interactions between path-specific effects can thus substantially facilitate interpretations of the targeted effects (Daniel et al., 2015).

Accordingly, in chapter 5, we indicate that flexible and parsimonious modeling and estimation approaches seem unavoidable to reduce increasing complexity in the face of multiple causally ordered mediators to more manageable proportions. These enable assessing evidence of interaction, and expressing effects on the scale at which the evidence of interaction is weak. Alternatively, if interest lies in only one – or fewer – path-specific effects, as is often the case in practice, one may redirect focus on coarser and therefore less ambitious decompositions that involve only those specific component effects of interest. These may often be identifiable under weaker conditions, as discussed in section 3.7.2, and moreover increase interpretability.

Chapter 4

Flexible mediation analysis with a single mediator

This chapter is based on the following paper: Steen, J., Loeys, T., Moerkerke, B., Vansteelandt, S. (2016). Medflex: An R Package for Flexible Mediation Analysis Using Natural Effect Models. *Journal of Statistical Software*, in press.

Mediation analysis is routinely adopted by researchers from a wide range of applied disciplines as a statistical tool to disentangle the causal pathways by which an exposure or treatment affects an outcome. The counterfactual framework provides a language for clearly defining path-specific effects of interest and has fostered a principled extension of mediation analysis beyond the context of linear models. This chapter describes medflex, an R package that implements some recent developments in mediation analysis embedded within the counterfactual framework. The medflex package offers a set of ready-made functions for fitting natural effect models, a novel class of causal models which directly parameterize the path-specific effects of interest, thereby adding flexibility to existing software packages for mediation analysis, in particular with respect to hypothesis testing and parsimony. In this chapter, we give a comprehensive overview of the functionalities of the medflex package.

4.1 Introduction

Empirical studies often aim at gaining insight into the underlying mechanisms by which an exposure or treatment affects an outcome of interest. Mediation analysis, as popularized in psychology and the social sciences by Judd and Kenny (1981) and Baron and Kenny (1986), has been widely adopted as a statistical tool to shed light on these mechanisms, by enabling the decomposition of total causal effects into an *indirect* effect through a hypothesized intermediate variable or mediator and the remaining *direct* effect. Although its initial formulations were restricted to the context of linear regression models, several attempts have been made to extend the application of traditional estimators for indirect effects (i.e. product-of-coefficients and difference-in-coefficients estimators) beyond linear settings (e.g. MacKinnon and Dwyer, 1993; MacKinnon et al., 2007; Hayes and Preacher, 2010; Iacobucci, 2012). However, these extensions lack formal justification and yield effect estimates that are often difficult to interpret (e.g. Pearl, 2012).

Recent advances from the causal inference literature (e.g. Albert, 2008; Albert and Nelson, 2011; Avin et al., 2005; Imai et al., 2010b; Pearl, 2001, 2012; Robins and Greenland, 1992; VanderWeele and Vansteelandt, 2009, 2010) have furthered these developments and improved both inference and interpretability of direct and indirect effect estimates in nonlinear settings by building on the central notion of counterfactual or potential outcomes. This notion provides a framework that has aided in (i) formally defining direct and indirect effects (in a way that is not tied to a specific statistical model), (ii) describing the conditions required for their identification (unveiling and formalizing often implicitly made causal assumptions) and (iii) assessing the robustness of empirical findings against violations of these identification conditions (i.e. sensitivity analysis).

For instance, Imai et al. (2010a) proposed mediation analysis techniques that can be applied within a larger class of nonlinear models. They implemented these in a user-friendly R package, called mediation (Tingley et al., 2014b; see Hicks and Tingley, 2011 for a version in Stata (StataCorp, 2013) with more limited functionality). More recently, Valeri and VanderWeele (2013) reviewed the latest developments in mediation analysis for non-

linear models, focusing on exposure-mediator interactions, and provided SAS (SAS Institute Inc., 2014) and SPSS (IBM Corporation, 2013) macros, enabling practitioners to easily conduct these methods using well-known commercial packages. Similarly, Emsley and Liu (2013) and Muthén and Asparouhov (2015) described how direct and indirect effects as defined in the counterfactual framework can be estimated in Stata and via extended types of structural equation models in Mplus (Muthén and Muthén, 2012), respectively.

In this chapter, we introduce medflex (Steen et al., 2015), an R package that enables flexible estimation of direct and indirect effects while accommodating some of the limitations of other available packages. More specifically, we make use of novel so-called *natural effect models* (Lange et al., 2012, 2014; Loeys et al., 2013; Vansteelandt et al., 2012b), which directly parameterize the target causal estimands on their most natural scale. This renders formal testing and interpretation more straightforward compared to other approaches as implemented in the aforementioned software applications. The medflex package is freely available from the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/package=medflex` (R Core Team, 2015).

Throughout, the functionalities of the medflex package will be illustrated using data from a survey study that was part of the Interdisciplinary Project for the Optimization of Separation trajectories (Ghent University and Catholic University of Louvain, 2010). This large-scale project involved the recruitment of individuals who divorced between March 2008 and March 2009 in four major courts in Flanders. It aimed to improve the quality of life in families during and after the divorce by translating research findings into practical guidelines for separation specialists (such as lawyers, judges, psychologists, welfare workers...) and by promoting evidence-based policy. The corresponding dataset `UPBdata` is included in the package and involves a subsample of 385 individuals who responded to a battery of questionnaires related to romantic *relationship* characteristics (such as adult attachment style) and *breakup* characteristics (such as breakup initiator status, experiencing negative affectivity and engaging in unwanted pursuit behaviors; UPB) (De Smet et al., 2012). Respondents were asked to imagine their former

partner as well as possible and to remember how they generally felt in their relationship *before* the breakup when completing the attachment style questionnaire. The mediation hypothesis of interest concerned the question whether the level of emotional distress or negative affectivity experienced *during* the breakup can be regarded as an intermediate mechanism (M) through which attachment style towards the ex-partner *before* the breakup (A) exerts its influence on displaying UPBs *after* the breakup (Y) (Loeys et al., 2013).

In the next section, we briefly introduce the *mediation formula* (Pearl, 2001, 2012; Petersen et al., 2006; Imai et al., 2010b), which is the predominant vehicle for effect decomposition within the counterfactual framework. Advantages of natural effect models over direct application of the mediation formula will also be discussed in more detail. We then focus on two missing data techniques for fitting these models and demonstrate how these approaches can be implemented in R using the medflex package (section 4.3). Next, we demonstrate how different types of exposure and mediator variables can be dealt with (section 4.4) and how to assess effect modification of natural effects (i.e. exposure-mediator interactions and moderated mediation) (section 4.5). Tools are provided for calculating and visualizing different causal effects estimates (section 4.6) and for estimating population-average natural effects (section 4.7) and natural indirect effects as defined through multiple intermediate pathways jointly (section 4.8). In section 4.9, we further elaborate on modeling demands and missing data, two aspects that may need to be taken into consideration by practitioners when choosing between the two main estimation approaches offered by the package. Finally, we conclude with some final remarks and list some extensions of the package which are planned to be implemented in the near future (section 4.10).

4.2 The mediation formula

4.2.1 Counterfactual outcomes and effect decomposition

A major appeal of the counterfactual framework is that it enables to decompose the total causal effect into a so-called *natural* direct and *natural* indirect effect, irrespective of the data distribution or scale of the effect. Readers familiar with counterfactual notation, definitions and assumptions for natural direct and indirect effects may wish to skip to section 4.2.2.

Let $Y_i(a)$ denote the potential outcome for subject i that had been observed if, possibly contrary to the fact, i had been assigned to treatment (or exposure level) a. For a binary exposure (with $A = 1$ for the exposed and $A = 0$ for the unexposed), the individual-level causal effect can then be expressed by comparing $Y_i(1)$ to $Y_i(0)$, whereas the population average total causal effect can be expressed as $E\{Y(1) - Y(0)\}$. Similarly, direct and indirect effects have been defined in terms of counterfactual outcomes. For instance, the definition of the so-called *controlled* direct effect reflects the traditional notion of measuring the effect of exposure while fixing the mediator M at the same value m for all subjects (Robins and Greenland, 1992). Using counterfactual notation, this effect can be expressed as

$$\text{CDE}(m) = E\{Y(1, m) - Y(0, m)\},$$

where $Y(a, m)$ denotes the potential outcome that would have been observed under exposure level a and mediator value m.

Robins and Greenland (1992) introduced an alternative definition that invokes so-called *composite* or *nested* counterfactuals, $Y(a, M(a'))$. For instance, the (pure) natural direct effect

$$\text{NDE}(0) = E\{Y(1, M(0)) - Y(0, M(0))\}$$

expresses the expected exposure-induced change in outcome when keeping the mediator fixed at the value that had naturally been observed if unexposed. By considering potential intermediate outcomes $M(a')$ rather than a fixed mediator value m, these authors offered a definition of direct

effect that both allows for natural variation in the mediator and provides a complementary operational definition for the indirect effect (which the definition of the controlled direct effect does not). That is, under the composition assumption, which states that $Y(a, M(a)) = Y(a)$ (VanderWeele and Vansteelandt, 2009), the difference between the average total effect $E\{Y(1) - Y(0)\}$ and the (pure) natural direct effect yields an expression for the (total) natural indirect effect

$$\text{NIE}(1) = E\{Y(1, M(1)) - Y(1, M(0))\}.$$

This reflects the expected difference in outcome if all subjects were exposed but their mediator value had changed to the value it would take if unexposed.

Adopting this counterfactual notation naturally leads to framing causal inference as a missing data problem (Holland, 1986): for each subject i, only one counterfactual outcome, i.e. $Y_i = Y_i(A_i, M_i(A_i))$, is observed. Consequently, identification of natural effects relies on rather strong causal assumptions. In the context of mediation analysis, the most commonly invoked conditions for identification can be encoded in a causal diagram (such as Figure 4.2) interpreted as a non-parametric structural equation model with independent error terms (NPSEM; Pearl, 2001). More specifically, upon adjustment for a given set of observed baseline covariates C, such model implies certain independencies among variables and potential outcomes which have been proposed as sufficient conditions for non-parametric or model-free identification of natural effects. However, this adjustment set C needs to be carefully selected, such that it is deemed sufficient to control for confounding (i) between exposure and outcome, thereby satisfying

$$Y(a, m) \perp\!\!\!\perp A|C \qquad \text{for all levels of } a \text{ and } m, \qquad \text{(A1)}$$

(ii) between exposure and mediator, thereby satisfying

$$M(a) \perp\!\!\!\perp A|C \qquad \text{for all levels of } a, \qquad \text{(A2)}$$

and (iii) between mediator and outcome (after adjustment for the exposure), thereby satisfying

$$Y(a,m) \perp\!\!\!\perp M|A=a,C \qquad \text{for all levels of } a \text{ and } m. \tag{A3}$$

In addition to these *no omitted confounders* assumptions, identification requires the further *cross-world independence* assumption (Pearl, 2001)

$$Y(a,m) \perp\!\!\!\perp M(a')|C \qquad \text{for all levels of } a, a' \text{ and } m, \tag{A4}$$

which is satisfied under a NPSEM when no confounders of the mediator-outcome relationship (whether observed or unobserved) are affected by the exposure (i.e. no intermediate or exposure-induced confounding).

Whereas the first two assumptions by definition hold in randomized experiments, the other two assumptions may not.[1] Although Judd and Kenny (1981) initially pointed to its importance, assumption (A3) since has largely been ignored in much of the social sciences literature, as evidenced by many mediation studies not adjusting for confounders of the mediator-outcome relationship. In recent years, however, this issue has been brought back to attention within the social sciences (e.g. Bullock et al., 2010; MacKinnon, 2008; Mayer et al., 2014).

Assumption (A4) is more difficult to grasp intuitively. It is a strong assumption because, in contrast to the other three conditions, it is impossible to design a study that would be able to validate it (Robins and Richardson 2010; although see Imai et al. 2013 for a notable attempt).

The interested reader may refer to VanderWeele and Vansteelandt (2009) for a more detailed and intuitive account of these identification assumptions (or to Petersen et al. 2006 or Imai et al. 2010b for alternative sets of assumptions).

[1] Note that assumption (A1) is sufficient for identifying total causal effects, whereas identification of controlled direct effects can be obtained under assumptions (A1) and (A3).

4.2.2 The mediation formula

The language of counterfactuals has enabled to clearly define causal effects in a more generic, non-parametric way, but has also promoted a more principled approach to estimating these effects than the one offered by the traditional SEM literature from the social sciences, which was mainly entrenched in parametric linear regression. The main identification result (Pearl, 2001; Imai et al., 2010b), which Pearl (2012) referred to as the *mediation formula*, has played a pivotal role in this regard. It prescribes estimating the expected value of nested counterfactuals by standardizing predictions from the outcome model corresponding to exposure level a under the mediator distribution corresponding to exposure level a':

$$E\left\{Y(a, M(a'))|C\right\} = \sum_m E(Y|A = a, M = m, C)P(M = m|A = a', C).$$

This weighted sum can be calculated based on any type of statistical model and has been shown to yield closed-form expressions for the natural indirect effect that encompass the traditional difference-in-coefficients and product-of-coefficient estimators when confined to strictly linear models (e.g. VanderWeele and Vansteelandt, 2009; Pearl, 2012). However, as soon as moving beyond linear settings, the latter estimators no longer coincide with their corresponding mediation formula expressions and no longer yield readily interpretable causal effect estimates (as formalized in the counterfactual framework).[2]

More recently, closed-form expressions for natural direct and indirect effects as defined on both additive and ratio scales have been derived for a limited number of nonlinear scenarios (VanderWeele and Vansteelandt, 2009, 2010; Valeri and VanderWeele, 2013).

[2]Muthén and Asparouhov (2015) give an intuitive account for SEM practitioners explaining why the product-of-coefficient estimator fails when applied in nonlinear settings or settings involving exposure-mediator interactions. Nonetheless, the product-of-coefficients method can still be useful for testing the null hypothesis of no indirect effect (VanderWeele, 2011a; Vansteelandt et al., 2012b).

4.2.3 Applying the mediation formula in practice

Software applications for obtaining closed-form solutions derived from the mediation formula, as well as their corresponding Delta method (or bootstrap) standard errors, have been made available as SPSS and SAS macros (Valeri and VanderWeele, 2013) and as the Stata module PARAMED (Emsley and Liu, 2013). More recently, Muthén and Asparouhov (2015) demonstrated how natural effect estimates can be obtained via extended types of structural equation models in Mplus, even in the presence of latent variables. However, such closed-form expressions can often not readily be obtained, for instance when combining a linear model for the mediator and a logistic model for the outcome.

Imai et al. (2010b) addressed this issue and instead suggested a more generic approach based on Monte-Carlo integration methods, which they implemented in the R package mediation (Tingley et al., 2014b). Whereas its lightweight version in Stata (Hicks and Tingley, 2011) and the Stata module gformula (Daniel et al., 2011), which adopts a similar simulation-based approach, are restricted to parametric models, this R package additionally allows to specify semi-parametric models for the mediator and outcome. Despite being computationally intensive, these offer more flexibility than the applications based on a purely analytical approach. In addition, the mediation package offers useful extensions, such as methods for dealing with multiple mediators and treatment noncompliance, while at the same time enabling users to evaluate the robustness of their findings to potential unmeasured confounding in a widely applicable sensitivity analysis.

A drawback of direct application of the mediation formula, however, is that combinations of simple models for the mediator and for the outcome often result in complex expressions for natural direct and indirect effects (Lange et al., 2012; Vansteelandt et al., 2012b). For instance, when using logistic regression models

$$\begin{aligned} \text{logit}P(M = 1|A, C) &= \alpha_0 + \alpha_1 A + \alpha_2 C \\ \text{logit}P(Y = 1|A, M, C) &= \beta_0 + \beta_1 A + \beta_2 M + \beta_3 C \end{aligned} \tag{0}$$

for binary mediators and outcomes, the mediation formula yields

$$\begin{aligned} &P(Y(a, M(a')) = 1|C) \\ &= \text{expit}\left(\beta_0 + \beta_1 a + \beta_2 + \beta_3 C\right) \text{expit}\left(\alpha_0 + \alpha_1 a' + \alpha_2 C\right) \\ &\quad + \text{expit}\left(\beta_0 + \beta_1 a + \beta_3 C\right) \left\{1 - \text{expit}\left(\alpha_0 + \alpha_1 a' + \alpha_2 C\right)\right\}, \end{aligned}$$

an expression which depends on exposure and covariate levels in a complicated way. Even though none of the postulated models include interaction terms reflecting effect modification, the corresponding direct and indirect effect estimates will vary with different exposure or covariate levels. This is also illustrated in Figure 4.1, which depicts estimates for the natural indirect effect odds ratio, as obtained by applying the mediation formula to these models fitted to our example dataset (using a dichotomized version of the mediator and baseline covariates C including gender, age and education level). As pointed out before by Lange et al. (2012) and Vansteelandt et al. (2012b), these convoluted expressions render results difficult to report and hypothesis testing (e.g. testing for moderated mediation) infeasible, as it may turn out impossible to find plausible models for the mediator and outcome that combine into effect expressions that do not depend on covariate levels. In certain cases, this complexity can pose a major impediment to routine application of the mediation formula.

Moreover, the mediation package only provides natural effect estimates on the additive scale. This may complicate estimation and inference in nonlinear outcome models, mainly when dealing with continuous exposures or covariates, because of induced nonadditivity. Specifically, because the indirect effect is not encoded by a single parameter, but may take on a different value for each level of a, the null hypothesis of no indirect effect over the entire range of exposure levels becomes difficult to test. Similarly, although the mediation package enables users to test for effect modification in nonlinear models (i.e. either treatment-mediator interactions or moderated mediation), these hypothesis tests probe research questions in terms of e.g. risk differences that are tied to pre-specified exposure or covariate levels. A concern is that these levels might, at least in some applications, need to be chosen in a rather arbitrary way (Loeys et al., 2013).

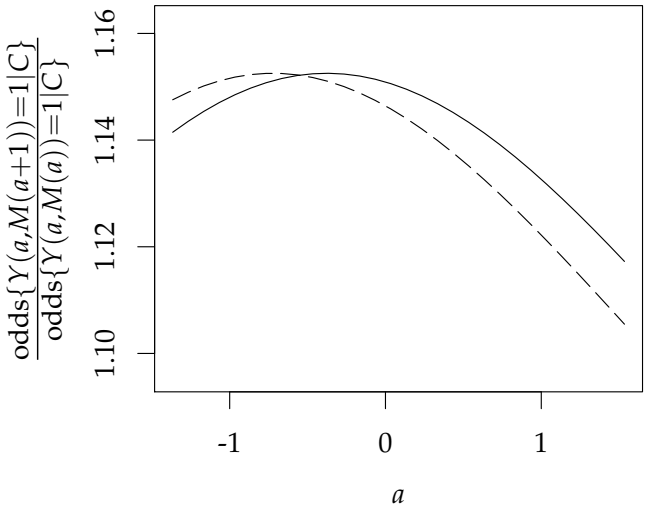


Figure 4.1: Estimated (total) natural indirect effect odds ratios corresponding to a one-unit change in anxious attachment level as a function of different reference levels for anxious attachment level a (as obtained through direct application of the mediation formula). These are conditional estimates for 43-year-old men (solid curve) and women (dashed curve) with intermediate education levels.

An approach that circumvents the aforementioned complexity but is closely related to application of the mediation formula was recently proposed by Lange et al. (2012) and Vansteelandt et al. (2012b). These authors proposed to directly model natural effects and introduced a novel class of mean models for nested counterfactuals, which they termed *natural effect models* (also see van der Laan and Petersen, 2008, for a similar approach). This approach is implemented in the medflex package and provides a viable alternative to the aforementioned software applications because

- it can handle a larger class of parametric models for the mediator and outcome than the software applications that rely on closed-form expressions (see section 4.4),

- estimates can be expressed on more natural effect scales (i.e. a scale that corresponds to the link-function of the outcome model), thereby avoiding potential induced dependence on exposure or covariate levels characteristic for the additive scale,

- natural effect models simplify testing since the hypotheses of interest can always be captured by a finite set of model parameters,
- for the most common types of parametric models robust standard errors (based on the sandwich estimator) are available as an alternative to more computer-intensive bootstrap standard errors.

In the next section, we describe this novel class of causal models together with two different estimation approaches that have been suggested in Lange et al. (2012) and Vansteelandt et al. (2012b).

4.3 Mediation analysis via natural effect models

Natural effect models are conditional mean models for nested counterfactuals $Y(a, M(a'))$:

$$E\{Y(a, M(a'))|C\} = g^{-1}\{\beta' W(a, a', C)\}$$

with $g(\cdot)$ a known link function (e.g. the identity or logit link), $W(a, a', C)$ a known vector with components that may depend on a, a' and C, and β a vector including parameters that encode the natural effects of interest. It can, for instance, easily be inferred that in model

$$E\{Y(a, M(a'))|C\} = \beta_0 + \beta_1 a + \beta_2 a' + \beta_3 C,$$

β_1 captures the natural direct effect whereas β_2 captures the natural indirect effect, both corresponding to a one-unit increase in the exposure level. With $g(\cdot)$ the log-link function, for example, the Poisson regression model

$$\log E\{Y(a, M(a'))|C\} = \beta_0 + \beta_1 a + \beta_2 a' + \beta_3 C,$$

enables to quantify the natural direct and indirect effect for count outcomes on a more natural, multiplicative scale. Specifically, in this model, $\exp(\beta_1)$ captures the natural direct effect rate ratio

$$\frac{E\{Y(a+1, M(a))|C\}}{E\{Y(a, M(a))|C\}}$$

whereas $\exp(\beta_2)$ captures the natural indirect effect rate ratio

$$\frac{E\{Y(a, M(a+1))|C\}}{E\{Y(a, M(a))|C\}},$$

corresponding to a one-unit increase in exposure level. Since each of the effects or quantities of interest are encoded by parameters indexing the natural effect model, the aforementioned limitations related to direct application of the mediation formula can be overcome. As will be illustrated, this facilitates interpretation and hypothesis testing in nonlinear settings.

4.3.1 Fitting natural effect models

Before describing the two main approaches for fitting natural effect methods, we first return to our motivating example. The corresponding dataset will then be used to both illustrate these approaches and to demonstrate how they can be implemented in R.

To install the most recent version of the medflex package available from CRAN, use the command

```
install.packages("medflex")
```

After loading the package, displaying the first few rows of the example dataset `UPBdata` provides some insight into the data:

```
library("medflex")
data("UPBdata")
head(UPBdata)
```

```
##      att attbin attcat negaff  initiator gender educ age UPB
## 1  1.001      1      M  0.840     myself      F    M  41   1
## 2 -0.709      0      L -1.257       both      M    M  42   0
## 3 -0.709      0      L -1.202       both      F    H  43   0
## 4  0.606      1      M -0.374 ex-partner      M    H  52   1
## 5  0.212      1      M  1.945 ex-partner      M    M  32   1
## 6  2.052      1      H -0.816 ex-partner      M    H  47   0
```

De Smet et al. (2012) and Loeys et al. (2013) proposed emotional distress or the amount of negative affectivity experienced during the breakup as a mediating variable for the effect of attachment style towards the ex-partner before the breakup on displaying unwanted pursuit behaviors after the breakup. Figure 4.2 depicts the causal diagram (Pearl, 1995a) that reflects this mediation hypothesis along with its aforementioned identification assumptions.

As direct and indirect effects are most easily understood for a binary exposure, we will use a dichotomized version of anxious attachment level (`attbin`) for didactive purposes. Moreover, negative affectivity (`negaff`) has been standardized to allow for easily interpretable effect estimates. The outcome variable unwanted pursuit behavior (`UPB`) indicates whether (=1) or not (=0) the respondent has engaged in any unwanted pursuit behaviors.

A relatively simple natural effect model is the logistic model

$$\text{logit}P\left\{Y(a, M(a')) = 1|C\right\} = \beta_0 + \beta_1 a + \beta_2 a' + \beta_3 C, \tag{4.1}$$

with a and a' corresponding to hypothetical levels of the dichotomized version of the anxious attachment variable (i.e. 0 for lower than average or 1 otherwise), $M(a')$ to the level of negative affectivity that would have been reported if anxious attachment level were set to a', and $Y(a, M(a'))$ to the UPB perpetration status that would have been observed if anxious attachment level were set to a and negative affectivity were set to the level that would have been reported if anxious attachment style were set to a'. To control for confounding, we condition on a set of baseline covariates C:

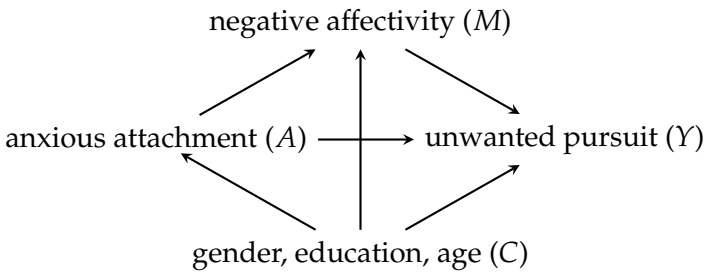


Figure 4.2: Causal diagram reflecting the mediation hypothesis.

age (in *years*), gender and education level (`educ`; with `H` or 'high' indicating having obtained at least a bachelor's degree, `M` or 'intermediate' indicating having finished secondary school and `L` or 'low' otherwise). As emphasized earlier, the selection of such an adjustment set needs careful consideration in order to meet assumptions (A1)-(A4). For illustrative purposes, the current set of baseline covariates C will, possibly contrary to the fact, be considered sufficient to control for confounding throughout the remainder of this chapter.

As an illustration, we schematically display the first two observations in Table 4.1. For each individual or observation unit i, only the counterfactual outcome $Y_i(A_i, M_i(A_i))$, corresponding to $Y_i(a, M_i(a'))$ with a and a' equal to the observed exposure level A_i, is observed. Postulating a model for nested counterfactuals that encodes both natural direct and indirect effects requires data in which either a or a' can be kept fixed within each individual while allowing the other variable to vary. Such a procedure amounts to expanding the data along unobserved (a, a') combinations, as illustrated by the grey entries in Table 4.1. Although, for the data at hand, three (a, a')

i	A_i	a	a'	$Y_i(a, M_i(a'))$
1	1	1	1	Y_1
1	1	1	**0**	.
1	1	**0**	1	.
1	1	**0**	**0**	.
2	0	0	0	Y_2
2	0	0	**1**	.
2	0	**1**	0	.
2	0	**1**	**1**	.
⋮	⋮	⋮	⋮	⋮

Table 4.1: Schematic display of the expanded dataset with missing counterfactual outcomes.

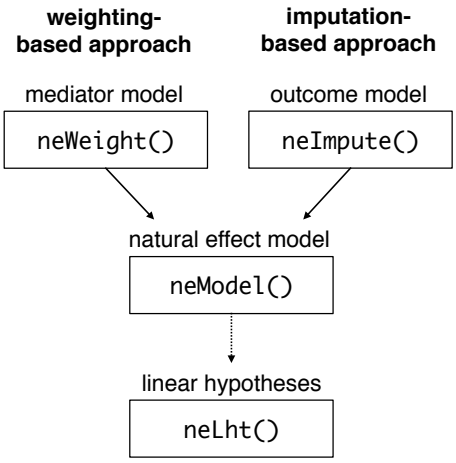


Figure 4.3: Workflow of the medflex package.

combinations are unobserved for each individual, to disentangle natural direct and indirect effects, it is sufficient to introduce only one additional observation corresponding to an unobserved combination for which a does not equal a'.

Fitting natural effect models then entails using well-established methods to deal with missingness in the outcome, which results from expanding the data. Throughout, we will describe a weighting- and an imputation-based approach, which, as outlined below, differ mainly in terms of the statistical working models on which they rely (Vansteelandt, 2012).

Data expansion is highly similar for both approaches, but subsequent algorithms for data preparation differ depending on the type of working model. In the medflex package, these two steps are implemented in the functions `neWeight` and `neImpute`. Both return an expanded dataset to which the natural effect model can be fitted using the central function `neModel` (see Figure 4.3). In the next two sections, we explain both approaches and give example code in R.

4.3.2 Weighting-based approach

One way to account for missingness in the expanded data is to standardize observed outcomes to the mediator distribution of the hypothetical exposure level a'. Building on Hong's (2010) ratio-of-mediator-probability

i	A_i	a	a'	$Y_i(a, M_i(a'))$	w_i
1	1	1	1	Y_1	1
1	1	1	**0**	Y_1	$\hat{p}_1(\mathbf{0})/\hat{p}_1(1)$
2	0	0	0	Y_2	1
2	0	0	**1**	Y_2	$\hat{p}_2(\mathbf{1})/\hat{p}_2(0)$
⋮	⋮	⋮	⋮	⋮	⋮

Table 4.2: Schematic display of the weighting-based approach.

weighting (RMPW) method, Lange et al. (2012) proposed to weight each observation in the expanded dataset by

$$w_i = \frac{p_i(a')}{p_i(a)} = \frac{P(M = M_i|A = a', C = C_i)}{P(M = M_i|A = a, C = C_i)}.$$

For instance, for a binary exposure, $E\{Y(0, M(0))|C\}$ and $E\{Y(1, M(1))|C\}$ can readily be estimated from the observed data (under assumption (A1)) without weighting (i.e. as $a = a'$ the corresponding weights equal 1). To enable estimation of $E\{Y(1, M(0))|C\}$ and $E\{Y(0, M(1))|C\}$ RMPW aims to construct a 'parallel' pseudo-population for each exposure group a (within each stratum of C) with mediator values that would have been observed if each subject had been a member of the opposite exposure group $a' = 1 - a$. This is done by up-weighting individuals whose observed mediator value is more typical for the opposite exposure group than the exposure group to which they originally belong. Similarly, individuals whose observed mediator value is relatively more typical for the original exposure group are down-weighted.[3]

Data expansion hence only requires a' to take on values different from the observed exposure to enable estimation of natural direct and indirect effects via the weighting-based approach, as illustrated in Table 4.2. Estimates can then be obtained by regressing the observed outcome on a, a' and

[3] Hong et al. (2015) gives a more detailed example which may provide more intuition into RMPW.

baseline covariates C, weighting each observation in the expanded dataset by its corresponding ratio-of-mediator-probability weight. This procedure easily extends to continuous exposures (see section 4.4.2) and/or mediators (provided probabilities are replaced by densities). The interested reader is referred to Technical appendix 4.A.1, where we illustrate that the corresponding estimator is the stratum-specific analog of the RMPW estimator discussed in section 3.4.2 of chapter 3.

Expanding the data and computing weights for the natural effect model

Using the medflex package, expanding the dataset and calculating weights can be done in a single run, using the `neWeight` function. To calculate the weights, a model for the mediator needs to be fitted. For instance, in R, the simple linear model

$$E(M|A,C) = \alpha_0 + \alpha_1 A + \alpha_2 C,$$

can be fitted using the `glm` function:

```
medFit <- glm(negaff ~ factor(attbin) + gender + educ + age,
  family = gaussian, data = UPBdata)
```

Next, this fitted object needs to be specified as the first argument in `neWeight`, which in turn codes the first predictor variable in the `formula` argument as the exposure and then expands the data along hypothetical values of this variable. It is important to note here that, for successful data expansion, categorical exposures should be explicitly coded as factors in the `formula` if they are not yet coded as such in the dataset.

```
expData <- neWeight(medFit)
```

Inspecting the first rows of the resulting expanded dataset shows that for each individual two replications have been created.

```
head(expData, 4)
```

```
##   id attbin0 attbin1    att attcat negaff initiator gender educ age UPB
## 1  1       1       1  1.001      M   0.84    myself      F    M  41   1
## 2  1       1       0  1.001      M   0.84    myself      F    M  41   1
## 3  2       0       0 -0.709      L  -1.26      both      M    M  42   0
## 4  2       0       1 -0.709      L  -1.26      both      M    M  42   0
```


The new variables `attbin0` and `attbin1` correspond to hypothetical exposure values a and a', respectively. By convention, the index '0' is used for parameters (and corresponding auxiliary variables) indexing natural direct effects, whereas the index '1' is used for parameters indexing natural indirect effects in the natural effect model.

To shorten code, one can instead choose to directly specify the `formula`, `family` and `data` arguments in `neWeight`.

```
expData <- neWeight(negaff ~ factor(attbin) + gender + educ + age,
  data = UPBdata)
```

By default, `glm` is used as internal model-fitting function. However, other model-fitting functions can be specified in the `FUN` argument (e.g. `vglm` from the VGAM package; Yee, 2015).[4]

Finally, the weights are stored as an attribute of the expanded dataset and can easily be retrieved using the generic `weights` function, e.g. for further inspection of their empirical distribution:

```
w <- weights(expData)
head(w, 10)
```

```
##  [1] 1.000 0.640 1.000 0.494 1.000 0.475 1.000 1.211 1.000 0.326
```

[4] In the current version of the package also `vglm` and `vgam` from the VGAM package and `gam` from the gam package (Hastie, 2015) are supported. When specifying model-fitting functions other than `glm` in the `FUN` argument, one might need to specify the `family` argument differently. That is, in a way that is consistent with argument specification of that specific model-fitting function.

Fitting the natural effect model on the expanded data

After expanding the data and calculating regression weights for each of the replicates, the natural effect model can be fitted using the `neModel` function. Argument specification for this function is similar to that of the `glm` function, which is called internally. However, the `formula` argument now must be specified in function of the variables from the expanded dataset. The latter, in turn, needs to be specified via the `expData` argument. `neModel` automatically extracts the regression weights from this expanded dataset and applies them for model fitting.

Default `glm` standard errors tend to be downwardly biased as the uncertainty inherent to prediction of the weights based on the estimated mediator model is not taken into account. For this reason, `neModel` returns bootstrapped standard errors. In order to approximate the sampling distribution of each of the natural effect model parameters, the applied nonparametric bootstrap procedure repeatedly resamples the original data with replacement. For each replication, all aforementioned steps are repeated and estimates of the natural effect model parameters are obtained. The resulting bootstrap distribution can then be used for statistical inference. By refitting the same model for the mediator distribution to each bootstrap sample and recalculating ratio-of-mediator-probability weights for the (subsequently) expanded bootstrap samples, uncertainty related to estimation of the mediator model is incorporated into the bootstrapped standard errors. The number of bootstrap replications defaults to 1000 and can be set in the `nBoot` argument:

```
set.seed(1234)
neMod1 <- neModel(UPB ~ attbin0 + attbin1 + gender + educ + age,
  family = binomial("logit"), expData = expData)
```

The `summary` table of the resulting natural effect model object provides these bootstrap standard errors along with corresponding Wald-type z statistics and p values.

```
summary(neMod1)
```

```
## Natural effect model
## with standard errors based on the non-parametric bootstrap
## ---
## Exposure: attbin
## Mediator(s): negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.92521    0.91389   -1.01    0.311
## attbin01     0.39592    0.23283    1.70    0.089 .
## attbin11     0.35197    0.08829    3.99  6.7e-05 ***
## genderM      0.27597    0.23954    1.15    0.249
## educM        0.16701    0.75404    0.22    0.825
## educH        0.42335    0.75101    0.56    0.573
## age         -0.00945    0.01283   -0.74    0.461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As an alternative, robust standard errors based on the sandwich estimator (Liang and Zeger, 1986) can be requested by setting `se = "robust"`. Calculation of these standard errors is less computer-intensive and is available for natural effect models with working models fitted via the `glm` function. Technical details on this variance estimator can be found in Technical appendix 4.A.2.

```
neMod1 <- neModel(UPB ~ attbin0 + attbin1 + gender + educ + age,
  family = binomial("logit"), expData = expData, se = "robust")
summary(neMod1)
```

```
## Natural effect model
## with robust standard errors based on the sandwich estimator
## ---
```

```
## Exposure: attbin
## Mediator(s): negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.92521    0.71463   -1.29    0.195
## attbin01     0.39592    0.21761    1.82    0.069 .
## attbin11     0.35197    0.08939    3.94  8.2e-05 ***
## genderM      0.27597    0.23370    1.18    0.238
## educM        0.16701    0.50065    0.33    0.739
## educH        0.42335    0.50917    0.83    0.406
## age         -0.00945    0.01227   -0.77    0.441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Interpreting model parameters

Exponentiating the model parameter estimates provides estimates that can be interpreted as odds ratios. For instance, for a subject with baseline covariate levels C, altering the level of anxious attachment from low (=0) to high (=1), while controlling negative affectivity at levels as naturally observed at any given level of anxious attachment a, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}^{\text{NDE}}_{1,0|C} = \frac{\text{odds}\{Y(1, M(a)) = 1|C\}}{\text{odds}\{Y(0, M(a)) = 1|C\}} = \exp(\hat{\beta}_1) = 1.49.$$

Altering levels of negative affectivity as observed at low anxious attachment scores to levels that would have been observed at high anxious attachment scores, while controlling their anxious attachment score at any given level a, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}^{\text{NIE}}_{1,0|C} = \frac{\text{odds}\{Y(a, M(1)) = 1|C\}}{\text{odds}\{Y(a, M(0)) = 1|C\}} = \exp(\hat{\beta}_2) = 1.42.$$

Wald-type confidence intervals can be obtained by applying the `confint` function to the natural effect model object. The confidence level defaults to 95%, but can be changed via the `level` argument. By exponentiating the intervals on the logit scale, we can obtain the corresponding 95% confidence intervals (based on the robust standard errors) on the odds ratio scale:

```
exp(confint(neMod1)[c("attbin01", "attbin11"), ])
```

```
##           95% LCL 95% UCL
## attbin01    0.97    2.28
## attbin11    1.19    1.69
```


If standard errors are obtained via the bootstrap procedure, bootstrap confidence intervals are returned. The default type is calculated based on a first order normal approximation (`type = "norm"`), but other types of bootstrap confidence intervals (such as basic bootstrap, bootstrap percentile and bias-corrected and accelerated confidence intervals) can be obtained by setting the `type` argument to the desired type.[5]

4.3.3 Imputation-based approach

The second approach avoids reliance on a model for the mediator distribution and instead requires fitting a working model for the outcome mean (Vansteelandt et al., 2012b). By setting a' (rather than a) equal to the observed exposure A, unobserved nested counterfactuals can be imputed using any appropriate model for the outcome mean. That is, since the potential intermediate outcome $M(a')$ equals the observed mediator M within the subgroup with exposure $A = a'$, $Y(a, M(a'))$ equals $Y(a, M)$ for all individuals in that exposure group. The latter can then be imputed using fitted values $\hat{E}(Y|A = a, M, C)$ based on an appropriate model for the outcome mean, henceforth referred to as the imputation model, with exposure A set to a and with mediator M and baseline covariates C set to their observed values. This approach easily accommodates missing outcomes in the orig-

[5]The `type` argument in `confint` corresponds to that of the `boot.ci` function from the `boot` package (Canty and Ripley, 2015), which is called internally.

inal dataset, as the corresponding nested counterfactuals can likewise be imputed.

In contrast to the weighting-based approach, data expansion only requires a to take on values different from the observed exposure to enable estimation of natural direct and indirect effects, as illustrated in Table 4.3. Estimates can finally be obtained upon fitting a natural effect model to the imputed dataset. For ease of implementation, observed nested counterfactuals are imputed as well in the medflex package.[6] Again, the interested reader is referred to Technical appendix 4.A.1, where we illustrate that the corresponding estimator is the stratum-specific analog of the imputation estimator discussed in section 3.4.2 of chapter 3.

i	A_i	a	a'	$Y_i(a, M_i(a'))$
1	1	1	1	Y_1
1	1	**0**	1	$\hat{Y}_1(\mathbf{0}, M_1)$
2	0	0	0	Y_2
2	0	**1**	0	$\hat{Y}_2(\mathbf{1}, M_2)$
⋮	⋮	⋮	⋮	⋮

Table 4.3: Schematic display of the imputation-based approach. $\hat{Y}_i(a, M_i)$ represent the imputed counterfactual outcomes.

Expanding the data and imputing nested counterfactuals

Although application of the imputation-based approach is similar to that of the weighting-based approach, it differs in some key respects. These differences are mainly captured by differences between the functions `neWeight` and `neImpute`. Argument specification of this function is identical to that of `neWeight`, unless indicated otherwise.

As for the weighted-based approach, the first step amounts to fitting a working model. Instead of a model for the mediator, the imputation-based

[6] Simulation studies (not shown here) have shown that this procedure does not lead to bias or loss of efficiency.

approach requires fitting a mean model for the outcome. Moreover, this model should at least reflect the structure of model (4.1) (i.e. it should at least contain all terms of this natural effect model with a' replaced by M). For instance, a simple logistic regression model

$$\text{logit}P(Y = 1|A, M, C) = \gamma_0 + \gamma_1 A + \gamma_2 M + \gamma_3 C,$$

can be fitted in R using the `glm` function:


```
impFit <- glm(UPB ~ factor(attbin) + negaff + gender + educ + age,
  family = binomial("logit"), data = UPBdata)
```

In order for `neImpute` to identify the predictor variables in the `formula` argument correctly as either exposure, mediator(s) or baseline covariates, they need to be entered in a particular order. That is, the first predictor variable again needs to point to the exposure and the second to the mediator. All other predictors are automatically coded as baseline covariates. It is important to adhere to this prespecified order to enable `neImpute` to create valid pointers to these different types of predictor variables. This requirement extends to the use of operators different from the `+` operator, such as the `:` and `*` operators (when e.g. adding interaction terms). For instance, the formula expressions below all impose the same structural form for the imputation model.

```
Y ~ A + M + C1 + C2 + A:C1 + M:C1
Y ~ A + M + A:C1 + M:C1 + C1 + C2
Y ~ (A + M) * C1 + C2
Y ~ A * C1 + M * C1 + C2
```

However, only for the former three expressions, correct pointers to exposure, mediator and baseline covariates will be created, as the order of occurence of each of the unique predictor variables is identical in all three specifications, but not in the latter.

This fitted object then needs to be entered as the first argument in `neImpute`:

```
expData <- neImpute(impFit)
```

Alternatively, the `formula`, `family` and `data` arguments can be directly specified in `neImpute`:

```
expData <- neImpute(UPB ~ factor(attbin) + negaff + gender + educ + age,
  family = binomial("logit"), data = UPBdata)
```


Similar to `neWeight`, `neImpute` first expands the data along hypothetical exposure values. Instead of calculating weights for these new observations, `neImpute` then imputes the nested counterfactual outcomes by fitted values based on the imputation model. As illustrated below, the resulting expanded dataset includes two imputed nested counterfactual outcomes for each subject. The outcomes are no longer binary, but are substituted by conditional mean imputations.

```
head(expData, 4)
```

```
##   id attbin0 attbin1    att attcat negaff initiator gender educ age   UPB
## 1  1       1       1  1.001      M   0.84    myself      F    M  41 0.492
## 2  1       0       1  1.001      M   0.84    myself      F    M  41 0.384
## 3  2       0       0 -0.709      L  -1.26      both      M    M  42 0.187
## 4  2       1       0 -0.709      L  -1.26      both      M    M  42 0.263
```

Fitting the natural effect model on the imputed data

After expanding and imputing the data, specifying the natural effect model can be done as for the weighting-based approach:

```
neMod1 <- neModel(UPB ~ attbin0 + attbin1 + gender + educ + age,
  family = binomial("logit"), expData = expData, se = "robust")
```

Again, bootstrap or robust standard errors are reported in the output of the `summary` function, in order to account for the uncertainty inherent to the working model (i.e. in this case, the imputation model):

```
summary(neMod1)

## Natural effect model
## with robust standard errors based on the sandwich estimator
## ---
## Exposure: attbin
## Mediator(s): negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9216     0.6892   -1.34     0.18
## attbin01      0.4015     0.2134    1.88     0.06 .
## attbin11      0.3407     0.0805    4.23  2.3e-05 ***
## genderM       0.2940     0.2250    1.31     0.19
## educM         0.3462     0.4817    0.72     0.47
## educH         0.5143     0.4878    1.05     0.29
## age          -0.0122     0.0119   -1.02     0.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Natural direct and indirect effect odds ratio estimates and their confidence intervals can be obtained as before.

4.4 Dealing with different types of variables

In the previous section, we used a dichotomized version of the continuous exposure variable `att`. However, the natural effect model framework easily extends to different types of exposure, mediator or outcome variables. In the following two sections, we give a detailed description on how to fit natural effect models with multicategorical (i.e. ordinal or nominal) and continuous exposures. In these sections, as well as throughout the remainder of this chapter, we will focus on the imputation-based approach when introducing new features of the medflex package. Unless indicated otherwise, the weighting-based approach can be applied analogously.

	Outcome type					
	Binary		Count		Continuous	
Mediator type	`neWeight`	`neImpute`	`neWeight`	`neImpute`	`neWeight`	`neImpute`
Binary	✓	✓	✓	✓	✓	✓
Count	✓	✓	✓	✓	✓	✓
Continuous	✓	✓	✓	✓	✓	✓
Ordinal		✓		✓		✓
Nominal	✓*	✓	✓*	✓	✓*	✓

Table 4.4: Types of variables that can be dealt with in the medflex package. Natural effect models are currently restricted to models that can be fitted with the `glm` function. '*' indicates that robust standard errors are not available.

An overview of the types of mediators and outcomes the medflex package can currently handle, is given in Table 4.4. When using the weighting-based approach, models for binary, count and continuous mediators can be fitted using the `glm` function or the `vglm` function from the VGAM package. Models for nominal mediators, on the other hand, can only be fitted using the `vglm` function (setting `family = multinomial`).[7] Although models for ordinal mediators are not compatible with the `neWeight` function, ordered factors can easily be treated as nominal variables. Finally, the imputation-based approach can deal with virtually any type of mediator as it does not require the specification of a mediator model.

4.4.1 Multicategorical exposures

Methods for dealing with multicategorical treatments or exposures, as encountered in e.g. multiple intervention studies, in which multiple experimental conditions are compared to a control condition, have rarely been described within the mediation literature (although see Hayes and Preacher,

[7] In the current version of the package, when using working models for weighting (either when adopting the weighting-based approach or when fitting population-average natural effect models), robust standard errors are only available if these working models are fitted using `glm` and their outcomes (i.e. either an exposure or a mediator) follow either a normal, binomial or Poisson distribution.

2014; Tingley et al., 2014b, for some notable exceptions).

In this section, we illustrate how to expand the dataset and fit natural effect models when using a multicategorical exposure. In this example, instead of using the binary exposure variable `attbin`, we use a discretized version of anxious attachment style, named `attcat` (with `L` indicating low, `M` indicating intermediate and `H` indicating high anxious attachment levels).

Inspecting the first rows of the expanded dataset shows that the number of replications for each subject again corresponds to the number of unique levels of the categorical exposure variable. That is, the auxiliary variable a' (`attcat1`) is fixed to the observed exposure, whereas the other, a (`attcat0`), enumerates all potential exposure levels.


```
expData <- neImpute(UPB ~ attcat + negaff + gender + educ + age,
  family = binomial, data = UPBdata)
head(expData)
```

```
##   id attcat0 attcat1    att attbin negaff initiator gender educ age   UPB
## 1  1       M       M  1.001      1   0.84    myself      F    M  41 0.468
## 2  1       H       M  1.001      1   0.84    myself      F    M  41 0.558
## 3  1       L       M  1.001      1   0.84    myself      F    M  41 0.366
## 4  2       L       L -0.709      0  -1.26      both      M    M  42 0.182
## 5  2       M       L -0.709      0  -1.26      both      M    M  42 0.253
## 6  2       H       L -0.709      0  -1.26      both      M    M  42 0.327
```

The `summary` table returns estimates for the natural direct and indirect effect log odds ratios comparing intermediate and high anxious attachment levels to low levels of anxious attachment (i.e. the reference level).

```
neMod <- neModel(UPB ~ attcat0 + attcat1 + gender + educ + age,
  family = binomial, expData = expData, se = "robust")
summary(neMod)
```

```
## Natural effect model
## with robust standard errors based on the sandwich estimator
## ---
```

```
## Exposure: attcat
## Mediator(s): negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9616     0.6976   -1.38  0.16807
## attcat0M      0.3921     0.2365    1.66  0.09729 .
## attcat0H      0.7239     0.3105    2.33  0.01975 *
## attcat1M      0.3012     0.0797    3.78  0.00016 ***
## attcat1H      0.5218     0.1314    3.97  7.2e-05 ***
## genderM       0.2700     0.2266    1.19  0.23336
## educM         0.3279     0.4817    0.68  0.49601
## educH         0.4826     0.4877    0.99  0.32239
## age          -0.0127     0.0121   -1.05  0.29510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overall assessment of natural effects (i.e. a joint comparison of all levels of the exposure) cannot be based on the default `summary` output, but instead requires an Anova table for the natural effect model, which can be obtained using the `Anova` function from the car package (Fox and Weisberg, 2011):

```
library("car")
Anova(neMod)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: UPB
##         Df Chisq Pr(>Chisq)
## attcat0  2  5.98       0.05 .
## attcat1  2 19.11    7.1e-05 ***
## gender   1  1.42       0.23
## educ     2  1.17       0.56
## age      1  1.10       0.30
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both type-II (the default) and type-III Anova tables can be requested by specifying the desired type via the `type` argument. This table includes corresponding Wald χ^2 tests for multivariate hypotheses which account for the uncertainty inherent to the working model. The output suggests that the natural direct and indirect effect odds differ significantly between the three exposure levels.

4.4.2 Continuous exposures

In contrast to the mediation package, hypothesis testing for natural direct and indirect effects along the entire support of continuous exposures is facilitated by defining causal effects on their most natural scale. In this section, we use the continuous variable `att`, a standardized version of the original anxious attachment variable.

For continuous variables, expanding the dataset along unobserved (a, a') combinations requires a slightly adapted approach than for categorical exposures. Instead of enumerating all exposure levels to construct auxiliary variables a and a' for each subject, Vansteelandt et al. (2012b) proposed to draw specific quantiles from the conditional density of the exposure given baseline covariates. By default, these hypothetical exposure levels are drawn from a linear model for the exposure, conditional on a linear combination of all covariates specified in the working model.[8]

Both `neWeight` and `neImpute` allow to choose the number of draws to sample from this conditional density via the `nRep` argument (which defaults to 5).[9]

[8] If one wishes to use another model for the exposure, this default model specification can be overruled by referring to a fitted model object in the `xFit` argument. Misspecification of this sampling model does not induce bias in the estimated coefficients and standard errors of the natural effect model.

[9] We recommend to use a minimum of 3 draws. Although finite sample bias and sampling variability can be reduced to some extent by choosing a larger number of draws, simulations have shown this gain to be ignorable when choosing more than 5 draws (Vansteelandt et al., 2012b).

```
expData <- neImpute(UPB ~ att + negaff + gender + educ + age,
  family = binomial("logit"), data = UPBdata, nRep = 3)
head(expData)
```

```
##   id      att0   att1 attbin attcat negaff initiator gender educ age   UPB
## 1  1 -1.64e+00  1.001      1      M   0.84    myself      F    M  41 0.309
## 2  1  8.02e-06  1.001      1      M   0.84    myself      F    M  41 0.429
## 3  1  1.64e+00  1.001      1      M   0.84    myself      F    M  41 0.557
## 4  2 -1.66e+00 -0.709      0      L  -1.26      both      M    M  42 0.149
## 5  2 -1.82e-02 -0.709      0      L  -1.26      both      M    M  42 0.227
## 6  2  1.63e+00 -0.709      0      L  -1.26      both      M    M  42 0.330
```

Specification of the natural effect model via `neModel` can be done as described before:

```
neMod1 <- neModel(UPB ~ att0 + att1 + gender + educ + age,
  family = binomial("logit"), expData = expData, se = "robust")
summary(neMod1)
```

```
## Natural effect model
## with robust standard errors based on the sandwich estimator
## ---
## Exposure: att
## Mediator(s): negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4873     0.6862   -0.71   0.4776
## att0          0.2923     0.1091    2.68   0.0074 **
## att1          0.2018     0.0470    4.29  1.8e-05 ***
## genderM       0.2671     0.2274    1.17   0.2402
## educM         0.2679     0.4894    0.55   0.5841
## educH         0.4103     0.4959    0.83   0.4080
## age          -0.0120     0.0122   -0.99   0.3236
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output illustrates that defining natural effects on the (log) odds ratio scale allows to capture each of these effects along the entire support of the exposure by a single parameter. For instance, for a subject with baseline covariate levels C, the direct and indirect effects of one standard deviation increase in anxious attachment level (i.e. from a to $a+1$) correspond to an increase in the odds of displaying unwanted pursuit behaviors by a factor

$$\widehat{\text{OR}}^{\text{NDE}}_{a+1,a|C} = \frac{\text{odds}\left\{Y(a+1, M(a)) = 1|C\right\}}{\text{odds}\left\{Y(a, M(a)) = 1|C\right\}} = \exp(\hat{\beta}_1) = \exp(0.29) = 1.34,$$

and

$$\widehat{\text{OR}}^{\text{NIE}}_{a+1,a|C} = \frac{\text{odds}\left\{Y(a, M(a+1)) = 1|C\right\}}{\text{odds}\left\{Y(a, M(a)) = 1|C\right\}} = \exp(\hat{\beta}_2) = \exp(0.2) = 1.22,$$

respectively, regardless of the initial level a. Defining natural effects on the risk difference scale (as in the mediation package) would not have enabled to capture these by a single parameter along the entire support of the exposure, because of induced non-additivity (an artificial example illustrating this induced non-additivity is given in Figure 4 of Loeys et al., 2013).

Throughout the remainder of this chapter, we will continue to use the original continuous exposure variable, `att`.

4.5 Effect modification of natural effects

4.5.1 Exposure-mediator interactions

So far, the considered natural effect models reflected the assumption that exposure and mediator do not interact in their effect on the outcome (on the scale defined by the link function). In particular, the natural direct effect odds ratio

$$\text{OR}^{\text{NDE}}_{1,0|C}(a') = \frac{\text{odds}\left\{Y(1, M(a')) = 1|C\right\}}{\text{odds}\left\{Y(0, M(a')) = 1|C\right\}}$$

was postulated to be the same for each choice of mediator level $M(a')$, and hence for each choice of reference exposure level a', at which the mediator is evaluated. Similarly, the natural indirect effect odds ratio

$$\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(a) = \frac{\mathrm{odds}\left\{Y(a, M(1)) = 1|C\right\}}{\mathrm{odds}\left\{Y(a, M(0)) = 1|C\right\}}$$

was postulated to be constant across different choices of a at which the outcome is evaluated. In other words, the effects Robins and Greenland (1992) referred to as the *pure* direct effect, $\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0)$, and *total* direct effect, $\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(1)$, were assumed to be equal. Likewise, the *pure* indirect effect, $\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0)$, and *total* indirect effect, $\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(1)$, were assumed to be equal. However, in many studies, these assumptions may not be plausible.

As pointed out by VanderWeele (2013), total causal effects can be decomposed into a pure direct effect, a pure indirect effect and a mediated interactive effect. On an additive scale, the latter can be described as either the difference between total direct and pure direct effects or as the difference between total indirect and pure indirect effects. Similarly, the total effect odds ratio

$$\mathrm{OR}_{1,0|C} = \frac{\mathrm{odds}\left\{Y(1, M(1)) = 1|C\right\}}{\mathrm{odds}\left\{Y(0, M(0)) = 1|C\right\}}$$

can be expressed as the product

$$\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0) \times \mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0) \times \frac{\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(1)}{\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0)} = \mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0) \times \mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0) \times \frac{\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(1)}{\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0)}$$

of the pure direct and pure indirect effect odds ratios and the mediated interaction odds ratio. Rather than reflecting the *difference* between total and pure direct or indirect effects, the mediated interaction odds ratio corresponds to the *ratio* of total and pure direct or indirect effect odds ratios.

In a logistic natural effect model, testing for exposure-mediator interaction amounts to testing whether the mediated interaction odds ratio differs from 1, or equivalently, on the scale of the linear predictor, whether the

corresponding log odds ratio, β_3' in natural effect model

$$\text{logit}P\left\{Y(a, M(a')) = 1|C\right\} = \beta_0' + \beta_1' a + \beta_2' a' + \beta_3' aa' + \beta_4' C, \qquad (4.2)$$

differs from 0. When including this interaction term in the outcome model, β_1' and β_2' encode the pure direct and indirect effect log odds ratios, respectively.

When applying the imputation-based approach, the working model needs to at least reflect the structure of the final natural effect model (as has been pointed out in section 4.3.3). This requires the user to first (re)fit the imputation model accordingly. For instance, a minimal imputation model for natural effect model (4.2) would be the logistic regression model

$$\text{logit}P(Y = 1|A, M, C) = \gamma_0' + \gamma_1' A + \gamma_2' M + \gamma_3' AM + \gamma_4' C.$$

The output of the corresponding natural effect model object suggests there is no evidence for mediated interaction at the 5% significance level ($p = .0541$).

```
expData <- neImpute(UPB ~ att * negaff + gender + educ + age,
  family = binomial("logit"), data = UPBdata)
neMod2 <- neModel(UPB ~ att0 * att1 + gender + educ + age,
  family = binomial("logit"), expData = expData, se = "robust")
summary(neMod2)
```

```
## Natural effect model
## with robust standard errors based on the sandwich estimator
## ---
## Exposure: att
## Mediator(s): negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3949     0.6800   -0.58   0.5614
## att0          0.2950     0.1102    2.68   0.0074 **
```

```
## att1          0.1817     0.0467    3.90  9.8e-05 ***
## genderM       0.2815     0.2263    1.24   0.2135
## educM         0.1798     0.4857    0.37   0.7113
## educH         0.3105     0.4929    0.63   0.5287
## age          -0.0139     0.0122   -1.14   0.2545
## att0:att1     0.0698     0.0363    1.93   0.0541 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


4.5.2 Effect modification by baseline covariates

One might additionally wish to determine whether direct or indirect effects generalize across different strata of the population and across different conditions.

In our example, researchers might for instance investigate whether the extent to which the effect of anxious attachment level on engaging in UPBs is mediated through the experience of negative affectivity differs between men and women or between people with different education levels (Muller et al., 2005; Preacher et al., 2007). This moderated mediation hypothesis can be probed by allowing the conditional indirect effect, as indexed by β_2 in expression (4.1), to depend on gender, C_1, as expressed in model (4.3):

$$\text{logit}P\left\{Y(a, M(a')) = 1|C\right\} = \beta_0'' + \beta_1'' a + \beta_2'' a' + \beta_3'' a' C_1 + \beta_4'' C. \quad (4.3)$$

The amount of effect modification by gender in this model is then simply captured by β_3''.

```
impData <- neImpute(UPB ~ (att + negaff) * gender + educ + age,
  family = binomial("logit"), data = UPBdata)
neMod3 <- neModel(UPB ~ att0 + att1 * gender + educ + age,
  family = binomial("logit"), expData = impData, se = "robust")
summary(neMod3)
```

```
## Natural effect model
```

```
## with robust standard errors based on the sandwich estimator
## ---
## Exposure: att
## Mediator(s): negaff
## ---
## Parameter estimates:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.4731     0.6860   -0.69   0.4904
## att0           0.2850     0.1069    2.67   0.0077 **
## att1           0.1441     0.0583    2.47   0.0134 *
## genderM        0.2591     0.2278    1.14   0.2553
## educM          0.2718     0.4903    0.55   0.5793
## educH          0.4166     0.4975    0.84   0.4024
## age           -0.0123     0.0122   -1.00   0.3153
## att1:genderM   0.1598     0.1016    1.57   0.1156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


The output suggests that the natural indirect effect does not differ significantly between men and women ($p = 0.1156$).

In a similar way, researchers can gauge effect modification by education level. Suppose, for instance, that one wishes to test whether education level moderates both the direct and indirect effect. This can be done by fitting the natural effect model

$$\begin{aligned}\text{logit}P\left\{Y(a, M(a'))=1|C\right\} = \beta_0^\star + \beta_1^\star a + \beta_2^\star a' + \beta_3^\star a C_{2,1} + \beta_4^\star a C_{2,2} \\ + \beta_5^\star a' C_{2,1} + \beta_6^\star a' C_{2,2} + \beta_7^\star C, \qquad (4.4)\end{aligned}$$

with $C_{2,1}$ and $C_{2,2}$ dummy variables encoding the three education levels. Effect modification of the natural indirect (direct) effect by education level in model (4.4) is then captured by $\beta_5^\star$ and $\beta_6^\star$ ($\beta_3^\star$ and $\beta_4^\star$).

```
impData <- neImpute(UPB ~ (att + negaff) * educ + gender + age,
  family = binomial("logit"), data = UPBdata)
neMod4 <- neModel(UPB ~ (att0 + att1) * educ + gender + age,
  family = binomial("logit"), expData = impData, se = "robust")
```

Testing for moderation by a multicategorical variable calls for a multivariate test, which can again be obtained by requesting an Anova table for the natural effect model.

4.6 Tools for calculating and visualizing causal effect estimates

In this section, we highlight tools that can aid in calculating and visualizing specific causal effect estimates of interest. These tools might prove useful for gaining insight, especially for more complex models including interaction terms involving natural effect parameters.

4.6.1 Linear combinations of parameter estimates

Although effect estimates for e.g. the total causal effect can easily be obtained from the `summary` table of a natural effect model, its standard error and confidence interval cannot. To this end, the function `neLht`, which exploits the functionality of the `glht` function from the multcomp package (Hothorn et al., 2008) can be of use. This function enables the calculation of linear combinations of parameter estimates as well as their corresponding standard errors and confidence intervals based on the bootstrap or robust variance-covariance matrix of the natural effect model.

For instance, in model (4.2), the total direct and indirect effect can be expressed on the log odds scale as $\beta'_1 + \beta'_3$ and $\beta'_2 + \beta'_3$, respectively. Similarly, the total causal effect log odds ratio is captured by $\beta'_1 + \beta'_2 + \beta'_3$. As the argument for the linear function, `linfct`, needs to be specified in terms of one or more linear hypotheses, these effects can be specified as illustrated below:

```
lht <- neLht(neMod2, linfct = c("att0 + att0:att1 = 0",
  "att1 + att0:att1 = 0", "att0 + att1 + att0:att1 = 0"))
```

The corresponding odds ratios and their confidence intervals can be requested by exponentiating the coefficients and confidence intervals of the resulting object:

```
exp(cbind(coef(lht), confint(lht)))
```

```
##                               95% LCL 95% UCL
## att0 + att0:att1        1.44    1.15    1.80
## att1 + att0:att1        1.29    1.15    1.43
## att0 + att1 + att0:att1 1.73    1.39    2.15
```

Separate univariate tests for linear hypothesis objects can be requested using the `summary` function:

```
summary(lht)
```

```
## Linear hypotheses for natural effect models
## with standard errors based on the sandwich estimator
## ---
##                          Estimate Std. Error z value Pr(>|z|)
## att0 + att0:att1           0.3648     0.1145    3.19   0.0014 **
## att1 + att0:att1           0.2515     0.0553    4.55  5.4e-06 ***
## att0 + att1 + att0:att1    0.5465     0.1118    4.89  1.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Univariate p-values reported)
```

In contrast to the `summary` table for `glht` objects, which yields p values that are adjusted for multiple testing, tests returned by the `summary` function applied to `neLht` objects report unadjusted univariate tests. Adjusted tests can be obtained by setting `test = adjusted()` (for more details consult the help page of the `adjusted()` function from the multcomp package; Hothorn et al., 2008).

4.6.2 Effect decomposition

If interest is mainly focused on the natural effect parameters, the convenience function `neEffdecomp` can be used instead of `neLht`. This function automatically retains the natural effect estimates and generates a linear hypothesis object that reflects the most suitable effect decomposition:

```
effdecomp <- neEffdecomp(neMod2)
summary(effdecomp)

## Effect decomposition on the scale of the linear predictor
## with standard errors based on the sandwich estimator
## ---
## conditional on: gender, educ, age
## with x* = 0, x = 1
## ---
##                       Estimate Std. Error z value Pr(>|z|)
## pure direct effect      0.2950     0.1102    2.68   0.0074 **
## total direct effect     0.3648     0.1145    3.19   0.0014 **
## pure indirect effect    0.1817     0.0467    3.90  9.8e-05 ***
## total indirect effect   0.2515     0.0553    4.55  5.4e-06 ***
## total effect            0.5465     0.1118    4.89  1.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Univariate p-values reported)
```

By default, reference levels for the exposure, a and a', are chosen to be 1 and 0, respectively. If one wishes to evaluate causal effects at different reference levels (e.g. if the natural effect model allows for mediated interaction or if it includes quadratic or higher-order polynomial terms for the exposure), these can be specified as a vector of the form `c(a*,a)` via the `xRef` argument.

The output indicates that, for a subject with baseline covariate levels C, a standard deviation increase from the average level of anxious attachment (=0), increases the odds of displaying unwanted pursuit behaviors with a

factor

$$\widehat{\text{OR}}^{\text{NDE}}_{1,0|C}(0) = \frac{\text{odds}\{Y(1, M(0)) = 1|C\}}{\text{odds}\{Y(0, M(0)) = 1|C\}} = \exp(\hat{\beta}'_1) = 1.34$$

when controlling negative affectivity at levels as naturally observed at average anxious attachment levels, or with a factor

$$\widehat{\text{OR}}^{\text{NDE}}_{1,0|C}(1) = \frac{\text{odds}\{Y(1, M(1)) = 1|C\}}{\text{odds}\{Y(0, M(1)) = 1|C\}} = \exp(\hat{\beta}'_1 + \hat{\beta}'_3) = 1.44$$

when controlling negative affectivity at levels as naturally observed at anxious attachment levels one standard deviation above the average level.

On the other hand, altering negative affectivity from levels that would have been observed at average levels of anxious attachment to levels that would have been observed at attachment scores of one standard deviation higher, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}^{\text{NIE}}_{1,0|C}(0) = \frac{\text{odds}\{Y(0, M(1)) = 1|C\}}{\text{odds}\{Y(0, M(0)) = 1|C\}} = \exp(\hat{\beta}'_2) = 1.20$$

when controlling their anxious attachment level at the average, or with a factor

$$\widehat{\text{OR}}^{\text{NIE}}_{1,0|C}(1) = \frac{\text{odds}\{Y(1, M(1)) = 1|C\}}{\text{odds}\{Y(1, M(0)) = 1|C\}} = \exp(\hat{\beta}'_2 + \hat{\beta}'_3) = 1.29$$

when controlling their anxious attachment level one standard deviation above the average.

The total causal effect odds ratio can be expressed as the product of the pure direct and indirect effect odds ratios and the mediated interaction odds ratio: a standard deviation increase from the average level of anxious attachment approximately doubles the odds of displaying unwanted pursuit behaviors.

$$\widehat{\text{OR}}_{1,0|C} = \frac{\text{odds}\{Y(1, M(1)) = 1|C\}}{\text{odds}\{Y(0, M(0)) = 1|C\}} = \exp(\hat{\beta}'_1 + \hat{\beta}'_2 + \hat{\beta}'_3) = 1.73.$$

If the model includes terms reflecting effect modification by baseline covariates (e.g. as in model (4.3)), effect decomposition is by default evaluated at covariate levels that correspond to 0 for continuous covariates and to the reference level for categorical covariates coded as factors. However, for this type of models, it might often be insightful to evaluate natural effect components at different covariate levels than the default levels. This can be done via the `covLev` argument, which requires a vector including valid levels for modifier covariates specified in the natural effect model. An example of effect decomposition for women (`gender = "F"`, the default covariate level) and men (`gender = "M"`) in model (4.3) is given in the R code below.


```
neEffdecomp(neMod3)
```

```
## Effect decomposition on the scale of the linear predictor
## ---
## conditional on: gender = F, educ, age
## with x* = 0, x = 1
## ---
##                          Estimate
## natural direct effect       0.285
## natural indirect effect     0.144
## total effect                0.429
```

```
neEffdecomp(neMod3, covLev = c(gender = "M"))
```

```
## Effect decomposition on the scale of the linear predictor
## ---
## conditional on: gender = M, educ, age
## with x* = 0, x = 1
## ---
##                          Estimate
## natural direct effect       0.285
## natural indirect effect     0.304
## total effect                0.589
```

4.6.3 Global hypothesis tests

Wald tests considering all specified linear hypotheses jointly can be requested by specifying `test = Chisqtest()`. For instance, in model (4.4), instead of using the `Anova` function, one could also test for moderated mediation by the multicategorical baseline covariate education level via a global hypothesis test involving the relevant parameters $\beta_5^\star$ and $\beta_6^\star$.

```
modmed <- neLht(neMod4, linfct = c("att1:educM = 0", "att1:educH = 0"))
summary(modmed, test = Chisqtest())
```

```
## Global linear hypothesis test for natural effect models
## with standard errors based on the sandwich estimator
## ---
##   Chisq DF Pr(>Chisq)
## 1   5.2  2     0.0742
```


4.6.4 Visualizing effect estimates and their uncertainty

Finally, the generic `plot` function can be applied to linear hypothesis objects to visualize (linear combinations of) effect estimates and their uncertainty by means of confidence interval plots. To obtain estimates and confidence intervals on the odds ratio scale, one can specify `transf = exp` in order to exponentiate the original parameter estimates (on the log odds ratio scale).

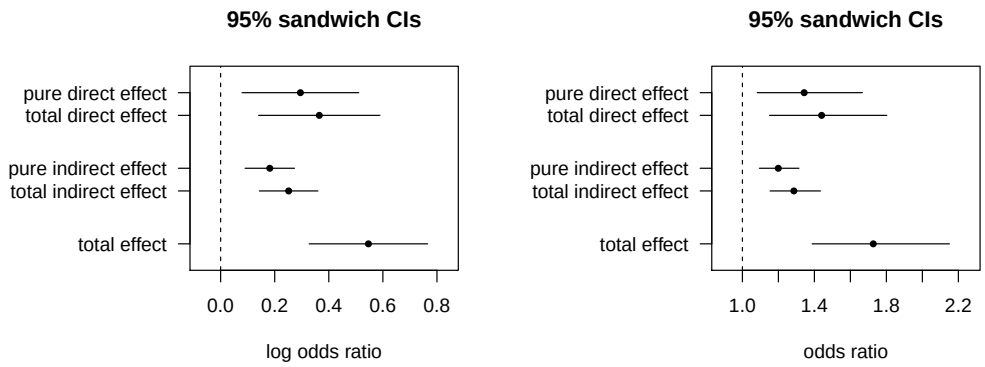


Figure 4.4: Effect decomposition on the log odds ratio and odds ratio scales.

Applying the `plot` function to a natural effect model object automatically retains the causal effect estimates of interest, generates a linear hypothesis object using `neEffdecomp` and then plots its corresponding estimates and confidence intervals, as shown in Figure 4.4. The default exposure reference and covariate levels for these plots are the same as for the `neEffdecomp` function, but can again be altered via the corresponding arguments `xRef` and `covLev`.

4.7 Population-average natural effects

In all previous sections, we defined natural effects as conditional or stratum-specific effects (i.e. conditional on baseline covariates). However, the medflex package additionally allows to estimate population-average natural effects. As demonstrated in section 3.4.2 of chapter 3, rewriting the mediation formula reveals that estimation of these population-average effects requires weighting by the reciprocal of the conditional exposure distribution in order to adjust for confounding (also see Albert, 2012; Vansteelandt, 2012).

As a consequence, a model for the exposure density needs to be fitted and specified as an additional working model, e.g.

```
expFit <- glm(att ~ gender + educ + age, data = UPBdata)
```

Since specifying population-average natural effect models using the `neModel` is equivalent for the weighting- and imputation-based approaches, in the remainder of this section, we demonstrate how to proceed when adhering to the imputation-based approach. Moreover, when estimating population-average natural effects, incoherence between imputation and natural effect models is less of a concern as the latter does not require modeling the relation between outcome and covariates. The (first) working model can again be fitted using the same commands as before:

```
impData <- neImpute(UPB ~ att + negaff + gender + educ + age,
  family = binomial("logit"), data = UPBdata)
```

Each observation in the expanded dataset to which the marginal natural effect model

$$\text{logit}P\left\{Y(a, M(a')) = 1\right\} = \theta_0 + \theta_1 a + \theta_2 a' \tag{4.5}$$

is fitted, needs to be weighted by the reciprocal of the exposure probability density, $P(A|C)$, evaluated at the observed exposure. The fitted model object that is used to calculate regression weights needs to be specified in the `xFit` argument of the `neModel` function:


```
neMod5 <- neModel(UPB ~ att0 + att1, family = binomial("logit"),
  expData = impData, xFit = expFit, se = "robust")
summary(neMod5)
```

```
## Natural effect model
## with robust standard errors based on the sandwich estimator
## ---
## Exposure: att
## Mediator(s): negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5793     0.1112   -5.21  1.9e-07 ***
## att0          0.2967     0.1082    2.74   0.0061 **
## att1          0.2294     0.0578    3.97  7.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both the marginal natural direct and indirect effect odds ratios again seem to be significantly different from 1: increasing the anxious attachment level from average to one standard error above average, while keeping negative affectivity fixed at levels corresponding to anxious attachment level a', increases the odds of displaying unwanted pursuit behaviors with

a factor

$$\widehat{\text{OR}}_{1,0}^{\text{NDE}} = \frac{\text{odds}\{Y(1, M(a')) = 1\}}{\text{odds}\{Y(0, M(a')) = 1\}} = \exp(\hat{\theta}_1) = 1.35.$$

A similar interpretation can again be made for the natural indirect effect.

4.8 Intermediate confounding: a joint mediation approach

In many settings multiple mediators may be of interest. In our example, one could argue that being anxiously attached to one's partner makes respondents more hesitant to end their relationship and that, in turn, not having initiated the break-up causes them to engage in unwanted pursuit behaviors more often. Initiator status (`initiator`: either `"both"`, `"ex-partner"`, or `"myself"`) can thus also be considered a mediator, which we denote L.

If hypothesized mediators are conditionally independent (given exposure and baseline covariates), separate natural effect models can be fitted (each with a different working model involving only one of the mediators) to assess the mediated effects through each of the mediators one at a time. Specifically, if the aforementioned ignorability conditions in assumptions (A1)-(A4) hold with respect to each mediator separately[10], natural indirect effects, as defined as causal pathways through single mediators, are identified since these conditions imply that the given mediators are independent given exposure and baseline covariates (Imai and Yamamoto, 2013; VanderWeele and Vansteelandt, 2013). Recently, Lange et al. (2014) demonstrated how independent intermediate pathways can be assessed in a single natural effect model using the weighting-based approach. Additionally, these authors proposed a regression-based approach for testing conditional dependence between mediators (also see Loeys et al., 2013; Imai and Yamamoto, 2013).

Often, however, mediators are interdependent and can be thought of as being linked in a sequential causal chain. For instance, not having initi-

[10] In addition to assumptions (A1)-(A4), we additionally assume that $Y(a, l) \perp\!\!\!\perp A|C$ (for all levels of a and l), $L(a) \perp\!\!\!\perp A|C$ (for all levels of a), $Y(a, l) \perp\!\!\!\perp L|A = a, C$ (for all levels of a and l) and $Y(a, l) \perp\!\!\!\perp L(a')|C$ (for all levels of a, a' and l).

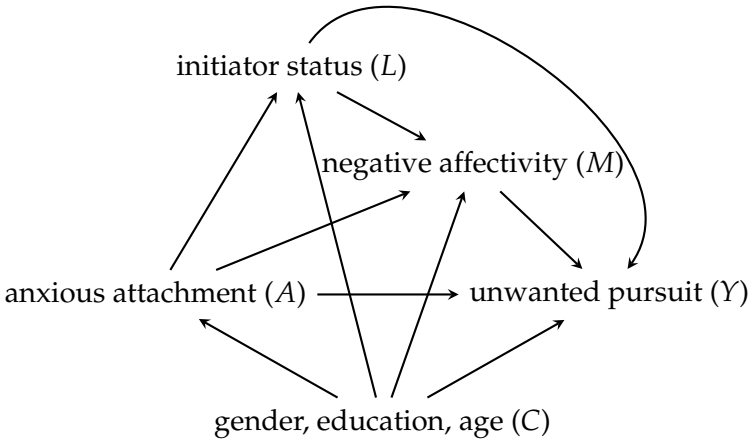


Figure 4.5: Causal diagram reflecting exposure-induced confounding.

ated the break-up could have made respondents more prone to feeling sad, jealous, angry, frustrated or hurt, as reflected in the causal diagram in Figure 4.5. Under this diagram, initiator status confounds the relation between the mediator and outcome (given that negative affectivity is the mediator of interest), while at the same time being affected by the exposure, hence violating assumption (A4). As a consequence, the natural indirect effect via negative affectivity is no longer identified under the NPSEM depicted in Figure 4.5 (although see Robins, 2003; Tchetgen Tchetgen and VanderWeele, 2014; Vansteelandt and VanderWeele, 2012, for additional (parametric) restrictions which enable identification). This non-identification can intuitively be appreciated by the fact that, in the presence of an intermediate confounder L, the natural indirect effect via M can be rewritten as

$$\frac{\text{odds}\left\{Y(a, L(a), M(1, L(1))) = 1 | C\right\}}{\text{odds}\left\{Y(a, L(a), M(0, L(0))) = 1 | C\right\}},$$

which involves blocking the causal path through L only ($A \rightarrow L \rightarrow Y$), while at the same time assessing the effect transmitted through L and M ($A \rightarrow L \rightarrow M \rightarrow Y$) (Didelez et al., 2006).

Alternatively, the total causal effect can be decomposed into the effect transmitted through L and M simultaneously and the effect not mediated

by any of the given mediators (VanderWeele and Vansteelandt, 2013; VanderWeele et al., 2014). Although such a joint mediation approach might not target the initial mediation hypothesis, it may still shed some light on the underlying causal mechanisms if there are reasons (either theoretical or empirical) to question the validity of assumption (A4) (with respect to a single mediator)[11], since this decomposition relies on a weaker set of ignorability assumptions. Specifically, if, as under the NPSEM depicted in Figure 4.5, we assume that a set of baseline covariates C satisfies 'no omitted confounders' assumptions (A1)-(A3) with respect to L and M jointly (rather than separately) and that no measured or unmeasured confounders of the $(L, M) - Y$ relation are affected by the exposure[12], the joint mediated effect and corresponding direct effect are identified. The appeal of this joint mediation approach is that by defining a natural indirect effect with respect to a set or vector of mediators (rather than a single mediator) assumption (A4) can be made more plausible by simply including mediator-outcome confounders that are deemed likely to be affected by the exposure in the joint set of mediators (VanderWeele and Vansteelandt, 2013).

For example, $\exp(\beta_1^{\star\star})$ in model (4.6)

$$\text{logit}P\left\{Y(a, L(a'), M(a', L(a'))) = 1|C\right\} = \beta_0^{\star\star} + \beta_1^{\star\star}a + \beta_2^{\star\star}a' + \beta_3^{\star\star}C, \tag{4.6}$$

captures the (newly defined) natural direct effect odds ratio

$$\text{OR}_{1,0|C}^{\text{NDE}} = \frac{\text{odds}\left\{Y(1, L(a'), M(a', L(a'))) = 1|C\right\}}{\text{odds}\left\{Y(0, L(a'), M(a', L(a'))) = 1|C\right\}},$$

whereas $\exp(\beta_2^{\star\star})$ captures the natural indirect effect odds ratio

$$\text{OR}_{1,0|C}^{\text{NIE}} = \frac{\text{odds}\left\{Y(a, L(1), M(1, L(1))) = 1|C\right\}}{\text{odds}\left\{Y(a, L(0), M(0, L(0))) = 1|C\right\}}$$

[11] In particular, it can be interesting to assess if the two mediators in combination lead to a null direct effect as this may signal that all important components in the causal chain from exposure to outcome have been identified.

[12] i.e. assuming that $Y(a, l, m) \perp\!\!\!\perp A|C$ (for all levels of a, l and m), $\{M(a), L(a)\} \perp\!\!\!\perp A|C$ (for all levels of a), $Y(a, l, m) \perp\!\!\!\perp \{L, M\}|A = a, C$ (for all levels of a, l and m) and $Y(a, l, m) \perp\!\!\!\perp \{L(a'), M(a')\}|C$ (for all levels of a, a', l and m).

through L and M jointly.

Fitting this natural effect model, however, requires both mediators to be taken into account in the working model(s). When applying the weighting-based approach, dealing with multiple mediators entails fitting a model for each of the mediators separately to calculate ratio-of-mediator probability weights, as in Lange et al. (2014). The imputation-based approach, on the other hand, is less demanding as it only requires one working model for the outcome. For this reason, estimation of joint mediated effects is implemented only for the imputation-based approach in the current version of the medflex package.

Hence, after expanding the data, nested counterfactual outcomes need to be imputed by fitted values from an imputation model conditional on both L and M. For instance, in the R code below, a logistic model

$$\text{logit}P(Y = 1|A, L, M, C) = \gamma_0^{\star\star} + \gamma_1^{\star\star}A + \gamma_2^{\star\star}L + \gamma_3^{\star\star}M + \gamma_4^{\star\star}LM + \gamma_5^{\star\star}C$$

is fitted that allows the mediators to interact in their effect on the outcome.

```
impData <- neImpute(UPB ~ att + initiator * negaff + gender + educ + age,
  family = binomial("logit"), nMed = 2, data = UPBdata)
```

The number of mediators to be considered jointly should be set via the `nMed` argument in the `neImpute` function. If `nMed = 2`, not only the second predictor variable, but the two predictor variables declared after the exposure variable are internally coded as mediators. Subsequently, natural effect model (4.6) can be fitted to the imputed dataset using the `neModel` function.

```
neMod6 <- neModel(UPB ~ att0 + att1 + gender + educ + age,
  family = binomial("logit"), expData = impData, se = "robust")
summary(neMod6)
```

```
## Natural effect model
## with robust standard errors based on the sandwich estimator
## ---
```

```
## Exposure: att
## Mediator(s): initiator, negaff
## ---
## Parameter estimates:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4919     0.6854   -0.72    0.473
## att0          0.2444     0.1114    2.19    0.028 *
## att1          0.2476     0.0538    4.60  4.2e-06 ***
## genderM       0.2629     0.2274    1.16    0.248
## educM         0.2780     0.4912    0.57    0.571
## educH         0.4223     0.4979    0.85    0.396
## age          -0.0121     0.0122   -0.99    0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Correct specification of the (number of) mediators can easily be checked in the `summary` output of the natural effect model object, which lists the names of the exposure and all mediators.

Although we have hypothesized that initiator status affects the level of experienced negative affectivity, this joint mediator approach does not necessarily require knowing the ordering of the mediators. VanderWeele and Vansteelandt (2013) and VanderWeele et al. (2014) described how additional insight into the causal mechanisms can be gained when the ordering is (assumed to be) known. These authors advocated a sequential approach which enables further effect decomposition of the total causal effect into multiple path-specific effects (Avin et al. 2005; also see Huber 2014 for an inverse-probability weighting approach and Albert and Nelson 2011 and Daniel et al. 2015 for a parametric g-computation approach for estimating some of these path-specific effects). Such sequential approach can easily be embedded in the natural effect model framework and is planned to be implemented in an upcoming version of the medflex package.

4.9 Weighting or imputing?

For both the weighting- and imputation-based approach, valid estimation of natural effects hinges on adequate specification of their corresponding nuisance working models and the natural effect model. In this section, we highlight the impact of model misspecification for each of the two proposed estimation approaches. The resulting trade-off in terms of modeling demands may serve as a guideline as to which of the two approaches is to be preferred in which particular setting. Moreover, certain missing data patterns might also favor one approach over the other, as discussed in more detail below.

4.9.1 Modeling demands

The proposed weighting-based approach yields consistent natural effects estimates if both the natural effect model and the conditional distribution of the mediator are correctly specified. The latter needs careful consideration, especially when exposure or baseline covariates are highly predictive of the mediator, for then even minor misspecifications in its conditional expectation can have a major impact on the weights and lead to heavily biased estimation of the target natural effects parameters. However, residual plots with scatterplot smoothers are often helpful to diagnose model inadequacy and can be requested, for instance, by passing the `expData`-class object to the `residualPlots` function from the car package. When dealing with continuous mediators, correct modeling not only demands adequate specification of the mediator's expectation, but also requires additional parametric assumptions on the mediator's conditional density (i.e. the distribution of the error terms).[13] Moreover, even under proper model specification, weights for continuous mediators typically tend to be unstable, leading to

[13] By default, the density function will correspond to the error distribution specified in the `family` argument for the mediator model (in turn specified via `neWeight`). QQ plots of the residuals can in this case be informative as to whether this parametric assumption is warranted for continuous mediators and can be obtained using the `qqnorm` function. The residuals can easily be obtained directly from the expanded dataset (as the working model is stored as an attribute in the expanded dataset object) by the command `residuals(expData)`.

less precise natural effect estimates and considerable finite sample bias. In particular, when the outcome is linear in the mediator, it might be sensible to avoid unnecessary parametric assumptions, since then the mediation formula prescribes only correct specification of the mediator's expectation.

In the light of these considerations, Vansteelandt et al. (2012b) recommended routine application of the imputation-based approach, especially when dealing with continuous mediators, since it avoids reliance on a model for the mediator. Despite this attraction, the imputation estimator does not come without limitations.

As in other imputation settings, one must pay due attention to coherent (or *congenial*) specification of the imputer's model and the analyst's model (i.e. in this case, the natural effect model) (Meng, 1994). This might be particularly challenging for nonlinear outcome models. For instance, when using logistic regression to model binary outcomes, the imputation model may be difficult or impossible to match with the natural effect model (VanderWeele and Vansteelandt, 2010; Tchetgen Tchetgen, 2014). To limit the impact of potential model uncongeniality in terms of misspecification bias, Vansteelandt et al. (2012b) and Loeys et al. (2013) advocated the use of a sufficiently rich imputation model.[14] To this end, the medflex package allows users to fit an imputation model using generalized additive models or machine learning techniques, such as the ensemble learner as implemented in the SuperLearner package (Polley and van der Laan, 2014).[15] Moreover, issues of uncongeniality can be avoided altogether by resorting to saturated natural effect models. In practice, models for conditional natural effects will rarely be saturated as either (some) baseline covariates or the exposure variable are continuous (or both). If the exposure is categorical, saturated models can be fitted for estimating population-average rather than stratum-specific natural effects (see section 4.7). However, for observational data, as op-

[14] A 'minimal' imputation model should thus at least reflect the structure of the natural effect model (e.g. also including exposure-mediator interactions when these are postulated as an aa' interaction in the natural effect model) to avoid attenuation of the estimates of effects that were precluded from the imputation model.

[15] An example is given in the help files of the package and can be consulted via `?neImpute.default`. Only bootstrap standard errors are available when fitting the imputation model using the `SuperLearner` function.

posed to data from experiments where the exposure is randomly assigned, adjustment for confounding in population-average natural effect models requires inverse weighting for the exposure.[16]

Second, as opposed to the weighting-based estimator, estimation by imputation requires modeling the mediator-outcome relation, which can be far from trivial whenever the exposure or baseline covariates are strongly associated with the mediator. In these scenarios, information about the effect of the mediator on the outcome may be sparse within certain strata defined by the exposure and covariates and, as a result, model misspecification may be difficult to diagnose and extrapolation bias becomes more likely (Vansteelandt, 2012). Whenever increased concerns of model extrapolation arise, the weighting-based approach may be indicated, as extrapolation uncertainty will typically be more honestly reflected in the corresponding standard errors (Vansteelandt et al., 2012a).[17]

Finally, it can be argued that, for both estimation approaches, if the working model is correctly specified (either via generalized linear models or via more advanced techniques), a parsimonious (and possibly misspecified) natural effect model will still provide some summary result tailored to answer the practitioner's main research questions (Vansteelandt et al., 2012b; Loeys et al., 2013). Suppose, for instance, that the logistic regression models in expression (0) are correctly specified. Fitting a natural effect model of the form

$$\text{logit}P(Y = 1|A, M, C) = \beta_0 + \beta_1 a + \beta_2 a' + \beta_3 C$$

to the expanded dataset using the imputation-based approach will then yield an estimated conditional natural indirect effect odds ratio of 1.143,

[16] Note that in both settings all baseline confounders still need to be adjusted for in the imputation model. Moreover, although the use of population-average natural effect models can, in some settings, avoid issues concerning potential model uncongeniality, it is up to the researcher to decide whether stratum-specific or population-average effects are the target of interest.

[17] Extrapolation might also affect estimation in the natural effect model, primarily when baseline covariates and exposure are highly correlated. This concern holds for both the weighting- and imputation-based estimator, since both require regression adjustment for covariates to estimate conditional natural effects.

which can be roughly considered as the mean conditional odds ratio across potential exposure levels (as depicted in Figure 4.1). If such an approach turns out to be unsatisfactory, users can again request residual plots to guide further model building and improve goodness-of-fit (by calling the `residualPlots` function). These diagnostics can be particularly helpful in the presence of certain non-linearities. For instance, when a continuous mediator is quadratic in the exposure, residual plots will indicate the need for a quadratic term for the indirect effect in the natural effect model, which will usually go unnoticed when fitting an imputation model for the outcome.

4.9.2 Missing data

As previously stated, when missingness occurs in the outcome, this is naturally dealt with when choosing the imputation-based approach, as missing outcomes in the original dataset are (by default) imputed in the expanded dataset, under the assumption that these outcomes are MAR (missing at random) given exposure, mediator(s) and baseline covariates.[18]

The weighting-based approach, on the other hand, is restricted to the analysis of complete cases and hence requires the more stringent MCAR (missing completely at random) assumption to hold in order to obtain unbiased estimation of the natural effect parameters. Whenever missingness occurs only in the outcome, we therefore advise to use the imputation-based approach. Alternatively, one might resort to multiple imputation, as also recommended if missingness occurs in either the exposure, mediator(s) or baseline covariates.

For instance, the `mice` function from the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011) can be used to obtain multiply imputed datasets (stored in a `mids`-class object). The working model can in turn be fitted to each of these datasets by passing them (or rather the object containing these datasets) to the `with.mids` function, which also processes the function (i.e. either `neWeight` or `neImpute`) and expression that needs

[18] It might be necessary to include additional covariates (that are both predictive of the outcome and missingness in the outcome, but are not included in the set of baseline covariates, C, that is chosen to meet assumptions (A1)-(A4)) in the imputation model to make the MAR assumption more plausible.

to be evaluated via the second argument. These steps are illustrated in the code below, in which `missdat` is a copy of the UPB dataset with artificially introduced missingness in each of the original variables.

```
library("mice")
library("mitools")

missdat <- UPBdata
for (i in 1:ncol(missdat))
  missdat[sample(nrow(missdat))[1:10], i] <- NA

multImp <- mice(missdat, m = 10)
expData <- with(multImp, neWeight(negaff ~ factor(attbin) + gender
  + educ + age))
```

Next, we use some functionalities from the mitools package (Lumley, 2014) to fit natural effect model (4.1) to each of the expanded multiply imputed datasets (stored in `expData$analyses`). The function `imputationList` can be used to transform the output containing these expanded datasets into a format that can be further passed to the `with.imputationList` function.

```
expData <- imputationList(expData$analyses)
neMod1 <- with(expData, neModel(UPB ~ attbin0 + attbin1 + gender
   + educ + age, family = binomial("logit"), se = "robust"))
```

Finally, the results can be pooled by using the `MIcombine` function.

```
MIcombine(neMod1)

## Multiple imputation results:
##       with(expData, neModel(UPB ~ attbin0 + attbin1 + gender + educ +
##     age, family = binomial("logit"), se = "robust"))
##       MIcombine.default(neMod1)
##                results     se
## (Intercept) -1.10743 0.7421
```

```
## attbin01     0.40150 0.2223
## attbin11     0.35112 0.0900
## gender2      0.26952 0.2374
## educ2        0.20981 0.5148
## educ3        0.36723 0.5299
## age         -0.00568 0.0125
```


4.10 Concluding remarks

In this chapter, we provided some theoretical background on the counterfactual framework, in particular on mediation analysis and natural direct and indirect effects, and described the functionalities of the R package medflex.

This package combines some important strengths of other (software) applications for mediation analysis that build on the mediation formula, while accommodating some of their respective weaknesses. The major appeal of this package is its flexibility in dealing with nonlinear parametric models and the functionalities it offers for hypothesis testing by resorting to natural effect models, which allow for direct parameterization of the target causal estimands on their most natural scale. Furthermore, for the most common parametric models, robust standard errors can be obtained, so the computer-intensive bootstrap can be avoided. A limitation of this package is that, at present, it does not offer any tools for assessing the sensitivity of one's results to possible violations of the identification assumptions of the causal estimands.

As mentioned in section 4.8, additional functionalities for dealing with exposure-induced confounding and multiple mediators are intended to be added to the package in the future, as well as extensions for survival models. Future developments within the realm of natural effect models (such as a generic framework for conducting sensitivity analyses) will be added in updates of the package.

4.A Technical appendices

4.A.1 Semi-parametric estimators

Weighting-based estimator

$$
\begin{aligned}
& E\left\{Y(a, M(a'))|C\right\} \\
&= \sum_m E(Y|A=a, M=m, C) P(M=m|A=a', C) \\
&= \sum_{y,m} y \cdot P(Y=y|A=a, M=m, C) P(M=m|A=a', C) \\
&= \sum_{y,m} y \cdot P(Y=y, M=m|A=a, C) \frac{P(M=m|A=a, C)}{P(M=m|A=a', C)} \\
&= E\left[Y \frac{P(M=m|A=a', C)}{P(M=m|A=a, C)} \,\middle|\, A=a, C \right]
\end{aligned}
$$

Imputation-based estimator

$$
\begin{aligned}
& E\left\{Y(a, M(a'))|C\right\} \\
&= \sum_m E(Y|A=a, M=m, C) P(M=m|A=a', C) \\
&= E\Big[E(Y|A=a, M, C) \,\Big|\, A=a', C \Big]
\end{aligned}
$$

4.A.2 Constructing sandwich estimators

In this section, we construct empirical sandwich estimators for the sampling variance of the stratum-specific analogs of the aforementioned semi-parametric estimators. Analytical expressions of sandwich estimators for population-average estimators are provided at
`https://cran.r-project.org/web/packages/medflex/vignettes/sandwich.pdf`.

Weighting-based estimator

Let $\mu_1(A, a', C; \beta)$ denote a natural effect model for $g\big[\mathrm{E}\{Y(A, M(a'))|C\}\big]$, $\mu_2(A, C; \theta)$ denote a nuisance working model for $g\big[\mathrm{E}(M|A, C)\big]$, and $g(\cdot)$ a canonical link function.

The stratum-specific weighting-based estimator then yields the following estimating equations:

$$U_1(A_i, a', C_i; \beta, \theta) = k^{-1} \sum_{j=1}^{k} \frac{\partial \mu_1(A_i, a'_{ij}, C_i; \beta)}{\partial \beta} \Sigma_{\mu_1}^{-1} \frac{P(M_i|a'_{ij}, C_i; \theta)}{P(M_i|A_i, C_i; \theta)} \cdot \left[Y_i - \mu_1(A_i, a'_{ij}, C_i; \beta)\right]$$

$$U_2(A_i, C_i; \theta) = \frac{\partial \mu_2(A_i, C_i; \theta)}{\partial \theta} \Sigma_{\mu_2}^{-1} \left[M_i - \mu_2(A_i, C_i; \theta)\right]$$

with k the number of replications or hypothetical values a' for each observation unit i and Σ_{μ_i} the residual variance-covariance matrix for model μ_i.

Let $\zeta = (\beta, \theta)$ and $\tilde{U} = (U_1, U_2)$. The sandwich estimator variance-covariance matrix can then be written as

$$n^{-1} \cdot \mathrm{E}\left(-\frac{\partial \tilde{U}}{\partial \zeta}\right)^{-1} \mathrm{Var}(\tilde{U}) \cdot \mathrm{E}\left(-\frac{\partial \tilde{U}}{\partial \zeta}\right)^{-T}$$

with n the total sample size of the original dataset,

$$\mathrm{E}\left(-\frac{\partial \tilde{U}}{\partial \zeta}\right)^{-1} = \begin{bmatrix} \mathrm{E}\left(-\frac{\partial U_1}{\partial \beta}\right) & \mathrm{E}\left(-\frac{\partial U_1}{\partial \theta}\right) \\ 0 & \mathrm{E}\left(-\frac{\partial U_2}{\partial \theta}\right) \end{bmatrix}^{-1} \text{ and}$$

$$\frac{\partial U_1}{\partial \beta} = -k^{-1} \sum_{j=1}^{k} \frac{\partial \mu_1(A_i, a'_{ij}, C_i; \beta)}{\partial \beta} \Sigma_{\mu_1}^{-1} \frac{P(M_i|a'_{ij}, C_i; \theta)}{P(M_i|A_i, C_i; \theta)} \frac{\partial \mu_1(A_i, a'_{ij}, C_i; \beta)}{\partial \beta}$$

$$\frac{\partial U_1}{\partial \theta} = k^{-1} \sum_{j=1}^{k} \frac{\partial \mu_1(A_i, a'_{ij}, C_i; \beta)}{\partial \beta} \Sigma_{\mu_1}^{-1} \frac{\partial}{\partial \theta} \left(\frac{P(M_i|a'_{ij}, C_i; \theta)}{P(M_i|A_i, C_i; \theta)}\right) \left[Y_i - \mu_1(A_i, a'_{ij}, C_i; \beta)\right]$$

$$\frac{\partial U_2}{\partial \theta} = -\frac{\partial \mu_2(A_i, C_i; \theta)}{\partial \theta} \Sigma_{\mu_2}^{-1} \frac{\partial \mu_2(A_i, C_i; \theta)}{\partial \theta}$$

Imputation-based estimator

Let $\mu_1(a, A, C; \beta)$ denote a natural effect model for $g\big[\mathrm{E}\{Y(a, M(A))|C\}\big]$, $\mu_2(A, M, C; \gamma)$ a nuisance working model for $g\big[\mathrm{E}(Y|A, M, C)\big]$, and $g(\cdot)$ a canonical link function.

The stratum-specific imputation-based estimator then yields the following estimating equations:

$$U_1(a, A_i, C_i; \beta, \gamma) = k^{-1} \sum_{j=1}^{k} \frac{\partial \mu_1(a_{ij}, A_i, C_i; \beta)}{\partial \beta} \Sigma_{\mu_1}^{-1} \cdot \big[\mu_2(a_{ij}, M_i, C_i; \gamma) - \mu_1(a_{ij}, A_i, C_i; \beta)\big]$$

$$U_2(A_i, M_i, C_i; \gamma) = \frac{\partial \mu_2(A_i, M_i, C_i; \gamma)}{\partial \gamma} \Sigma_{\mu_2}^{-1} \big[Y_i - \mu_2(A_i, M_i, C_i; \gamma)\big]$$

Let $\zeta = (\beta, \gamma)$ and $\tilde{U} = (U_1, U_2)$. The sandwich estimator variance-covariance matrix can then be written as

$$n^{-1} \cdot \mathrm{E}\left(-\frac{\partial \tilde{U}}{\partial \zeta}\right)^{-1} \mathrm{Var}(\tilde{U}) \cdot \mathrm{E}\left(-\frac{\partial \tilde{U}}{\partial \zeta}\right)^{-T}$$

with

$$\mathrm{E}\left(-\frac{\partial \tilde{U}}{\partial \zeta}\right)^{-1} = \begin{bmatrix} \mathrm{E}\left(-\frac{\partial U_1}{\partial \beta}\right) & \mathrm{E}\left(-\frac{\partial U_1}{\partial \gamma}\right) \\ 0 & \mathrm{E}\left(-\frac{\partial U_2}{\partial \gamma}\right) \end{bmatrix}^{-1}$$

and

$$\frac{\partial U_1}{\partial \beta} = -k^{-1} \sum_{j=1}^{k} \frac{\partial \mu_1(a_{ij}, A_i, C_i; \beta)}{\partial \beta} \Sigma_{\mu_1}^{-1} \frac{\partial \mu_1(a_{ij}, A_i, C_i; \beta)}{\partial \beta}$$

$$\frac{\partial U_1}{\partial \gamma} = k^{-1} \sum_{j=1}^{k} \frac{\partial \mu_1(a_{ij}, A_i, C_i; \beta)}{\partial \beta} \Sigma_{\mu_1}^{-1} \frac{\partial \mu_2(a_{ij}, M_i, C_i; \gamma)}{\partial \gamma}$$

$$\frac{\partial U_2}{\partial \gamma} = -\frac{\partial \mu_2(A_i, M_i, C_i; \beta)}{\partial \gamma} \Sigma_{\mu_2}^{-1} \frac{\partial \mu_2(A_i, M_i, C_i; \beta)}{\partial \gamma}$$

Chapter 5

Flexible mediation analysis with multiple mediators

This chapter is based on the following paper: Steen, J., Loeys, T., Moerkerke, B., Vansteelandt, S. (2016). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, in press.

The advent of counterfactual-based mediation analysis has triggered enormous progress on how, and under what assumptions, one may disentangle path-specific effects upon combining arbitrary (possibly nonlinear) models for mediator and outcome. However, current developments have largely focused on single mediators because required identification assumptions prohibit simple extensions to settings with multiple mediators that may depend on one another.

In this chapter, we propose a procedure for obtaining fine-grained decompositions that may still be recovered from the data in such complex settings. We first show that existing analytical approaches target specific instances of a more general set of decompositions and may therefore fail to provide a comprehensive assessment of the processes that underpin cause-effect relations between exposure and outcome. We then outline conditions for obtaining the remaining set of decompositions. Because the number of targeted decompositions increases rapidly with the number of mediators,

we introduce natural effect models along with estimation methods that allow for flexible and parsimonious modeling.

Our procedure can easily be implemented using off-the-shelf software and is illustrated in a re-analysis of the World Health Organization's Large Analysis and Review of European Housing and Health Status (WHO-LARES) study on the effect of mold exposure on mental health (2002-2003).

5.1 Introduction

Mediation analysis is widely conducted to deepen understanding of the mechanisms behind established cause-effect relationships. It does so by separating the indirect effect that operates through a given intermediate (or mediator) from the remaining direct effect and by quantifying their respective contributions to the overall exposure effect. Epidemiologists often focus on multiple mediators, either because interest lies in multiple mechanisms or because the association between the mediator of interest and the outcome is confounded by an earlier intermediate. However, as the number of definable causal pathways from exposure to outcome grows exponentially with an increasing number of mediators being considered, so does the complexity related to their identification and estimation (Daniel et al., 2015).

Although analyses with multiple mediators have a long tradition in the structural equation models (SEM) literature, complications related to effect decomposition have long been obscured as SEM-based definitions of path-specific effects rely on stringent parametric constraints (Taylor et al., 2007). Recent contributions building on the counterfactual framework have helped to unveil intricacies related to non-parametric identification of path-specific effects (Avin et al., 2005). Accordingly, counterfactual-based approaches to effect decomposition in the presence of causally ordered mediators have been put forward. These approaches have mainly illustrated that progress can be made either by incorporating sensitivity analyses to obtain the finest possible decomposition (Daniel et al., 2015; Albert and Nelson, 2011) or by focusing on coarser decompositions that require weaker assumptions (VanderWeele and Vansteelandt, 2013; VanderWeele et al., 2014).

In the current chapter, we extend this second line of research by proposing a simple estimation procedure for effect decomposition in the presence of causally ordered mediators. Such settings give rise to a large number of possible decompositions (Daniel et al., 2015). For instance, applications with only three sequential mediators already yield 24 possible ways of partitioning the total causal effect into path-specific effects that can be identified, under certain conditions, without imposing parametric restrictions. Existing approaches (VanderWeele and Vansteelandt, 2013) are limited as they recover only a subset of all such targeted decompositions. They may therefore give an incomplete assessment of the processes that underlie cause-effect relations, especially in the presence of interaction. The multitude of possible decompositions, however, calls for parsimonious modeling strategies. We therefore extend so-called natural effect models (Lange et al., 2012; Vansteelandt et al., 2012b), a class of marginal structural models for mediation analysis, along with accompanying fitting strategies. Besides parsimony, our procedure offers greater modeling flexibility than prevailing Monte Carlo approaches (Daniel et al., 2015; Albert and Nelson, 2011). For didactic purposes, we present our approach for two sequential mediators, although it easily extends to more mediators (see Technical appendices 5.A.1 and 5.A.2).

5.2 Effect decomposition into path-specific effects

5.2.1 Decomposition in a single mediator setting

Notation, definitions and identification

Within the counterfactual framework, causal effects are defined by comparing counterfactual outcomes under different exposure regimes. The total effect of a binary exposure ($A = 1$ for exposed, $A = 0$ for unexposed) on an outcome Y is obtained by contrasting $Y(1)$ and $Y(0)$, with $Y(a)$ the counterfactual outcome that would be observed if A were set, possibly contrary to the fact, to a. The population-average exposure effect can then be expressed in terms of mean differences $E\{Y(1) - Y(0)\}$, relative risks $P\{Y(1) = 1\}/P\{Y(0) = 1\}$, etc.

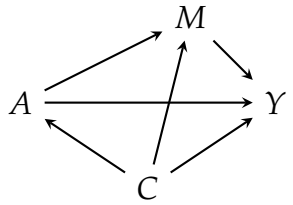


Figure 5.1: Causal diagram with a single mediator M.

Expressions for direct and mediated effects can similarly be obtained by invoking nested counterfactuals $Y(a, M(a'))$. For instance, one can isolate part of the effect that is transmitted by M by leaving the exposure unchanged at $A = 1$, but changing the mediator from $M(1)$, the natural value it would have taken under exposure, to $M(0)$, the value it would have taken under no exposure. Comparison of nested counterfactuals $Y(1, M(1))$ and $Y(1, M(0))$ is central to the definition of natural indirect effects (Pearl, 2001; Robins and Greenland, 1992). Definitions of natural direct effects can similarly be obtained by comparing $Y(1, M(0))$ and $Y(0, M(0))$. This contrast captures the intuitive notion of blocking the exposure's effect on the mediator by keeping the latter fixed at the level it would have taken in the absence of exposure.

Natural effects combine to produce the total effect, irrespective of the scale of interest or the presence of interactions or nonlinearities. For instance, on the additive scale, the total causal effect decomposes into the sum of the natural direct and indirect effect

$$\begin{aligned} E\{Y(1) - Y(0)\} = & E\{Y(1, M(0)) - Y(0, M(0))\} \\ & + E\{Y(1, M(1)) - Y(1, M(0))\}, \end{aligned}$$

given the composition assumption that $Y(a, M(a)) = Y(a)$.

Non-parametric identification of natural effects can be obtained under a set of sufficient conditions (VanderWeele and Vansteelandt, 2009), which state that for any value of a, a' and m

$$Y(a, m) \perp\!\!\!\perp A | C \tag{5.1}$$

$$Y(a, m) \perp\!\!\!\perp M | A = a, C \tag{5.2}$$

$$M(a) \perp\!\!\!\perp A|C \tag{5.3}$$
$$Y(a,m) \perp\!\!\!\perp M(a')|C, \tag{5.4}$$

where $U \perp\!\!\!\perp V|W$ denotes that U and V are independent conditional on W.

These conditions require a set of measured baseline covariates C that suffices to deconfound not only (i) the effect of exposure A on outcome Y and (ii) the effect of mediator M on outcome Y conditional on exposure A, as dictated in the SEM literature (Judd and Kenny, 1981), but also (iii) the effect of exposure A on mediator M. Assumption (5.4) is a strong assumption, commonly referred to as Pearl (2001)'s 'cross-world' independence assumption. If the data are assumed to be generated from a non-parametric structural equation model with independent errors (NPSEM) (Robins and Richardson, 2010), assumptions (5.1)-(5.4) can be shown to hold if, in addition to (i)-(iii), (iv) none of the mediator-outcome confounders are affected by exposure. In this chapter, we will further discuss identification conditions, such as (i)-(iv), as represented in causal diagrams (such as Figure 5.1) interpreted as NPSEMs.

Natural effect models

Natural direct and indirect effects can be parameterized by so-called natural effect models (Lange et al., 2012; Steen et al., 2016b; Vansteelandt et al., 2012b). These express the mean of nested counterfactuals in terms of hypothetical exposure levels a and a' and therefore naturally extend marginal structural models to allow for effect decomposition. For instance, in the following saturated model for a binary exposure A

$$E\{Y(a, M(a'))\} = \beta_0 + \beta_1 a + \beta_2 a' + \beta_3 a a', \tag{5.5}$$

for a, a' equal to 0 or 1, β_1 and $\beta_2 + \beta_3$ respectively capture the natural direct and indirect effect as expressed above, that is,

$$E\{Y(1, M(0)) - Y(0, M(0))\} = \beta_1,$$
$$E\{Y(1, M(1)) - Y(1, M(0))\} = \beta_2 + \beta_3.$$

This two-way decomposition of the total effect ($\beta_1 + \beta_2 + \beta_3$) into the so-called pure direct and total indirect effect is not unique (Robins and Greenland, 1992). A different decomposition into the so-called total direct and pure indirect effect arises from differently apportioning the interaction term β_3 as follows:

$$E\{Y(1, M(1)) - Y(0, M(1))\} = \beta_1 + \beta_3,$$
$$E\{Y(0, M(1)) - Y(0, M(0))\} = \beta_2.$$

Model (5.5) is a special case of the wider class of generalized linear natural effect models

$$E\{Y(a, M(a'))|C^*\} = g^{-1}\{\beta^\top W(a, a', C^*)\},$$

with $W(a, a', C^*)$ a known vector with components that may depend on a, a' and (possibly) a set of baseline covariates C^* (with $C^* \in C$), β an unknown parameter vector and link function $g(\cdot)$. In model (5.5), which encodes population-average rather than stratum-specific natural effects, C^* is the empty set, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$, $W(a, a', C^*) = (1, a, a', aa')^\top$, and $g(\cdot)$ is the identity link. The inclusion of a non-empty set C^* additionally enables parameterizing effect modification by baseline covariates.

5.2.2 Decomposition in a setting with two sequential mediators

In most mediation analyses, even when interest lies in a single mediator, one cannot ignore the possible presence of multiple mediators, as the following motivating example illustrates.

Motivating example

For illustrative purposes, we revisit previous analyses (VanderWeele and Vansteelandt, 2010; Vansteelandt et al., 2012b) on survey data from 5,882 adult respondents from the Large Analysis and Review of European Housing and Health Status (LARES) project conducted by the World Health Organization (Shenassa et al., 2007). These analyses focused on the effect

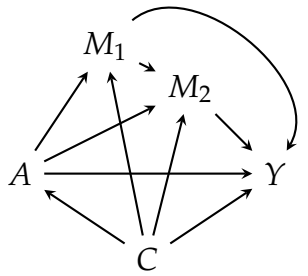


Figure 5.2: Causal diagram with two sequential mediators M_1 and M_2.

of living in damp and moldy conditions (binary exposure A) on the risk of depression (binary outcome Y) and put forward perceived control over one's home as a putative mediating mechanism (M). Corresponding natural direct and indirect effects (via perceived control) were estimated under the assumption that available individual and housing characteristics (C) were sufficient to control for confounding so that conditions (i)-(iii) were met (as reflected by the DAG in Figure 5.1). Kaufman (2010), however, indicated that mold exposure is likely to also cause physical illness, which may, in turn, compromise both one's sense of control and mental health. This hypothetical scenario (as reflected by the DAG in Figure 5.2) therefore violates assumption (iv) and thus hinders identification of the targeted natural effects discussed earlier. It moreover implies that both physical illness (M_1) and perceived control (M_2) act as sequential mediators, giving rise to a finest possible decomposition that involves four distinct pathways from exposure to outcome (i.e. pathways $A \to Y$, $A \to M_1 \to Y$, $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$).

In the remainder of this section, we first outline a sequential approach that bears close resemblance to VanderWeele and Vansteelandt (2013), starting from a coarse two-way decomposition which is next refined into a three-way decomposition. We then demonstrate how natural effect models can be extended to parameterize component effects of the resulting and alternative decompositions and articulate required identification conditions.

A sequential approach

Let $Y(a, M_1(a'), M_2(a', M_1(a')))$ be the counterfactual outcome that would be observed if A were set to a and M_1 and M_2 were set to the natural value they would have taken if A had been a'. The first stage then corresponds to a two-way decomposition with respect to the joint mediator $\{M_1, M_2\}$, separating pathway $A \to Y$ from the remaining pathways as follows:

$$
\begin{aligned}
&E\{Y(1) - Y(0)\} \\
&= E\{Y(1, M_1(1), M_2(1, M_1(1))) - Y(1, M_1(0), M_2(0, M_1(0)))\} \qquad (5.6) \\
&\quad + E\{Y(1, M_1(0), M_2(0, M_1(0))) - Y(0, M_1(0), M_2(0, M_1(0)))\}. \qquad (5.7)
\end{aligned}
$$

That is, the effect transmitted along either one or both mediators, or so-called joint natural indirect effect (expression (5.6)), is separated from the remaining effect through neither of the mediators, or the joint natural direct effect (expression (5.7)), denoted $E_{A\to Y}(0, 0)$ (see Table 5.1).

In a second stage, a more fine-grained, three-way decomposition can be obtained by further partitioning expression (5.6) into the entire effect transmitted along M_1 and the effect transmitted along M_2 only, respectively denoted $E_{A\to M_1 Y}(1, 1)$ and $E_{A\to M_2\to Y}(1, 0)$ (see Table 5.1):

$$
\begin{aligned}
&E\{Y(1, M_1(1), M_2(1, M_1(1))) - Y(1, M_1(0), M_2(0, M_1(0)))\} \\
&= E\{Y(1, M_1(1), M_2(1, M_1(1))) - Y(1, M_1(0), M_2(1, M_1(0)))\} \qquad (5.8) \\
&\quad + E\{Y(1, M_1(0), M_2(1, M_1(0))) - Y(1, M_1(0), M_2(0, M_1(0)))\}. \qquad (5.9)
\end{aligned}
$$

The first contrast (expression (5.8)) captures the notion of activating all paths along M_1 that feed into Y, either directly or indirectly via M_2, while blocking all other pathways. It corresponds to the natural indirect effect as defined with respect to M_1 (i.e. along the combined pathways $A \to M_1 \to Y$ and $A \to M_1 \to M_2 \to Y$), under the composition assumption that $Y(a, M_1(a'), M_2(a, M_1(a'))) = Y(a, M_1(a'))$. The second contrast (expression (5.9)) expresses the so-called semi-natural indirect effect (Pearl, 2014) or partial indirect effect (Huber, 2014) with respect to M_2 (i.e. $A \to M_2 \to Y$), as it only captures part of the effect mediated by M_2 that bypasses M_1.

$$
\begin{aligned}
E_{A\to Y}(a', a'') &= g\left(E\{Y(1, M_1(a'), M_2(a'', M_1(a')))\}\right) \\
&\qquad - g\left(E\{Y(0, M_1(a'), M_2(a'', M_1(a')))\}\right) \\
&= \theta_1 + \theta_4 a' + \theta_5 a'' + \theta_7 a' a'' \\
E_{A\to M_1 Y}(a, a'') &= g\left(E\{Y(a, M_1(1), M_2(a'', M_1(1)))\}\right) \\
&\qquad - g\left(E\{Y(a, M_1(0), M_2(a'', M_1(0)))\}\right) \\
&= \theta_2 + \theta_4 a + \theta_6 a'' + \theta_7 a a'' \\
E_{A\to M_2\to Y}(a, a') &= g\left(E\{Y(a, M_1(a'), M_2(1, M_1(a')))\}\right) \\
&\qquad - g\left(E\{Y(a, M_1(a'), M_2(0, M_1(a')))\}\right) \\
&= \theta_3 + \theta_5 a + \theta_6 a' + \theta_7 a a'
\end{aligned}
$$

Table 5.1: Shorthand notation for the component effects from a three-way decomposition in the presence of two causally ordered mediators M_1 and M_2 and their parameterization in model (5.10), for which the link function $g(\cdot)$ is the identity link.

Further decomposition will generally fail without imposing strong parametric constraints, as in the linear SEM framework (Avin et al., 2005) (although see Daniel et al. (2015) for a sensitivity analysis approach). Likewise, alternative decompositions of expression (5.6) that involve the natural indirect effect with respect to M_2 (instead of M_1; i.e. along the combined pathways $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$) cannot be recovered without making certain no-interaction assumptions (Huber, 2014; Imai and Yamamoto, 2013; Petersen et al., 2006; Robins, 2003; Tchetgen Tchetgen and VanderWeele, 2014). These decompositions are beyond the scope of this chapter (see Technical appendix 5.A.1 for a detailed overview and comparison of targeted decompositions).

Natural effect models

Natural effect models can be extended to characterize the three-way decomposition of the previous section. For instance, in the following saturated natural effect model for a binary exposure A

$$E\{Y(a, M_1(a'), M_2(a'', M_1(a')))\} = \theta_0 + \theta_1 a + \theta_2 a' + \theta_3 a''$$

$$+ \theta_4 aa' + \theta_5 aa'' + \theta_6 a'a'' + \theta_7 aa'a'', \tag{5.10}$$

for a, a' and a'' equal to 0 or 1, the total effect, $\sum_{i=1}^{7} \theta_i$, can be partitioned into the joint natural direct effect

$$E_{A\to Y}(0,0) = \theta_1,$$

and the joint natural indirect effect

$$E_{A\to M_1 Y}(1,1) + E_{A\to M_2\to Y}(1,0) = \sum_{i=2}^{7} \theta_i.$$

The latter can be further partitioned into the natural indirect effect with respect to M_1

$$E_{A\to M_1 Y}(1,1) = \theta_2 + \theta_4 + \theta_6 + \theta_7,$$

and the partial indirect effect with respect to M_2 (see Table 5.1)

$$E_{A\to M_2\to Y}(1,0) = \theta_3 + \theta_5.$$

Model (5.10) is a special case of the wider class of generalized linear natural effect models for three-way decomposition

$$E\{Y(a, M_1(a'), M_2(a'', M_1(a')))|C^*\} = g^{-1}\{\theta^\top W(a, a', a'', C^*)\},$$

with $W(a, a', a'', C^*)$ a known vector with components that may depend on a, a', a'' and (possibly) covariates C^*.

Different ways of accounting for the interaction terms θ_4 to θ_7 yield another five possible decompositions, listed in Table 5.2. For instance, θ_4 can be apportioned to either $E_{A\to Y}$ or $E_{A\to M_1 Y}$. Similarly, θ_5 can be apportioned to $E_{A\to Y}$ or $E_{A\to M_2\to Y}$, θ_6 to $E_{A\to M_1 Y}$ or $E_{A\to M_2\to Y}$ and θ_7 to either of the three components. VanderWeele and Vansteelandt (2013) focus only

$$
\begin{aligned}
(1)\quad & E_{A\to Y}(0,0) + E_{A\to M_1Y}(1,1) + E_{A\to M_2\to Y}(1,0) \\
& = (\theta_1) + (\theta_2 + \theta_4 + \theta_6 + \theta_7) + (\theta_3 + \theta_5) \\
(2)\quad & E_{A\to Y}(1,1) + E_{A\to M_1Y}(0,0) + E_{A\to M_2\to Y}(0,1) \\
& = (\theta_1 + \theta_4 + \theta_5 + \theta_7) + (\theta_2) + (\theta_3 + \theta_6) \\
(3)\quad & E_{A\to Y}(0,0) + E_{A\to M_1Y}(1,0) + E_{A\to M_2\to Y}(1,1) \\
& = (\theta_1) + (\theta_2 + \theta_4) + (\theta_3 + \theta_5 + \theta_6 + \theta_7) \\
(4)\quad & E_{A\to Y}(1,1) + E_{A\to M_1Y}(0,1) + E_{A\to M_2\to Y}(0,0) \\
& = (\theta_1 + \theta_4 + \theta_5 + \theta_7) + (\theta_2 + \theta_6) + (\theta_3) \\
(5)\quad & E_{A\to Y}(0,1) + E_{A\to M_1Y}(1,1) + E_{A\to M_2\to Y}(0,0) \\
& = (\theta_1 + \theta_5) + (\theta_2 + \theta_4 + \theta_6 + \theta_7) + (\theta_3) \\
(6)\quad & E_{A\to Y}(1,0) + E_{A\to M_1Y}(0,0) + E_{A\to M_2\to Y}(1,1) \\
& = (\theta_1 + \theta_4) + (\theta_2) + (\theta_3 + \theta_5 + \theta_6 + \theta_7)
\end{aligned}
$$

Table 5.2: All six possible three-way decompositions and their parameterization in model (5.10). Each component on the lefthand side of the equation is represented by a linear combination of parameters on the righthand side (grouped in parentheses).

on the first two decompositions in Table 5.2 as their sequential approach builds on identification of $E_{A\to Y}(0,0)$ and $E_{A\to M_1Y}(1,1)$, as outlined in the previous section. The remaining four decompositions involve instances of $E_{A\to Y}(a', a'')$ with $a' \neq a''$, and instances of $E_{A\to M_1Y}(a, a'')$ with $a \neq a''$, which require slightly stronger identification assumptions, as discussed next.

Identification

Two-way decomposition into joint natural direct and indirect effects can be obtained if assumptions (5.1)-(5.4) hold with respect to the joint mediator $\{M_1, M_2\}$. We refer to the corresponding conditions in NPSEMs as (i')-(iv').

Such first-stage decomposition can be obtained for the DAG in Figure 5.2, but also for the DAGs in Figures 5.3A and 5.3B. This may come as a surprise since the effect of M_1 on M_2 is confounded either by an unmeasured

confounder U (Figure 5.3A) or (measured) intermediate confounder L (Figure 5.3B). However, this does not hinder identification of the joint natural direct and indirect effect because (i')-(iv') do not impose restrictions on the structural relation between the mediators. The other DAGs, however, do not enable such two-way decomposition. In Figures 5.3C and 5.3D, (ii') and (iv') are violated because of unmeasured confounding by U and intermediate confounding by L, respectively.

All six three-way decompositions in Table 5.2 can be recovered under NPSEMs if, in addition to (i')-(iv'), (v') the effect of M_1 on M_2 is unconfounded within strata of $\{A, C\}$ and (vi') none of the $M_1 - M_2$ confounders are affected by exposure. In contrast to assumptions (i')-(iv'), (v') and (vi') do not allow for unmeasured or intermediate confounding of the effect of M_1 on M_2. Consequently, these assumptions are violated in all discussed DAGs (except the one in Figure 5.2). However, decomposition with respect to the three sequential mediators L, M_1 and M_2 becomes possible under

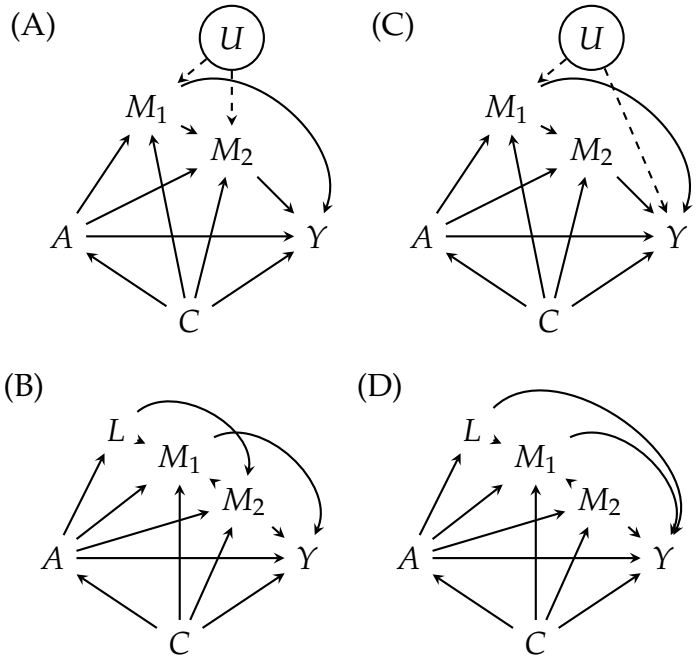


Figure 5.3: Causal diagrams with two sequential mediators M_1 and M_2 and unmeasured confounder U (in panels A and C) or intermediate confounder L (in panels B and D).

more general identification conditions for multiple mediators (see Technical appendix 5.A.2).

Finally, consistent with VanderWeele and Vansteelandt (2013), we show in Technical appendix 5.A.2.5 that the first two decompositions in Table 5.2 necessitate slightly weaker assumptions than (i')-(vi'). In Technical appendix 5.A.2, we also provide a more detailed and formal discussion of identification assumptions, as well as extensions to more than two mediators. Importantly, we generalize the adjustment criterion for two-way decomposition in a single mediator setting (Shpitser and VanderWeele, 2011) to $(k+1)$-way decompositions in settings with k causally ordered mediators.

5.3 Estimation approach

Vansteelandt et al. (2012b) proposed an imputation procedure for fitting natural effect models for single mediators (also see Steen et al. (2016b); Loeys et al. (2013)). Below we describe how this procedure can be extended to recover all possible three-way decompositions in Table 5.2 in settings with a binary exposure (coded 0/1) and two sequential mediators. We first focus on estimation of component effects as defined within strata of C, a covariate set assumed to be sufficient for conditions (i')-(vi') to be met, and next describe how population-average analogs can be obtained. In Technical appendix 5.A.3 we provide some intuition as to why this procedure works and how it relates to Monte Carlo procedures based on generalizations of Pearl (2001, 2012)'s mediation formula (Albert and Nelson, 2011; Daniel et al., 2015).

1. Fit a suitable model for the probability (density) of either

   (i) the first mediator conditional on exposure and covariate set C, for instance, a logistic regression model for binary M_1

   $$\text{logit}P(M_1 = 1|A, C) = \beta_0 + \beta_1 A + \beta_2^\top C, \tag{5.11}$$

   (ii) or the second mediator conditional on exposure, the first media-

tor and covariate set C, for instance, a linear regression model for normally distributed M_2 with constant variance σ^2

$$f(M_2|A, M_1, C) = N\Big(\gamma_0 + \gamma_1 A + \gamma_2 M_1 + \gamma_3 A M_1 + \gamma_4^\top C, \sigma^2\Big). \tag{5.12}$$

2. Fit a suitable model for the outcome mean conditional on exposure, both mediators and covariate set C, for instance, a logistic regression model for binary outcome Y

$$\begin{aligned} &\text{logit}P(Y = 1|A, M_1, M_2, C) \\ &= \delta_0 + \delta_1 A + \delta_2 M_1 + \delta_3 M_2 + \delta_4 A M_1 + \delta_5 A M_2 \\ &\quad + \delta_6 M_1 M_2 + \delta_7 A M_1 M_2 + \delta_8^\top C \end{aligned} \tag{5.13}$$

3. Construct an extended data set by replicating the observed data set 4 times. A similar step has previously been described by Lange et al. (2014) and is best understood in terms of sequential duplication. For the first duplication, add three auxiliary variables a, a' and a''. Let a take on the value of the observed exposure A_i for the first replication and of the counterfactual exposure $1 - A_i$ for the second replication (for each individual i). Let both a' and a'' take on the observed exposure level for both replications. Next, duplicate the resulting extended data once again, now letting a' (a'') take on counterfactual exposure level $1 - A_i$ if model (5.11) ((5.12)) is selected as working model (as illustrated in Tables 5.3 and 5.4, respectively).

4. If model (5.11) is selected, compute weights

$$W_{1i,a'} = \frac{\hat{P}(M_1 = M_{1i}|A = a', C_i)}{\hat{P}(M_1 = M_{1i}|A = a'', C_i)} = \frac{\hat{P}(M_1 = M_{1i}|A = a', C_i)}{\hat{P}(M_1 = M_{1i}|A = A_i, C_i)},$$

or, if model (5.12) is selected, compute weights

$$W_{2i,a''} = \frac{\hat{f}(M_2 = M_{2i}|A = a'', M_{1i}, C_i)}{\hat{f}(M_2 = M_{2i}|A = a', M_{1i}, C_i)} = \frac{\hat{f}(M_2 = M_{2i}|A = a'', M_{1i}, C_i)}{\hat{f}(M_2 = M_{2i}|A = A_i, M_{1i}, C_i)}$$

for each row in the extended data set.

5. Impute nested counterfactuals $Y_i(a, M_{1i}(a'), M_{2i}(a'', M_{1i}(a')))$ as fitted values $\hat{E}(Y_i|A = a, M_{1i}, M_{2i}, C_i)$ from outcome model (5.13) in step 2, for each row in the extended data set.

6. Fit a natural effect model of interest for

$$E\{Y(a, M_{1i}(a'), M_{2i}(a'', M_{1i}(a')))|C\}$$

to the extended data by regressing the imputed outcomes on a, a', a'' and C, weighting by the weights obtained in step 4.

In contrast to direct application of the generalized mediation formula (Albert and Nelson, 2011; Daniel et al., 2015), which relies on a model for the distribution of each of the mediators, our procedure requires only one of these models. This allows investigators to weight by the ratio of densities of the mediator whose corresponding model they believe is less prone to

i	A_i	a	a'	a''	$Y_i(a \cdot a' \cdot a'')$	$\hat{Y}_{i,a}$	$W_{1i,a'}$
1	1	1	1	1	Y_1	$\hat{Y}_{1,1}$	$W_{11,1}$
	1	**0**	1	1	.	$\hat{Y}_{1,0}$	$W_{11,1}$
	1	1	**0**	1	.	$\hat{Y}_{1,1}$	$W_{11,0}$
	1	**0**	**0**	1	.	$\hat{Y}_{1,0}$	$W_{11,0}$
2	0	0	0	0	Y_2	$\hat{Y}_{2,0}$	$W_{12,0}$
	0	**1**	0	0	.	$\hat{Y}_{2,1}$	$W_{12,0}$
	0	0	**1**	0	.	$\hat{Y}_{2,0}$	$W_{12,1}$
	0	**1**	**1**	0	.	$\hat{Y}_{2,1}$	$W_{12,1}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 5.3: Data extension for working models $E(Y|A, M_1, M_2, C)$ and $P(M_1|A, C)$. We use $Y(a \cdot a' \cdot a'')$ and $\hat{Y}_{i,a}$ as shorthand notation for $Y(a, M_1(a'), M_2(a'', M_1(a')))$ and $\hat{E}(Y_i|A_i = a, M_{1i}, M_{2i}, C_i)$, respectively. Imputed nested counterfactuals $\hat{Y}_{i,a}$ for which $a' \neq a''$ (in dark gray) need to be weighted by $W_{1i,a'} = \hat{P}(M_1 = M_{1i}|A = a', C_i)/\hat{P}(M_1 = M_{1i}|A = a'', C_i)$.

i	A_i	a	a'	a''	$Y_i(a \cdot a' \cdot a'')$	$\hat{Y}_{i,a}$	$W_{2i,a''}$
1	1	1	1	1	Y_1	$\hat{Y}_{1,1}$	$W_{21,1}$
	1	**0**	1	1	.	$\hat{Y}_{1,0}$	$W_{21,1}$
	1	1	1	**0**	.	$\hat{Y}_{1,1}$	$W_{21,0}$
	1	**0**	1	**0**	.	$\hat{Y}_{1,0}$	$W_{21,0}$
2	0	0	0	0	Y_2	$\hat{Y}_{2,0}$	$W_{22,0}$
	0	**1**	0	0	.	$\hat{Y}_{2,1}$	$W_{22,0}$
	0	0	0	**1**	.	$\hat{Y}_{2,0}$	$W_{22,1}$
	0	**1**	0	**1**	.	$\hat{Y}_{2,1}$	$W_{22,1}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 5.4: Data extension for working models $E(Y|A, M_1, M_2, C)$ and $f(M_2|A, M_1, C)$. We use $Y(a \cdot a' \cdot a'')$ and $\hat{Y}_{i,a}$ as shorthand notation for $Y(a, M_1(a'), M_2(a'', M_1(a')))$ and $\hat{E}(Y_i|A_i = a, M_{1i}, M_{2i}, C_i)$, respectively. Imputed nested counterfactuals $\hat{Y}_{i,a}$ for which $a' \neq a''$ (in dark gray) need to be weighted by $W_{2i,a''} = \hat{P}(M_2 = M_{2i}|A = a'', M_{1i}, C_i)/\hat{P}(M_2 = M_{2i}|A = a', M_{1i}, C_i)$.

misspecification. If, for instance, M_1 is binary and M_2 continuous, as in the examples given for models (5.11) and (5.12), weighting for M_1 would be most appropriate, since it allows analysts to refrain from modeling the (conditional) relationship between the mediators and making distributional assumptions.

The natural effect model from step 6 can be fitted to the weighted imputations to obtain estimates for stratum-specific component effects. If both exposure A and confounders C are discrete, saturated models can be fitted as long as C is not high-dimensional. In all other cases, our approach demands model restrictions. This improves interpretability of the results, but also increases the risk of misspecification of the natural effect model which may, in turn, lead to biased estimation of the component effects. However, as long as the structure of the imputation model is chosen sufficiently rich so as to minimize the risk of it being misspecified, results from an overly restrictive natural effect model may still be viewed as a useful summary (Vansteelandt et al., 2012b).

Component effects within strata of C^*, a subset of C, can be obtained by fitting a natural effect model for $E\{Y(a, M_1(a'), M_2(a'', M_1(a')))|C^*\}$ conditional on a, a', a'' and C^* upon multiplying the weights from step 4 by $\hat{P}(A = A_i|C_i^*)/\hat{P}(A = A_i|C_i)$. If C^* is empty, the corresponding natural effect model encodes population-average rather than stratum-specific effects and the numerator can simply be replaced by 1. Inverse weighting then enables transporting results to the general population as it accounts for the possibly selective nature of subjects with observed exposure $A = A_i$.

Finally, standard errors and confidence intervals for this imputation estimator can be obtained using a bootstrap procedure (including steps 1-6). Bootstrapping is preferred over use of default standard errors for parameter estimates of natural effect models returned by statistical software as the latter fail to account for uncertainty due to estimation of the working models.

Technical appendix 5.A.4 provides a detailed description on how to adapt the above procedure to continuous exposures (building on Vansteelandt et al. (2012b)), and to settings without interactions between component effects (building on an estimation procedure similar to the one described in VanderWeele et al. (2014)). It also explains how to implement our procedure and obtain bootstrap-based standard errors and confidence intervals in R.

In the next section, we reassess the mediating mechanisms from the empirical example introduced earlier by applying our suggested procedure to obtain a three-way decomposition of the total effect of dampness or mold exposure (A) on the presence of depressive symptoms (Y).

5.4 Motivating example revisited

Following Kaufman (2010), we allow for the possibility that mold-related illness ($M_1 = 1$ in the presence of at least one physical condition known to be related to mold exposure or 0 otherwise), affects perceived control (M_2), as measured on a 5-point Likert scale (reverse coded), but not vice versa. The available set of covariates (C) was assumed sufficient for conditions

(i')-(vi') to be met. A logistic natural effect model

$$\begin{aligned}&\text{logit}P\{Y(a, M_1(a'), M_2(a'', M_1(a'))) = 1|C\} \\ &= \eta_0 + \eta_1 a + \eta_2 a' + \eta_3 a'' + \eta_4 aa' + \eta_5 aa'' + \eta_6 a'a'' + \eta_7 aa'a'' + \eta_8^\top C\end{aligned} \tag{5.14}$$

was fitted to decompose the total effect odds ratio (OR) of dampness or mold exposure A on the presence of depressive symptoms Y (conditional on baseline covariates C), which was estimated to be 1.38 (95% confidence interval (CI): 1.09, 1.73). This was done following steps 1-6 of the previous section. First, mediator models (5.11) (for the probability of mold-related illness M_1) and (5.12) (for the density of perceived control M_2) and an extended version of outcome model (5.13) were fitted to the original data. The latter was used to impute nested counterfactuals in the data set that was extended according to whether model (5.11) or (5.12) was chosen to calculate regression weights for natural effect model (5.14). Each of the working models was specified to include all possible two- and three-way interactions between exposure and mediators to ensure that different decompositions resulting from model (5.14) appropriately reflected differences dictated by the data. For simplicity of exposition, we excluded interaction or polynomial terms involving baseline covariates. A more elaborate model focusing on effect modification by covariates as well as a marginal natural effect model are described in further detail in Empirical analysis section 5.B, which provides a more detailed report of the analyses of this section. 1000 bootstrap samples were drawn to calculate 95% standard normal approximation bootstrap confidence intervals.

Results for all possible three-way decompositions are displayed in Figure 5.4. Since different choices of working models yielded similar estimates, we only report estimates obtained upon weighting by the ratio of probabilities of M_1. The joint natural direct effect OR, $\exp\left(E_{A\to Y}(0, 0|C)\right)$, was 1.25 (95% CI: 0.99, 1.57). The odds of depression within a population (with specific individual and housing characteristics as defined within strata of C) would thus increase by 24% if all individuals were to be moved from a dry dwelling to a damp and moldy residence without their physical condition

nor their sense of control over one's living environment being affected by it. Its complement, the joint natural indirect effect OR was 1.10 (95% CI: 1.03, 1.19). That is, if all individuals were exposed to residential dampness and mold, then the effect of changing both their physical condition and perceived control to what it would be if they were not to live under such poor housing conditions, would be to reduce the odds of depression by 9%. A reduction of 5% would be attributed to changing their physical condition; $\exp\left(\hat{E}_{A\to M_1 Y}(1,1|C)\right) = 1.05$ (95% CI: 1.02, 1.09). Another reduction of about 4% would be attributed to additionally changing their perceived control, in as far as earlier changes in their physical condition would not yet have done so; $\exp\left(\hat{E}_{A\to M_2\to Y}(1,0|C)\right) = 1.05$ (95% CI: 0.98, 1.12).

Natural effect model (5.14) not only permits estimation of the component effects, but also enables probing potential interactions between causal mechanisms. For instance, a multivariate Wald test based on the bootstrap normal approximation indicated the mediating mechanisms captured by $E_{A\to M_1 Y}$ and $E_{A\to M_2\to Y}$ did not interact in their effect on the outcome, i.e. the null that $\eta_6 = \eta_7 = 0$ could not be rejected at the 5% level ($\chi^2 = 1.35, p = 0.51$).

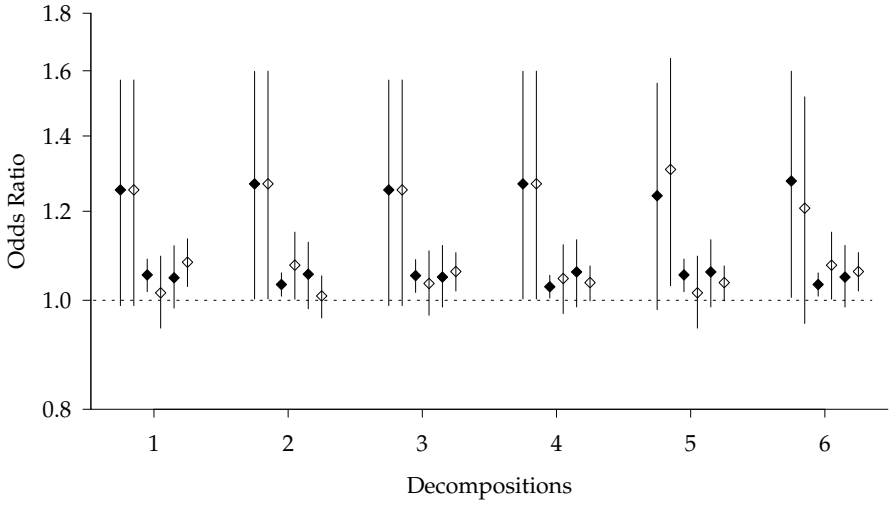


Figure 5.4: Odds ratio estimates and corresponding 95% confidence intervals for each of the stratum-specific analogs of the component effects displayed in Table 5.2 (on the log odds ratio scale). Components are grouped per decomposition and displayed in the same order as in Table 5.2. Estimates are based on natural effect model (5.14), fitted upon weighting by $W_{1i,a'}$ (black) or $W_{2i,a''}$ (white).

	Weighted by $W_{1i,a'}$		Weighted by $W_{2i,a''}$	
Component	Estimate	95% CI	Estimate	95% CI
$\exp\left(\hat{E}_{A\to Y}\right)$	1.260	1.000, 1.573	1.259	1.000, 1.571
$\exp\left(\hat{E}_{A\to M_1 Y}\right)$	1.042	1.015, 1.069	1.041	0.995, 1.089
$\exp\left(\hat{E}_{A\to M_2\to Y}\right)$	1.052	1.008, 1.098	1.048	1.016, 1.079

Table 5.5: Estimates and 95% confidence intervals of the component effects odds ratios, with component effects as parameterized in the logistic natural effect model $\text{logit}P\{Y(a, M_1(a'), M_2(a'', M_1(a'))) = 1|C\} = \zeta_0 + \zeta_1 a + \zeta_2 a' + \zeta_3 a'' + \zeta_4^\top C$.

In addition, there were no substantial differences between decompositions in Figure 5.4, i.e. the null that $\eta_4 = \eta_5 = \eta_6 = \eta_7 = 0$ could not be rejected at the 5% level ($\chi^2 = 3.43$, $p = 0.49$). The absence of such interactions not only facilitates interpretation of the component effects, it may also lead to more precise estimates when fitting a natural effect model that excludes these interaction terms. However, as the estimates and their 95% confidence intervals in Table 5.5 suggest, this did not result in the anticipated efficiency gain. Interestingly, in the absence of interactions, one may refrain from modeling mediator densities altogether by adopting a fully imputation-based estimation procedure (see Technical appendix 5.A.4.3).

Finally note that this illustrative analysis is likely oversimplistic as the assumptions encoded in the DAG in Figure 5.2 may well be violated. For instance, possible attempts to control mold growth, such as cleaning or ventilating the house, are possibly affected by the level of mold exposure and may, in turn, influence both mold-related illness and perceived control over one's home. The level of exposure may therefore be inherently time-varying, adding another level of complexity.

5.5 Discussion

In this chapter, we focused on the finest decomposition that can be obtained in settings with multiple causally ordered mediators without introducing sensitivity parameters (Albert and Nelson, 2011; Daniel et al., 2015; Imai and

Yamamoto, 2013) or parametric assumptions, as in the SEM tradition (see De Stavola et al. (2014) for a review). We pointed out that previous approaches with a similar focus yield only a subset of all possible decompositions (VanderWeele and Vansteelandt, 2013). Moreover, we proposed a flexible approach for estimating component effects and derived sufficient conditions for their identification.

Our estimation approach combines imputation- and weighting-based methods to fit a novel class of natural effect models (Lange et al., 2012; Loeys et al., 2013; Steen et al., 2016b; Vansteelandt et al., 2012b), for multiple mediators. As opposed to Monte Carlo approaches (Albert and Nelson, 2011; Daniel et al., 2015), which dictate modeling the joint density of the mediators, our approach necessitates modeling the density of only one of the mediators, enabling practitioners to bet on the mediator they feel most confident about modeling correctly. In the absence of interactions between component effects, one may even avoid modeling mediator densities altogether, at the expense of an additional model for the outcome, as discussed in Technical appendix 5.A.4.3. This may be particularly attractive in settings with large numbers of mediators as it dramatically reduces modeling demands. Nonetheless, when the joint density is correctly specified, fully parametric Monte Carlo approaches yield more efficient estimators for the component effects. Alternatively, one could refrain from modeling the outcome and, instead, opt for an approach that exclusively relies on weighting. However, this then requires correct specification of the joint density of the mediators, as in Lange et al. (2014) and Taguri et al. (2015) for settings with multiple causally unrelated mediators (also see VanderWeele et al. (2014) for a similar approach in settings with intermediate confounding). Unless there are major concerns for model extrapolation due to inadequate modeling of the outcome (Vansteelandt et al., 2012b), we generally discourage such approach, especially when dealing with continuous mediators, because typical issues of instability, characteristic for weighting methods, tend to exacerbate when combining density weights for each of the mediators.

In addition to added flexibility in choice of working models, natural effect modeling owes much of its attractiveness to its parsimonious parameterization. It enables testing certain hypotheses of interest (especially those

concerning effect modification by baseline covariates) which, in particular settings, cannot be tested by direct application of the mediation formula (Loeys et al., 2013; Vansteelandt et al., 2012b). In our illustration, we have demonstrated that differences between decompositions listed in Table 5.2, captured by the interaction terms of the natural effect model, can be formally tested in a straightforward manner.

Although we have restricted our presentation to applications with only two sequential mediators, results can straightforwardly be extended to settings with more mediators. In Technical appendix 5.A.2, we illustrate that, in such complex settings, our set of assumptions leads to a manageable and piecemeal identification procedure. Moreover, in settings where the structural dependence between certain subsequent mediators is unclear, these groups of mediators can simply be treated as joint mediators in order to render identification assumptions of the corresponding component effects more plausible (Taguri et al., 2015; VanderWeele and Vansteelandt, 2013).

5.A Technical appendices

5.A.1 Targeted decompositions

In the main text, we have focused on the finest possible decomposition(s) that can be obtained in settings with two sequential mediators without imposing parametric restrictions. We have demonstrated that such settings lead to 6 such possible decompositions. More generally, settings with k sequential mediators lead to $(k+1)!$ such possible decompositions along $k+1$ distinct pathways that can (possibly) be estimated from the data without making parametric assumptions. If the exposure is binary, saturated natural effect models with 2^{k+1} parameters can be fitted for estimating population-average component effects (see Table 5.6).

k	2^k	$(2^k)!$	$k+1$	$(k+1)!$	2^{k+1}
2	4	24	3	6	8
3	8	40,320	4	24	16
4	16	2.092×10^{13}	5	120	32

Table 5.6: Finest possible decomposition in the presence of k sequential mediators that can be non-parametrically obtained with (Daniel et al., 2015) and without introducing sensitivity parameters (current chapter).

In contrast, Daniel et al. (2015) focus on the finest possible decomposition into 2^k distinct pathways. However, identification results for the corresponding component effects make reference to unknown sensitivity parameters which capture cross-world correlations that cannot be identified from the data. As illustrated in Table 5.6, the total number of such possible decompositions $(2^k)!$ increases exponentially with increasing k. Moreover, it does so at a rate that far exceeds that of the number of decompositions that are targeted in the current chapter.

Finally, VanderWeele and Vansteelandt (2013) describe a decomposition in terms of the same $k+1$ pathways as in the current chapter, but because of the sequential treatment of the mediators only 2 out of the $(k+1)!$ decompositions are obtained.

To illustrate both the differences in types of decompositions described in these papers and extensions of our approach to more than two sequential mediators, we will throughout focus on a setting with three sequential mediators M_1, M_2

and M_3 (as depicted in the causal diagram in Figure 5.5). The finest possible decomposition in such a setting involves 8 distinct pathways from exposure to outcome (i.e. pathways $A \to Y$, $A \to M_1 \to Y$, $A \to M_2 \to Y$, $A \to M_3 \to Y$, $A \to M_1 \to M_2 \to Y$, $A \to M_1 \to M_3 \to Y$, $A \to M_2 \to M_3 \to Y$ and $A \to M_1 \to M_2 \to M_3 \to Y$). However, this decomposition can never be recovered from the observed data without making assumptions about the joint distribution of the same counterfactual at different exposure levels. Under certain conditions (articulated in section 5.A.2) we can, however, obtain a four-way decomposition of the total causal effect into

1. a direct effect with respect to $\{M_1, M_2, M_3\}$ as joint mediator (capturing pathway $A \to Y$),
2. an indirect effect with respect to M_1 as mediator (capturing all pathways along M_1 that feed into Y, i.e. $A \to M_1 \to Y$, $A \to M_1 \to M_2 \to Y$, $A \to M_1 \to M_3 \to Y$, and $A \to M_1 \to M_2 \to M_3 \to Y$),
3. a partial indirect with respect to M_2 as mediator (capturing all pathways along M_2 that do not first pass M_1, i.e. $A \to M_2 \to Y$ and $A \to M_2 \to M_3 \to Y$)
4. a partial indirect with respect to M_3 as mediator (capturing all pathways along M_3 that do not first pass earlier intermediates M_1 or M_2, i.e. only $A \to M_3 \to Y$)

Identification of the corresponding components depends on identification of

$$E\{Y(a, M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a'))))\}. \tag{5.15}$$

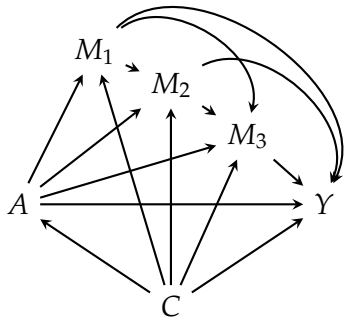


Figure 5.5: Causal diagram with three sequential mediators.

This expected nested counterfactual involves 4 hypothetical exposure levels a, a', a'' and a''', corresponding to the 4 pathways or components. Moreover, it is a specific case of the more general nested counterfactual

$$Y(a^0, M_1(a^1), M_2(a^2, M_1(a^3)), M_3(a^4, M_1(a^5), M_2(a^6, M_1(a^7)))) \tag{5.16}$$

that imposes the restriction that no conflicting hypothetical exposure levels can be associated to a single mediator (i.e. $a^1 = a^3 = a^5 = a^7 = a'$ and $a^2 = a^6 = a''$). This restriction implies that contrasts based on expectations of nested counterfactuals of the form in expression (5.15) only involve path-specific effects transmitted via *all* paths through a given mediator (or set of mediators) *excluding* those that pass earlier intermediates. It thus allows deriving all component effects described earlier, just as any path-specific effect that is a combination of these components (such as e.g. the joint indirect effect with respect to $\{M_1, M_2\}$ as a joint mediator, which combines the second and third component effect). It does not, however, allow to derive path-specific effects such as the one transmitted only through pathway $A \to M_1 \to Y$, since this effect does not capture all pathways that pass through M_1. Neither does it allow obtaining e.g. the natural indirect effect with respect to M_2 (which captures pathways $A \to M_2 \to Y$, $A \to M_1 \to M_2 \to Y$, $A \to M_2 \to M_3 \to Y$ and $A \to M_1 \to M_2 \to M_3 \to Y$) as it captures all paths through M_2, but also those that first pass M_1, which is an earlier intermediate. Note that the latter path-specific effects, could, however, be derived by contrasts of expected counterfactuals of the form in expression (5.16). Such contrasts would enable obtaining the finest possible decomposition as described in Daniel et al. (2015). However, non-parametric identification of expected counterfactuals of this form cannot be obtained without introducing sensitivity parameters, as opposed to identification of expected counterfactuals of the form in expression (5.15).

Finally, the decomposition approach described in VanderWeele and Vansteelandt (2013) would also result in a four-way decomposition into the same components as previously listed. These components are derived from a series of (in this case three) single mediator analyses in which respectively, M_1, $\{M_1, M_2\}$ and $\{M_1, M_2, M_3\}$ are considered as the mediator of interest. The first analysis thus requires identification of $E\{Y(a, M_1(a'))\}$, whereas the second and third analysis require identification of $E\{Y(a, M_1(a'), M_2(a'))\}$ and $E\{Y(a, M_1(a'), M_2(a'), M_3(a'))\}$, respectively. Under certain composition assumptions, the involved expectations of

nested counterfactuals can respectively be rewritten as

$$E\{Y(a, M_1(a'), M_2(a, M_1(a')), M_3(a, M_1(a'), M_2(a, M_1(a'))))\},$$
$$E\{Y(a, M_1(a'), M_2(a', M_1(a')), M_3(a, M_1(a'), M_2(a', M_1(a'))))\}$$

and

$$E\{Y(a, M_1(a'), M_2(a', M_1(a')), M_3(a', M_1(a'), M_2(a', M_1(a'))))\},$$

which can easily be seen to be special cases of the form in expression (5.15). It can be shown that the additional restrictions on these nested counterfactuals (which arise because of the sequential nature of this approach) imply that, when using this sequential approach, only 2 out of the 24 decompositions can be derived (since, for each of the component effects only 2 instances can be derived).

More generally, for settings with k sequential mediators, one can, under certain conditions (discussed below), identify mediated effects with respect to a mediator M_j in as far that they act over and above the effect transmitted by earlier intermediates $M_1, ..., M_{j-1}$. Identification of the component effects depends on identification of

$$E\{Y(a^0, M_1^*, ..., M_k^*)\} \tag{5.17}$$

with $M_1^* = M_1(a^1)$ and $M_j^* = M_j(a^j, M_1^*, ..., M_{j-1}^*)$. This expected nested counterfactual involves $k+1$ hypothetical exposure levels a^0 to a^k, corresponding to the $k+1$ pathways or components.

5.A.2 Identification

In this section, we focus on identification of the component effects of all possible four-way decompositions in the presence of three sequential mediators. Identification assumptions and results can, however, easily be shown to extend to settings with more mediators.

5.A.2.1 Identification assumptions

As discussed in section 5.A.1, identification of the component effects can be obtained upon identifying expression (5.15) which is established if for all a, a', a'', a''',

m_1, m_2 and m_3

$$M_1(a) \perp\!\!\!\perp A|C \tag{B0}$$
$$M_2(a, m_1) \perp\!\!\!\perp \{A, M_1\}|C \tag{B1}$$
$$M_2(a, m_1) \perp\!\!\!\perp M_1(a')|C \tag{C1}$$
$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{A, M_1, M_2\}|C \tag{B2}$$
$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{M_1(a'), M_2(a'', m_1)\}|C \tag{C2}$$
$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{A, M_1, M_2, M_3\}|C \tag{B3}$$
$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{M_1(a'), M_2(a'', m_1), M_3(a''', m_1, m_2)\}|C. \tag{C3}$$

5.A.2.2 Identification result

Below, we derive the identification result for the conditional equivalent of expression (5.15) given baseline covariates C, under (B0)-(B3) and (C1)-(C3). The result for expression (5.15) can be obtained upon marginalizing over the joint density of C. Moreover, identification results for expression (5.17) can more generally be obtained in settings with k sequential mediators upon extending the assumption set from section 5.A.2.1 accordingly. These results can be seen as a generalization of the identification result for $E\{Y(a, M(a'))\}$ (in single mediator settings) commonly referred to as Pearl's mediation formula (Pearl, 2001, 2012).

$$\begin{aligned}
&E\{Y(a, M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a'))))|C\} \\
&= \int E\big\{Y(a, m_1, m_2, m_3)\big|M_1(a') = m_1, M_2(a'', M_1(a')) = m_2, \\
&\qquad\qquad M_3(a''', M_1(a'), M_2(a'', M_1(a'))) = m_3, C\big\} \\
&\qquad \times dF_{M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a')))|C}(m_1, m_2, m_3) \\
&= \int E\big\{Y(a, m_1, m_2, m_3)\big|M_1(a') = m_1, M_2(a'', m_1) = m_2, M_3(a''', m_1, m_2) = m_3, C\big\} \\
&\qquad \times dF_{M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a')))|C}(m_1, m_2, m_3) \\
&\overset{(C3)}{=} \int E\big\{Y(a, m_1, m_2, m_3)\big|C\big\} \\
&\qquad \times dF_{M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a')))|C}(m_1, m_2, m_3) \\
&\overset{(B3)}{=} \int E\big\{Y\big|A = a, M_1 = m_1, M_2 = m_2, M_3 = m_3, C\big\} \\
&\qquad \times dF_{M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a')))|C}(m_1, m_2, m_3) \\
&= \int E\big\{Y\big|A = a, M_1 = m_1, M_2 = m_2, M_3 = m_3, C\big\} dF_{M_1(a')|C}(m_1) \\
&\qquad \times dF_{M_2(a'', M_1(a'))|M_1(a') = m_1, C}(m_2)
\end{aligned}$$

$$\times dF_{M_3(a''',M_1(a'),M_2(a'',M_1(a')))|M_1(a')=m_1,M_2(a'',M_1(a'))=m_2,C}(m_3)$$

$$= \int E\{Y|A=a,M_1=m_1,M_2=m_2,M_3=m_3,C\}dF_{M_1(a')|C}(m_1)$$
$$\times dF_{M_2(a'',m_1)|M_1(a')=m_1,C}(m_2)$$
$$\times dF_{M_3(a''',m_1,m_2)|M_1(a')=m_1,M_2(a'',M_1(a'))=m_2,C}(m_3)$$

$$\overset{(B0)}{=} \int E\{Y|A=a,M_1=m_1,M_2=m_2,M_3=m_3,C\}dF_{M_1|A=a',C}(m_1)$$
$$\times dF_{M_2(a'',m_1)|M_1(a')=m_1,C}(m_2)$$
$$\times dF_{M_3(a''',m_1,m_2)|M_1(a')=m_1,M_2(a'',M_1(a'))=m_2,C}(m_3)$$

$$\overset{(C1)}{=} \int E\{Y|A=a,M_1=m_1,M_2=m_2,M_3=m_3,C\}dF_{M_1|A=a',C}(m_1)$$
$$\times dF_{M_2(a'',m_1)|C}(m_2)$$
$$\times dF_{M_3(a''',m_1,m_2)|M_1(a')=m_1,M_2(a'',M_1(a'))=m_2,C}(m_3)$$

$$\overset{(B1)}{=} \int E\{Y|A=a,M_1=m_1,M_2=m_2,M_3=m_3,C\}dF_{M_1|A=a',C}(m_1)$$
$$\times dF_{M_2|A=a'',M_1=m_1,C}(m_2)$$
$$\times dF_{M_3(a''',m_1,m_2)|M_1(a')=m_1,M_2(a'',M_1(a'))=m_2,C}(m_3)$$

$$\overset{(C2)}{=} \int E\{Y|A=a,M_1=m_1,M_2=m_2,M_3=m_3,C\}dF_{M_1|A=a',C}(m_1)$$
$$\times dF_{M_2|A=a'',M_1=m_1,C}(m_2)dF_{M_3(a''',m_1,m_2)|C}(m_3)$$

$$\overset{(B2)}{=} \int E\{Y|A=a,M_1=m_1,M_2=m_2,M_3=m_3,C\}dF_{M_1|A=a',C}(m_1)$$
$$\times dF_{M_2|A=a'',M_1=m_1,C}(m_2)dF_{M_3|A=a''',M_1=m_1,M_2=m_2,C}(m_3)$$

It can easily be seen that identification can still be obtained upon relaxing assumptions (B0), (B1), (B2) and (B3) as below:

$$M_1(a) \perp\!\!\!\perp I(A=a)|C \tag{B0'}$$
$$M_2(a,m_1) \perp\!\!\!\perp \{I(A=a),I(M_1=m_1)\}|C \tag{B1'}$$
$$M_3(a,m_1,m_2) \perp\!\!\!\perp \{I(A=a),I(M_1=m_1),I(M_2=m_2)\}|C \tag{B2'}$$
$$Y(a,m_1,m_2,m_3) \perp\!\!\!\perp \{I(A=a),I(M_1=m_1),I(M_2=m_2),I(M_3=m_3)\}|C \tag{B3'}$$

Note that, in contrast to (B0)-(B3), the weaker subset (B0')-(B3') does not encode any cross-world independencies.[1] Simplifications under NPSEMs (in section 5.A.2.3), however, only apply to the stronger subset (B1), (B2) and (B3), as identification of the component effects critically hinges on certain cross-world

[1] It can easily be shown that this subset is implied by Robins and Richardson (2010)'s Minimal Causal Model (MCM) associated with the DAG in Figure 5.5.

independence assumptions.

5.A.2.3 Simplifications under NPSEMs

Note that assumptions (B0), (B1) and (C1) roughly correspond to assumptions (5.1)-(5.4) (for identification of natural effects in single mediator settings) in the main text, with M_1 being the mediator and M_2 the outcome. It is easily seen that (B0) corresponds to assumption (5.3) in the main text, (C1) corresponds to assumption (5.4) and (B1) combines assumption (5.1) and $Y(a,m) \perp\!\!\!\perp M|A,C$ (a stronger variant of assumption (5.2)).

Below we show that cross-world assumptions (C1)-(C3) are redundant under NPSEMs, as they follow from (B1)-(B3). This proof builds on the composition axiom (Dawid, 1979; Pearl, 1988), which holds for all counterfactuals under NPSEMs (Richardson and Robins, 2013) and states that if $X \perp\!\!\!\perp Y|Z$ and $W \perp\!\!\!\perp Y|Z$, then $\{X,W\} \perp\!\!\!\perp Y|Z$.[2]

First, it follows from (B1) that

$$M_2(a,m_1) \perp\!\!\!\perp A|C$$

and

$$M_2(a,m_1) \perp\!\!\!\perp M_1(a')|A=a',C.$$

Together, these imply

$$M_2(a,m_1) \perp\!\!\!\perp \{A=a', M_1(a')\}|C,$$

which, in turn, implies (C1)

$$M_2(a,m_1) \perp\!\!\!\perp M_1(a')|C.$$

This result has already been proven in Shpitser and VanderWeele (2011).

Similarly, it follows from (B2) that

$$M_3(a,m_1,m_2) \perp\!\!\!\perp A|C$$

[2]Not to be confused with the composition assumption for counterfactuals (also called generalized consistency assumption) which states that for each a, $Y(a,M(a)) = Y(a)$ with probability 1.

and

$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{M_1(a'), M_2(a')\} | A = a', C.$$

Relying on the composition assumption that $M_2(a') = M_2(a', M_1(a'))$, the latter conditional independency, in turn, implies

$$M_3(a, m_1, m_2) \perp\!\!\!\perp M_1(a') | A = a', C$$

and

$$M_3(a, m_1, m_2) \perp\!\!\!\perp M_2(a', m_1') | A = a', M_1(a') = m_1', C,$$

which, together with the first conditional independency, combine into

$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{A = a', M_1(a')\} | C$$

and

$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{A = a', M_1(a') = m_1', M_2(a', m_1')\} | C.$$

From these conditional independencies, we get

$$M_3(a, m_1, m_2) \perp\!\!\!\perp M_1(a') | C$$

and

$$M_3(a, m_1, m_2) \perp\!\!\!\perp M_2(a', m_1) | C,$$

which, by the composition axiom, combine into (C2)

$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{M_1(a'), M_2(a'', m_1)\} | C.$$

Finally, it follows from (B3) that

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp A | C$$

and

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{M_1(a'), M_2(a'), M_3(a')\} | A = a', C.$$

Relying on the composition assumptions that $M_2(a') = M_2(a', M_1(a'))$ and $M_3(a') = M_3(a', M_1(a'), M_2(a'))$, the latter conditional independency implies

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp M_1(a') | A = a', C,$$

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp M_2(a', m_1')|A = a', M_1(a') = m_1', C$$

and

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp M_3(a', m_1', m_2')|A = a', M_1(a') = m_1', M_2(a', m_1') = m_2', C,$$

which, together with the first conditional independency (i.e. $Y(a, m_1, m_2, m_3) \perp\!\!\!\perp A|C$), combine into

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{A = a', M_1(a')\}|C,$$

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{A = a', M_1(a') = m_1', M_2(a', m_1')\}|C$$

and

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{A = a', M_1(a') = m_1', M_2(a', m_1') = m_2', M_3(a', m_1', m_2')\}|C.$$

From these conditional independencies, we get

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp M_1(a')|C,$$

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp M_2(a', m_1)|C$$

and

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp M_3(a', m_1, m_2)|C,$$

which, by the composition axiom, combine into (C3)

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{M_1(a'), M_2(a'', m_1), M_3(a''', m_1, m_2)\}|C.$$

5.A.2.4 A sequential identification approach via the adjustment criterion

As demonstrated in section 5.A.2.3, under NPSEMs, assumptions (B0)-(B3) are sufficient for identifying expression (5.15) and thus recovering all possible targeted four-way decompositions (as listed in section 5.A.1). These assumptions are satisfied under certain no unmeasured or intermediate confounding assumptions which lead to a piecemeal sequential identification procedure that easily extends to more complex settings with more than three sequential mediators. That is, in general, expression (5.17) is identified (and thus all $(k+1)!$ targeted $(k+1)$-way

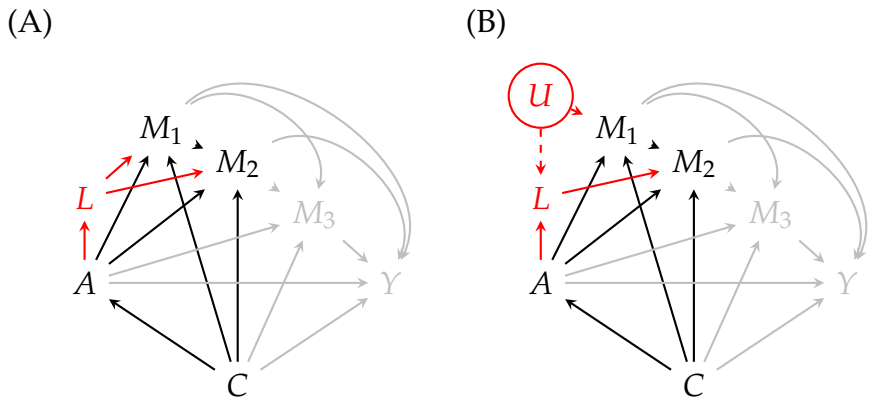


Figure 5.6: Causal diagrams with three sequential mediators and $M_1 - M_2$ confounder L affected by A.

decompositions can be recovered from the observed data) if there is a measured set of covariates C that satisfies the adjustment criterion[3] (Shpitser et al., 2010; Shpitser and VanderWeele, 2011) relative to each of the joint effects listed below.

First, under a NPSEM, (B0) requires that the effect of A on M_1 is unconfounded within strata of C. This is equivalent to the requirement that C satisfies the adjustment criterion relative to (A, M_1), which enables identification of the total causal effect of A on M_1.

Second, (B1) requires that the joint effect of $\{A, M_1\}$ on M_2 is unconfounded within strata of C. In addition, it requires that no $M_1 - M_2$ confounders[4] are descendants of A. This is equivalent to requiring that C satisfies the adjustment criterion relative to $(\{A, M_1\}, M_2)$, which enables identification of the joint effect of $\{A, M_1\}$ on M_2. Furthermore, under NPSEMs, (B1) implies (C1). (B0) and (B1) (i.e. C satisfies the adjustment criterion relative to both (A, M_1) and $(\{A, M_1\}, M_2)$) therefore identify the natural direct and indirect effect of A on M_2 with respect to M_1 as mediator. This result was established in Shpitser and VanderWeele (2011).

[3] Refer to section 2.4.2 of chapter 2.

[4] Note that in the remainder of the text, we denote $M_i - M_{i+1}$ confounders to be the set of covariates that is sufficient to adjust for confounding of the association between M_i and M_{i+1} (conditional on A and earlier intermediates in the DAG). These do not necessarily need to be common causes, as long as they render the relation between M_1 and M_2 unconfounded. For instance, in the DAG in the right panel of Figure 5.6, L would be considered to be an $M_1 - M_2$ confounder, as it is sufficient to render the association between M_1 and M_2 (given A) unconfounded by blocking the spurious pathway $M_1 \leftarrow U \rightarrow L \rightarrow M_2$, even though L does not affect both M_1 and M_2 as in the DAG in the left panel of Figure 5.6.

An example. For instance, in the DAG in the left panel of Figure 5.6, the directed paths $A \to M_2$, $A \to L \to M_2$ and $M_1 \to M_2$ are proper causal with respect to $(\{A, M_1\}, M_2)$. In contrast, the directed paths $A \to M_1 \to M_2$ and $A \to L \to M_1 \to M_2$ are not proper causal with respect to $(\{A, M_1\}, M_2)$ as they intersect $\{A, M_1\}$ at M_1. Furthermore, L is a confounder of the relation between M_1 and M_2 and should therefore be included in the adjustment set Z (which also includes C). However, by including L in the adjustment set Z, the proper causal path $A \to L \to M_2$ is blocked and the condition that no element of Z is a descendant in $\mathcal{G}_{\overline{A,M_1}}$ of any $W \notin \{A, M_1\}$ which lies on a proper causal path from $\{A, M_1\}$ to Y is violated (since, by definition, L is a descendant of itself). We cannot have it both ways (i.e. fully adjusting for $M_1 - M_2$ confounding and not blocking any proper causal paths with respect to $\{A, M_1\}$). Thus, we conclude that, at least under this DAG, no covariate set Z satisfies the adjustment criterion relative to both (A, M_1) and $(\{A, M_1\}, M_2)$, so that we cannot obtain identification of the natural direct and indirect effect of A on M_2 with respect to M_1 as mediator. Likewise, in the DAG in the right panel of Figure 5.6, identification of these natural effects cannot be obtained. In this DAG, the directed paths $A \to M_2$, $A \to L \to M_2$ and $M_1 \to M_2$ are proper causal with respect to $(\{A, M_1\}, M_2)$. In contrast, the directed path $A \to M_1 \to M_2$ is not proper causal with respect to $(\{A, M_1\}, M_2)$ as it intersects $\{A, M_1\}$ at M_1. Although L is not a common cause of M_1 and M_2, its adjustment does enable blocking the spurious (non-causal) pathway $M_1 \leftarrow U \to L \to M_2$ and therefore, L should again be included in the adjustment set Z. However, again, since L is affected by A, including it in the adjustment set Z leads to violation of the condition that no element of Z is a descendant in $\mathcal{G}_{\overline{A,M_1}}$ of any $W \notin \{A, M_1\}$ which lies on a proper causal path from $\{A, M_1\}$ to Y.

Third, (B2) requires that the joint effect of $\{A, M_1, M_2\}$ on M_3 is unconfounded within strata of C. In addition, it requires that no $\{M_1, M_2\} - M_3$ confounders are descendants of A.[5] This is equivalent to requiring that C satisfies the adjustment criterion relative to $(\{A, M_1, M_2\}, M_3)$, which enables identification of the joint effect of $\{A, M_1, M_2\}$ on M_3. Together with the previous assumptions, (B2) identifies the joint natural direct and indirect effect of A on M_3 with respect to $\{M_1, M_2\}$ as joint mediator. This can be seen upon noting that, under NPSEMs,

[5]Note that this also requires that no $M_2 - M_3$ confounders are descendants of M_1.

their identification conditions

$$\begin{aligned} M_3(a, m_1, m_2) &\perp\!\!\!\perp A|C \\ M_3(a, m_1, m_2) &\perp\!\!\!\perp \{M_1, M_2\}|A = a, C \\ \{M_1(a), M_2(a)\} &\perp\!\!\!\perp A|C \\ M_3(a, m_1, m_2) &\perp\!\!\!\perp \{M_1(a'), M_2(a')\}|C \end{aligned}$$

are implied by (B0)-(B2). The first two conditional independencies are implied by (B2). The third conditional independency follows from (B0) and (B1), since, by the composition axiom, these imply

$$\{M_1(a), M_2(a, m_1)\} \perp\!\!\!\perp A|C.$$

5

This conditional independency, in turn, implies

$$M_2(a, m_1) \perp\!\!\!\perp A|M_1(a) = m_1, C$$

and hence

$$M_2(a) \perp\!\!\!\perp A|M_1(a) = m_1, C,$$

which, together with (B0), implies

$$\{M_1(a) = m_1, M_2(a)\} \perp\!\!\!\perp A|C.$$

By the composition axiom this result yields the third conditional independency, when yet again combined with (B0). The fourth conditional independency follows directly from

$$M_3(a, m_1, m_2) \perp\!\!\!\perp A|C$$

and

$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{M_1(a'), M_2(a')\}|A = a', C,$$

which are direct implications of (B2). Combined, they imply

$$M_3(a, m_1, m_2) \perp\!\!\!\perp \{A = a', M_1(a'), M_2(a')\}|C$$

and thus the fourth conditional independency.

Finally, (B3) requires that the joint effect of $\{A, M_1, M_2, M_3\}$ on Y is unconfounded within strata of C. In addition, it requires that no $\{M_1, M_2, M_3\} - Y$

confounders are descendants of A. This is equivalent to requiring that C satisfies the adjustment criterion relative to $(\{A, M_1, M_2, M_3\}, Y)$, which enables identification of the joint effect of $\{A, M_1, M_2, M_3\}$ on Y. Together with the previous assumptions, (B3) identifies the joint natural direct and indirect effect of A on Y with respect to $\{M_1, M_2, M_3\}$ as joint mediator. This can again be seen upon noting that, under NPSEMs, their identification conditions

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp A|C \tag{A1''}$$

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{M_1, M_2, M_3\}|A = a, C \tag{A2''}$$

$$\{M_1(a), M_2(a), M_3(a)\} \perp\!\!\!\perp A|C \tag{A3''}$$

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{M_1(a'), M_2(a'), M_3(a')\}|C \tag{A4''}$$

are implied by (B0)-(B3). The proof boils down to a simple extension of the proof in the previous paragraph.

Note that, as previously mentioned, Shpitser and VanderWeele (2011) demonstrated that if, under a NPSEM, a measured covariate set C can be found that satisfies the adjustment criterion for both (A, M) and $(\{A, M\}, Y)$, the natural direct and indirect effect of A on Y wrt M as a mediator are identified from the observed data, and hence $E\{Y(a, M(a'))\}$ is identified. We have illustrated that, by extension, if, under a NPSEM, a measured covariate set C can be found that simultaneously satisfies the adjustment criterion for (A, M_1), $(\{A, M_1\}, M_2)$, $(\{A, M_1, M_2\}, M_3)$ and $(\{A, M_1, M_2, M_3\}, Y)$, expression (5.15) is identified and thus all possible targeted four-way decompositions can be obtained from the data. More generally, for settings with k sequential mediators $M_1, ..., M_k$, expression (5.17) is identified if there is a covariate set C that satisfies the adjustment criterion for $\{\overline{M}_j, M_{j+1}\}$ for all j from 0 to k, with $M_0 = A$, $M_{k+1} = Y$ and $\overline{M}_j = \{M_0, ..., M_j\}$.

We have thus shown that the adjustment criterion for natural direct and indirect effects, as outlined in Shpitser and VanderWeele (2011), can be extended to settings with k sequential mediators to recover all $(k+1)!$ targeted $(k+1)$-decompositions. Moreover, as shown above, the sequential nature of assumption set B (or the generalized adjustment criterion described above) implies that each additional step from $j-1$ to j (for all j from 1 to k) enables identification of the natural direct and indirect effect of A on M_{j+1} with respect to $\{M_1, ..., M_j\}$ as joint mediator. Conversely, it is easily shown that, under NPSEMs, identification of each of these natural direct and indirect effects (along with the total causal effect of A on M_1) by means of adjustment for a common covariate set C also enables identification

of expression (5.17) by means of adjustment for C. For instance, for settings with $k = 3$, as described above, identification of

- the total causal effect of A on M_1
  or $E\{M_1(a)\}$,
- the natural effects of A on M_2 wrt M_1 as mediator
  or $E\{M_2(a, M_1(a'))\}$,
- the natural effects of A on M_3 wrt $\{M_1, M_2\}$ as (joint) mediator
  or $E\{M_3(a, M_1(a'), M_2(a', M_1(a')))\}$ and
- the natural effects of A on Y wrt $\{M_1, M_2, M_3\}$ as (joint) mediator
  or $E\{Y(a, M_1(a'), M_2(a', M_1(a')), M_3(a', M_1(a'), M_2(a', M_1(a'))))\}$

by means of adjustment for a common covariate set[6] C implies identification of expression (5.15) by means of adjustment for C and thus enables to recover all possible targeted four-way decompositions. For instance, (B3) follows from combining (A1″) and (A4″). By the composition axiom, this combination yields

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{A, M_1(a'), M_2(a'), M_3(a')\}|C,$$

which implies

$$Y(a, m_1, m_2, m_3) \perp\!\!\!\perp \{M_1, M_2, M_3\}|A = a', C$$

for every a'. Combined with (A1″) this then results in (B3). Similarly (B1) and (B2) can be obtained from the other identification conditions.

[6]We wish to distinguish here between identification by means of adjustment for a common covariate set and identification by other means (e.g. by means of adjustment for separate covariate sets or application of the front-door criterion, as described in Pearl (2014)). Identification of the effects listed above only leads to identification of expression (5.15) under NPSEMs, if it is obtained by means of adjustment by a common covariate set C. In particular, it can be shown that identification of expression (5.17) can more generally be obtained under NPSEMs when (i) the total causal effect is identifiable and (ii) the recanting district criterion (Shpitser, 2013) (i.e. a generalization of the recanting witness criterion (Avin et al., 2005) to longitudinal settings with unobserved confounding), does not hold with respect to any of the $k + 1$ constituent components from the targeted $(k + 1)$-way decompositions (i.e. there does not exist a recanting district for any of the component effects).

5.A.2.5 Comparison with sequential identification approach of VanderWeele and Vansteelandt (2013)

Note that the sequential identification approach from section 5.A.2.4 differs from the one described in VanderWeele and Vansteelandt (2013) (henceforth referred to as VWV). Their approach also refers to the same nested set of mediators, i.e. $M_1 \in \{M_1, M_2\} \in ... \in \{M_1, M_2, ..., M_k\}$. However, each of their steps refer to the endpoint Y as outcome, whereas our approach refers to M_{j+1} as outcome when $\{M_1, ..., M_j\}$ is considered the joint set of mediators (with $M_{k+1} = Y$ in the presence of k mediators).

As already suggested in section 5.A.1, our approach yields slightly stronger identification results, leading to all $(k+1)!$ targeted $(k+1)$-way decompositions instead of only 2 such decompositions. Below we illustrate that, for a simple setting with $k = 2$, the conditions outlined in VWV do not imply the corresponding identification assumptions from section 5.A.2.1 for three-way decomposition and are thus not sufficient for identifying all component effects of all possible three-way decompositions. However, we argue that in most realistic settings the differences between identification conditions are subtle and of limited practical relevance.

VWV require the natural direct and indirect effect of A on Y with respect to M_1 as a mediator to be identified, or

$$Y(a, m_1) \perp\!\!\!\perp A|C \tag{A1}$$
$$Y(a, m_1) \perp\!\!\!\perp M_1|A = a, C \tag{A2}$$
$$M_1(a) \perp\!\!\!\perp A|C \tag{A3}$$
$$Y(a, m_1) \perp\!\!\!\perp M_1(a')|C. \tag{A4}$$

Additionally, they require the natural direct and indirect effect of A on Y with respect to the joint mediator set $\{M_1, M_2\}$ to be identified, or

$$Y(a, m_1, m_2) \perp\!\!\!\perp A|C \tag{A1'}$$
$$Y(a, m_1, m_2) \perp\!\!\!\perp \{M_1, M_2\}|A = a, C \tag{A2'}$$
$$\{M_1(a), M_2(a)\} \perp\!\!\!\perp A|C \tag{A3'}$$
$$Y(a, m_1, m_2) \perp\!\!\!\perp \{M_1(a'), M_2(a')\}|C. \tag{A4'}$$

Since, under NPSEMs, (A1) and (A4) can be combined into

$$Y(a, m_1) \perp\!\!\!\perp \{A, M_1\}|C$$

(cf section 5.A.2.4), which subsumes (A2), and, likewise, (A1') and (A4') can be combined into

$$Y(a, m_1, m_2) \perp\!\!\!\perp \{A, M_1, M_2\}|C,$$

which subsumes (A2'), these two assumption sets can be summarized by the assumption set V

$$\{M_1(a), M_2(a)\} \perp\!\!\!\perp A|C \tag{V1}$$
$$Y(a, m_1) \perp\!\!\!\perp \{A, M_1\}|C \tag{V2}$$
$$Y(a, m_1, m_2) \perp\!\!\!\perp \{A, M_1, M_2\}|C. \tag{V3}$$

Corresponding identification conditions from section 5.A.2.1 (under NPSEMs) are summarized in the assumption set S

$$M_1(a) \perp\!\!\!\perp A|C \tag{S1}$$
$$M_2(a, m_1) \perp\!\!\!\perp \{A, M_1\}|C \tag{S2}$$
$$Y(a, m_1, m_2) \perp\!\!\!\perp \{A, M_1, M_2\}|C. \tag{S3}$$

Although (S1) is implied by (V1), and (V3) and (S3) are identical, (S2) is not implied by assumption set V. This can be seen upon rewriting (S2) as an assumption set that consists of (S2a) and (S2b)

$$M_2(a, m_1) \perp\!\!\!\perp M_1|A, C \tag{S2a}$$
$$M_2(a, m_1) \perp\!\!\!\perp A|C. \tag{S2b}$$

(S2b) follows from (V1). This is because (V1) implies

$$M_2(a) \perp\!\!\!\perp A|M_1(a), C$$

and thus

$$M_2(a, m_1) \perp\!\!\!\perp A|M_1(a) = m_1, C,$$

which, when combined with

$$M_1(a) \perp\!\!\!\perp A|C,$$

another implication of (V1), implies

$$\{M_1(a) = m_1, M_2(a, m_1)\} \perp\!\!\!\perp A|C.$$

(S2a), on the other hand, does not follow from assumption set V. The proof follows by contradiction in that one can construct a causal diagram for which set V is satisfied, but not (S2a).

For instance, in the causal diagram in Figure 5.7, assumption set V is satisfied because there is no unmeasured confounding of (i) the effect of A on M_1, (ii) the effect of A on M_2, (iii) the joint effect of A and M_1 on Y or (iv) the joint effect of A, M_1 and M_2 on Y, nor is there intermediate confounding of (v) the effect of M_1 on Y or (vi) the joint effect of M_1 and M_2 on Y. However, (S2a) is violated since there is unmeasured confounding of the effect of M_1 on M_2 by U. In other words, V would need to be complemented with (S2a) in order to obtain all possible three-way decompositions. Alternatively, A' can be complemented with (S2a) (since it also incorporates assumptions (V1) and (V3)), leading to assumptions (v') and (vi') for NPSEMs in the main text.

It should be stressed, however, that causal diagrams such as Figure 5.7 may not reflect realistic settings in practice, since, if researchers were to possess the knowledge that M_2 does not causally affect Y, it would simply not be treated as a mediator in the first place. Consequently, differences between identification assumptions outlined by VWV (such as set V) and those put forward in section 5.A.2.1 (such as set S) may be of little practical relevance.

Nonetheless, it can similarly be shown that, in settings with k sequential media-

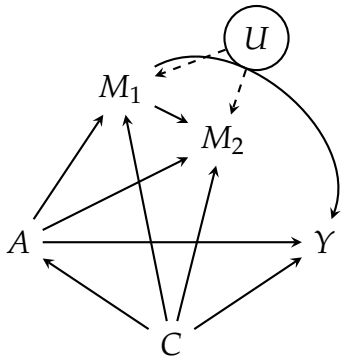


Figure 5.7: Causal diagram that satisfies assumption set V, but not assumption set S.

tors, the conditions outlined by VWV for $(k+1)$-way decompositions, would need to be complemented by

$$M_j(a, m_1, ..., m_{j-1}) \perp\!\!\!\perp \{M_1, ..., M_{j-1}\} | A, C$$

for all j from 2 to k, in order to obtain all $(k+1)!$ possible targeted decompositions.

5.A.2.6 Coping with mediator confounding

An advantage of our identification approach that is shared with the approach of VanderWeele and Vansteelandt (2013), is that, whenever in doubt about the structural dependence between certain mediators, one can simply treat them as a joint mediator in order to render identification assumptions of the corresponding component effects more plausible.

For instance, one may know that M_1 affects mediators M_2 and M_3, but the relation between M_2 and M_3 may be unclear, as depicted in the causal diagram in Figure 5.8. M_2 may affect M_3, but the causal effect may also be reverse. In addition, it may be that, irrespective of the direct causal link between M_2 and M_3, they share an unmeasured common cause U. Under these conditions, one cannot obtain a four-way decomposition as described in section 5.A.1. However, by simply treating $\{M_2, M_3\}$ as a joint mediator, one may still obtain a three-way decomposition of the total causal effect. Such strategy would, however, not work if there are additional mediators on the path between M_2 and M_3 (or between M_3

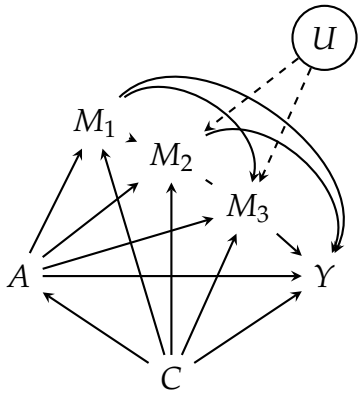


Figure 5.8: Causal diagram with three mediators. The dashed edge indicates the uncertainty of the causal relation between M_2 and M_3: M_2 may causally affect M_3 or vice versa, or they may only be related because of common cause U.

and M_2) that are explicitly taken into account in the decomposition.

5.A.3 Relation between weighted imputation and direct application of the generalized mediation formula

In this section, we provide a more detailed account of the relation between direct application of a generalization of Pearl's mediation formula (Pearl, 2001, 2012) by Monte Carlo integration (as discussed in e.g. Daniel et al. (2015) and Albert and Nelson (2011)) and our estimation procedure that relies on a combination of weighting and imputation. We first illustrate this link for three-way decompositions in settings with two sequential mediators M_1 and M_2, as in the main text. Throughout, we refer to steps and models from section 5.3.

Two-way decomposition into joint natural direct and indirect effects (with respect to the joint mediator $\{M_1, M_2\}$), conditional on baseline covariates C, involves estimation of the conditional expectation of nested counterfactuals

$$Y(a, M_1(a'), M_2(a'', M_1(a')))$$

for which $a' = a''$, i.e. $E\{Y(a, M_1(a'), M_2(a', M_1(a')))|C\}$, which, under assumptions (i')-(iv'), is non-parametrically identified by

$$\int E(Y|A = a, M_1 = m_1, M_2 = m_2, C) \times f(M_1 = m_1, M_2 = m_2|A = a', C)dm_1dm_2. \tag{5.18}$$

This expression, commonly referred to as Pearl's 'mediation formula' (Pearl, 2001, 2012), involves a form of standardization of the mean outcome in each stratum defined by mediators M_1 and M_2 and confounders C among individuals exposed at level $A = a$, to the mediator distribution of individuals exposed at level $A = a'$.

A single duplication of the original data set in step 3 (corresponding to the first two entries for each individual in Tables 5.3 and 5.4) is, in fact, sufficient to obtain such coarse two-way decomposition. Moreover, estimation of the constituent effects doesn't require weighting (i.e. as all corresponding weights in step 4 reduce to 1) and consequently, steps 1 and 4 can be omitted. Distributional and functional form assumptions for the mediators can be avoided since the suggested imputation estimator is constructed in such a way that averaging is done over the empirical distribution of the joint mediator (also see Albert, 2012). This can be seen upon noting that, in the duplicated data set, the latter corresponds to the

joint mediator distribution evaluated at $A = a'$ (as auxiliary variable a' is set equal to the observed exposure level A), and, hence, $E\{Y(a, M_1(a'), M_2(a', M_1(a')))|C\}$ can be estimated by averaging predicted outcomes $\hat{E}(Y|A = a, M_1, M_2, C)$ (based on an imputation model such as model (5.13)) in each stratum of C. As a result, this imputation strategy offers an attractive alternative to direct application of the mediation formula, which additionally relies on a model for the joint density of the mediators, as can clearly be seen upon inspecting expression (5.18).

Three-way effect decomposition as described in section 5.A.2.2 and parameterized by natural effect model (5.14) involves the conditional expectation of nested counterfactuals

$$Y(a, M_1(a'), M_2(a'', M_1(a')))$$

(for which possibly $a' \neq a''$), i.e. $E\{Y(a, M_1(a'), M_2(a'', M_1(a')))|C\}$. Under assumptions (i')-(vi'), this expectation is non-parametrically identified by

$$\int E(Y|A = a, M_1 = m_1, M_2 = m_2, C) \times f(M_1 = m_1|A = a', C) f(M_2 = m_2|A = a'', M_1 = m_1, C) dm_1 dm_2,$$

which can be rewritten as

$$\int E(Y|a, m_1, m_2, C) \frac{f(M_1 = m_1|a', C)}{f(M_1 = m_1|a'', C)} f(M_1 = m_1, M_2 = m_2|a'', C) dm_1 dm_2,$$

or

$$\int E(Y|a, m_1, m_2, C) f(M_1 = m_1, M_2 = m_2|a', C) \frac{f(M_2 = m_2|a'', m_1, C)}{f(M_2 = m_2|a', m_1, C)} dm_1 dm_2.$$

with $E(Y|a, m_1, m_2, C)$ shorthand notation for $E(Y|A = a, M_1 = m_1, M_2 = m_2, C)$. Accordingly, $E\{Y(a, M_1(a'), M_2(a'', M_1(a')))|C\}$ can be estimated from the fully extended data set by calculating a weighted average of predicted outcomes $\hat{E}(Y|A = a, M_1, M_2, C)$ (based on an imputation model such as model (5.13)) in each stratum of C, using either of the weights from step 4, depending on whether a model for M_1 or for M_2 is chosen as additional working model in step 1. Since this procedure again relies on averaging over the empirical joint distribution $f(M_1, M_2|A, C)$, data expansion is restricted to cases where either a'' or a' correspond to the observed exposure level A, depending on whether, respectively, a model for M_1 or for M_2 is chosen as additional working model in step 1, as illustrated in Tables 5.3 and 5.4 in the main text.

This estimation approach can easily be extended to settings with more than two sequential mediators. We give a short sketch of how to extend this procedure to obtain a four-way decomposition in the presence of three sequential mediators, which involves the conditional expectation of nested counterfactuals

$$Y(a, M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a'))))$$

(for which possibly $a' \neq a'' \neq a'''$), i.e.

$$E(Y(a, M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a'))))|C).$$

Under assumptions (B0)-(B3) and (C1)-(C3), this expectation is non-parametrically identified by

$$\begin{aligned}\int E(Y|a, m_1, m_2, m_3, C) f(M_1 = m_1|a', C) f(M_2 = m_2|a'', m_1, C) \\ \times f(M_3 = m_3|a''', m_1, m_2, C) dm_1 dm_2 dm_3,\end{aligned}$$

which can be rewritten as

$$\begin{aligned}&\int E(Y|a, m_1, m_2, m_3, C) \frac{f(M_1 = m_1|a', C)}{f(M_1 = m_1|a''', C)} \frac{f(M_2 = m_2|a'', m_1, C)}{f(M_2 = m_2|a''', m_1, C)} \\ &\quad \times f(M_1 = m_1, M_2 = m_2, M_3 = m_3|a''', C) dm_1 dm_2 dm_3, \\ &\int E(Y|a, m_1, m_2, m_3, C) \frac{f(M_1 = m_1|a', C)}{f(M_1 = m_1|a'', C)} \frac{f(M_3 = m_3|a''', m_1, m_2, C)}{f(M_3 = m_3|a'', m_1, m_2, C)} \\ &\quad \times f(M_1 = m_1, M_2 = m_2, M_3 = m_3|a'', C) dm_1 dm_2 dm_3,\end{aligned}$$

or

$$\begin{aligned}\int E(Y|a, m_1, m_2, m_3, C) \frac{f(M_2 = m_2|a'', m_1, C)}{f(M_2 = m_2|a', m_1, C)} \frac{f(M_3 = m_3|a''', m_1, m_2, C)}{f(M_3 = m_3|a', m_1, m_2, C)} \\ \times f(M_1 = m_1, M_2 = m_2, M_3 = m_3|a', C) dm_1 dm_2 dm_3.\end{aligned}$$

Accordingly, $E(Y(a, M_1(a'), M_2(a'', M_1(a')), M_3(a''', M_1(a'), M_2(a'', M_1(a'))))|C)$ can be estimated from the fully extended data set by calculating a weighted average of predicted outcomes $\hat{E}(Y|A = a, M_1, M_2, M_3, C)$ (based on an extended imputation model for the outcome) in each stratum of C, using weights that are calculated by combining working models for M_1 and M_2, M_1 and M_3, or M_2 and M_3, respectively.

In order to obtain a four-way decomposition, the extended data set needs to include an additional auxiliary variable a'''. Moreover, it needs to be constructed by replicating the observed data set 8 times (or sequentially duplicating it three times). The way it is constructed again depends on the choice of working models for the mediators. This can be seen upon noting that, again, this procedure relies on averaging over the empirical joint distribution of mediators $f(M_1, M_2, M_3|A, C)$, such that data expansion needs to be restricted to cases where either a''', a'' or a' correspond to the observed exposure level A, depending on whether weights are calculated by combining working models for M_1 and M_2, M_1 and M_3, or M_2 and M_3, respectively, as illustrated in Table 5.7.

More specifically, for the first duplication, four auxiliary variables need to be added:

(i) a, which, for each individual i, equals the observed exposure A_i for the first replication but the counterfactual exposure $1 - A_i$ for the second replication,

(ii) a', which equals the observed exposure for both replications,

(iii) a'', which also equals the observed exposure for both replications and

(iv) a''', which also equals the observed exposure for both replications.

If M_1 and M_2 are selected to be modelled, let a' take on counterfactual exposure level $1 - A_i$ in the second duplication and a'' take on counterfactual exposure level $1 - A_i$ in the third duplication. If M_1 and M_3 are selected to be modelled, let a' take on counterfactual exposure level $1 - A_i$ in the second duplication and a''' take on counterfactual exposure level $1 - A_i$ in the third duplication. Finally, if M_2 and M_3 are selected to be modelled, let a'' take on counterfactual exposure level $1 - A_i$ in the second duplication and a''' take on counterfactual exposure level $1 - A_i$ in the third duplication.

Note that, again, one mediator does not need to be modelled. Although this flexibility can be advantageous in settings with two sequential mediators, this advantage diminishes in settings with large number of mediators, as it only reduces a relatively small part of modeling demands.

i	A_i	a	a'	a''	a'''
1	1	1	1	1	1
	1	**0**	1	1	1
	1	1	**0**	1	1
	1	**0**	**0**	1	1
	1	1	1	**0**	1
	1	**0**	1	**0**	1
	1	1	**0**	**0**	1
	1	**0**	**0**	**0**	1
2	0	0	0	0	0
	0	**1**	0	0	0
	0	0	**1**	0	0
	0	**1**	**1**	0	0
	0	0	0	**1**	0
	0	**1**	0	**1**	0
	0	0	**1**	**1**	0
	0	**1**	**1**	**1**	0
⋮	⋮	⋮	⋮	⋮	

i	A_i	a	a'	a''	a'''
1	1	1	1	1	1
	1	**0**	1	1	1
	1	1	**0**	1	1
	1	**0**	**0**	1	1
	1	1	1	1	**0**
	1	**0**	1	1	**0**
	1	1	**0**	1	**0**
	1	**0**	**0**	1	**0**
2	0	0	0	0	0
	0	**1**	0	0	0
	0	0	**1**	0	0
	0	**1**	**1**	0	0
	0	0	0	0	**1**
	0	**1**	0	0	**1**
	0	0	**1**	0	**1**
	0	**1**	**1**	0	**1**
⋮	⋮	⋮	⋮	⋮	

i	A_i	a	a'	a''	a'''
1	1	1	1	1	1
	1	**0**	1	1	1
	1	1	1	**0**	1
	1	**0**	1	**0**	1
	1	1	1	1	**0**
	1	**0**	1	1	**0**
	1	1	1	**0**	**0**
	1	**0**	1	**0**	**0**
2	0	0	0	0	0
	0	**1**	0	0	0
	0	0	0	**1**	0
	0	**1**	0	**1**	0
	0	0	0	0	**1**
	0	**1**	0	0	**1**
	0	0	0	**1**	**1**
	0	**1**	0	**1**	**1**
⋮	⋮	⋮	⋮	⋮	

Table 5.7: Data extension when selecting working models for $f(M_1|A,C)$ and $f(M_2|A,M_1,C)$ (left), $f(M_1|A,C)$ and $f(M_3|A,M_1,M_2,C)$ (middle), $f(M_2|A,M_1,C)$ and $f(M_3|A,M_1,M_2,C)$ (right).

However, in settings without interactions between component effects, one may refrain from modeling mediator densities altogether, at the expense of additional working models for the outcome (conditional on a nested set of mediators). This adaptation is described in further detail in section 5.A.4, in which we also give a detailed step-by-step overview of the estimation approach described in section 5.3, accompanied by corresponding R code.

5.A.4 Estimation procedure

The imputation algorithm for fitting natural effect models for two-way decomposition in single mediator settings (Vansteelandt et al., 2012b) has been implemented in the R package medflex (Steen et al., 2016b), freely available from CRAN: `https://cran.r-project.org/web/packages/medflex/`.

Below, we describe how to fit natural effect models for three-way decomposition in the presence of two sequential mediators, as described in the main text.

5.A.4.1 Dichotomous exposure

In order to illustrate this procedure, we use an artificial data set of sample size $n = 1000$ simulated from a linear structural equation model represented in Figure 5.9. For the sake of illustration, we simulate from strictly linear models without interactions, so that the component effects are analytically tractable.

A binary exposure A was drawn from a binomial distribution

$$P(A = 1|C) = \text{expit}(\alpha_0 + \alpha_1 C), \tag{5.19}$$

with $\alpha = (0.25, -0.5)$, C standard normal and expit(x) = exp(x)/{1+exp(x)}. The first mediator M_1 was drawn from a normal distribution

$$f(M_1|A, C) = N(\beta_0 + \beta_1 A + \beta_2 C, \sigma^2_\beta), \tag{5.20}$$

with $\beta = (3, 1.2, 0.8)$ and $\sigma^2_\beta = 1$, as was the second mediator M_2

$$f(M_2|A, M_1, C) = N(\gamma_0 + \gamma_1 A + \gamma_2 M_1 + \gamma_3 C, \sigma^2_\gamma), \tag{5.21}$$

with $\gamma = (2, 1.6, 2, 0.9)$ and $\sigma^2_\gamma = 1$. Finally, the outcome Y was also drawn from a

normal distribution with

$$E(Y|A, M_1, M_2, C) = \delta_0 + \delta_1 A + \delta_2 M_1 + \delta_3 M_2 + \delta_4 C, \tag{5.22}$$

with $\delta = (1.6, 0.4, 0.6, 1.2, 1.4)$ and $\sigma_\delta^2 = 1$. This linear SEM yields component effects of size

$$\begin{aligned} E_{A\to Y}(a', a'') &= \delta_1 = 0.4 \\ E_{A\to M_1 Y}(a, a'') &= \beta_1(\delta_2 + \gamma_2\delta_3) = 1.2(0.6 + 2 \times 1.2) = 3.6 \\ E_{A\to M_2\to Y}(a, a') &= \gamma_1\delta_3 = 1.6 \times 1.2 = 1.92 \end{aligned}$$

for all a, a' and a''.

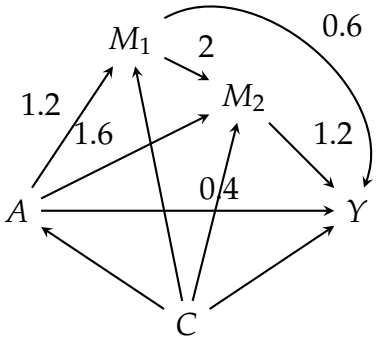


Figure 5.9: Causal diagram representing the causal data generating mechanism of the simulated data set.

First, the data set is simulated in R.

```
expit <- function(x) exp(x)/(1+exp(x))

n  <- 10^3
C  <- rnorm(n)
A  <- rbinom(n, size = 1, prob = expit(0.25 - 0.5*C))
M1 <- rnorm(n, mean = 3 + 1.2*A + 0.8*C, sd = 1)
M2 <- rnorm(n, mean = 2 + 1.6*A + 2*M1 + 0.9*C, sd = 1)
Y  <- rnorm(n, mean = 1.6 + 0.4*A + 0.6*M1 + 1.2*M2 + 1.4*C, sd = 1)

dat <- data.frame(id = 1:n, C, A, M1, M2, Y)
head(dat)
```

```
  id      C A   M1    M2    Y
1  1 -1.802 1 2.69  7.23 10.5
2  2  0.316 0 3.94  9.38 15.4
3  3 -0.560 0 2.54  8.21 13.2
4  4  0.524 1 5.48 14.33 23.7
5  5  0.544 0 4.38 11.62 16.9
6  6 -0.664 0 1.82  6.24  8.4
```

Next, we follow steps 1-6 as described in section 5.3. Since we are ignorant as to the functional form of the working models, we should ideally do some model building at this stage. For the sake of illustration, however, we fit models of the form of models (5.19), (5.20), (5.21) and (5.22), leaving out interactions or polynomials involving C, but including interactions between A, M_1 and M_2 in order to ensure that possible differences between decompositions parameterized in the final natural effect model appropriately reflect differences dictated by the data.

1. Fit a suitable model for the probability (density) of (one of) the mediators conditional on A, potential earlier intermediates and C.

```
fitM1 <- glm(M1 ~ A + C,
             family = gaussian("identity"), data = dat)
fitM2 <- glm(M2 ~ A * M1 + C,
             family = gaussian("identity"), data = dat)
```

2. Fit a suitable model for the outcome mean conditional on A, M_1, M_2 and C.

```
fitY <- glm(Y ~ A * M1 * M2 + C,
            family = gaussian("identity"), data = dat)
```

3. Construct the extended dataset. Below we illustrate exactly how the suggested replication procedure in section 5.3 (step 3) can be implemented in R. Let `a0`, `a1` and `a2` correspond to a, a' and a'' in the main text. Furthermore, let `extdat1` and `extdat2` correspond to the extended data sets that are extended according to whether a model for the density of M_1 or M_2 is selected as working model, respectively. In practice, only one of the extended data sets is needed.

```r
# create auxiliary variables
dat$a0 <- dat$A
dat$a1 <- dat$A
dat$a2 <- dat$A

# first duplication
# i.e. extract rownames of the original dataset
# and use two replicates of each rowname as indices to duplicate
extdat1 <- dat[rep(rownames(dat), times = 2), ]
extdat2 <- dat[rep(rownames(dat), times = 2), ]

# create duplication indicator
extdat1$dup <- rep(1:2, each = nrow(dat))
extdat2$dup <- rep(1:2, each = nrow(dat))

# let a0 take on counterfactual exposure level 1-A for the second duplicate
extdat1$a0 <- ifelse(extdat1$dup == 2, 1-extdat1$A, extdat1$A)
extdat2$a0 <- ifelse(extdat2$dup == 2, 1-extdat2$A, extdat2$A)

# second duplication
extdat1 <- extdat1[rep(rownames(extdat1), times = 2), ]
extdat2 <- extdat2[rep(rownames(extdat2), times = 2), ]

# create updated duplication indicator
extdat1$dup <- rep(1:2, each = 2*nrow(dat))
extdat2$dup <- rep(1:2, each = 2*nrow(dat))

# let a1 or a2 take on counterfactual exposure level 1-A for the second duplicate
extdat1$a1 <- ifelse(extdat1$dup == 2, 1-extdat1$A, extdat1$A)
extdat2$a2 <- ifelse(extdat2$dup == 2, 1-extdat2$A, extdat2$A)

# order by id
extdat1 <- extdat1[order(extdat1$id), ]
extdat2 <- extdat2[order(extdat2$id), ]

# check the result
head(extdat1, 4)
```

```
      id    C A   M1   M2    Y a0 a1 a2 dup
1      1 -1.8 1 2.69 7.23 10.5  1  1  1   1
1.1    1 -1.8 1 2.69 7.23 10.5  0  1  1   1
```

```
1.2    1 -1.8 1 2.69 7.23 10.5  1  0  1   2
1.1.1  1 -1.8 1 2.69 7.23 10.5  0  0  1   2
```

```
head(extdat2, 4)
```

```
      id    C A   M1   M2    Y a0 a1 a2 dup
1      1 -1.8 1 2.69 7.23 10.5  1  1  1   1
1.1    1 -1.8 1 2.69 7.23 10.5  0  1  1   1
1.2    1 -1.8 1 2.69 7.23 10.5  1  1  0   2
1.1.1  1 -1.8 1 2.69 7.23 10.5  0  1  0   2
```

Alternatively, one can use the `expand.grid` function, as explained below. The procedure below yields the exact same result as above. Although more technical, it allows us to more easily generalize data extension to categorical or continuous exposures (as discussed in section 5.A.4.4).

```
# remove auxiliary variables from the original data set
dat$a0 <- dat$a1 <- dat$a2 <- NULL

# create extended data set with 4 replicates for each subject
extdat1 <- dat[rep(dat$id, each = 4), ]
extdat2 <- dat[rep(dat$id, each = 4), ]
```

The rationale behind using the `expand.grid` function is that, for each subject, we want to obtain all possible pairwise combinations of hypothetical exposure levels a and a' (when weighting for M_1) or a and a'' (when weighting for M_2). Therefore, we need to enumerate all possible exposure levels for each subject (starting with the observed exposure level A_i, followed by counterfactual exposure level $1 - A_i$) and pass this to the `expand.grid` function, as illustrated below for the first subject.

```
levels <- c(dat$A[1], 1-dat$A[1])
expand.grid(levels, levels)
```

```
  Var1 Var2
1    1    1
2    0    1
3    1    0
4    0    0
```

The same result can more generally be obtained by putting `levels` in a list and repeating it twice.

```
expand.grid(rep(list(levels), 2))
```

```
  Var1 Var2
1    1    1
2    0    1
3    1    0
4    0    0
```

However, as we need to do this for each subject (`dat$id`), we pass this to the `lapply` function and stack the resulting matrices using the `rbind` function.

```
tmp <- lapply(dat$id, function(x) expand.grid(rep(list(c(dat$A[x], 1-dat$A[x])), 2)))
tmp <- do.call(rbind, tmp)

# check result
head(tmp)
```

```
  Var1 Var2
1    1    1
2    0    1
3    1    0
4    0    0
5    0    0
6    1    0
```

The resulting stacked matrices can now be merged with the extended data sets, and `Var1` and `Var2` can be renamed as `a0` and `a1` (or `a2`), respectively. A third auxiliary variable `a2` (or `a1`) needs to be added that is a copy of the observed exposure level A_i.

```
extdat1 <- data.frame(extdat1, a0 = tmp$Var1, a1 = tmp$Var2, a2 = extdat1$A)
extdat2 <- data.frame(extdat2, a0 = tmp$Var1, a1 = extdat2$A, a2 = tmp$Var2)

# check the result
head(extdat1, 4)
```

```
    id    C A   M1   M2    Y a0 a1 a2
1    1 -1.8 1 2.69 7.23 10.5  1  1  1
```

```
1.1  1 -1.8 1 2.69 7.23 10.5  0  1  1
1.2  1 -1.8 1 2.69 7.23 10.5  1  0  1
1.3  1 -1.8 1 2.69 7.23 10.5  0  0  1
```

```
head(extdat2, 4)
```

```
    id    C A   M1   M2    Y a0 a1 a2
1    1 -1.8 1 2.69 7.23 10.5  1  1  1
1.1  1 -1.8 1 2.69 7.23 10.5  0  1  1
1.2  1 -1.8 1 2.69 7.23 10.5  1  1  0
1.3  1 -1.8 1 2.69 7.23 10.5  0  1  0
```


4. Calculate regression weights W_{1i} or W_{2i} depending on whether a working model for M_1 or M_2 is selected. Since both mediators are normally distributed in the simulated data set, we use the `dnorm` function in order to obtain the densities $f(M_1 = M_{1i}|A = a', C_i)$, $f(M_1 = M_{1i}|A = A_i, C_i)$, $f(M_2 = M_{2i}|A = a'', M_{1i}, C_i)$ and $f(M_2 = M_{2i}|A = A_i, M_1 = M_{1i}, C_i)$.

```
# calculate W1
meanM1a1 <- predict(fitM1, newdata = data.frame(A = extdat1$a1, extdat1),
                    type = "response")
meanM1A <- predict(fitM1, newdata = data.frame(A = extdat1$A, extdat1),
                   type = "response")
sdM1 <- sqrt(summary(fitM1)$dispersion)

num1 <- dnorm(extdat1$M1, mean = meanM1a1, sd = sdM1)
denom1 <- dnorm(extdat1$M1, mean = meanM1A, sd = sdM1)
extdat1$W1 <- num1/denom1

# calculate W2
meanM2a2 <- predict(fitM2, newdata = data.frame(A = extdat2$a2, extdat2),
                    type = "response")
meanM2A <- predict(fitM2, newdata = data.frame(A = extdat2$A, extdat2),
                   type = "response")
sdM2 <- sqrt(summary(fitM2)$dispersion)

num2 <- dnorm(extdat2$M2, mean = meanM2a2, sd = sdM2)
denom2 <- dnorm(extdat2$M2, mean = meanM2A, sd = sdM2)
extdat2$W2 <- num2/denom2
```

```
# check result
head(extdat1, 4)
```

```
    id    C A   M1   M2    Y a0 a1 a2    W1
1    1 -1.8 1 2.69 7.23 10.5  1  1  1 1.000
1.1  1 -1.8 1 2.69 7.23 10.5  0  1  1 1.000
1.2  1 -1.8 1 2.69 7.23 10.5  1  0  1 0.413
1.3  1 -1.8 1 2.69 7.23 10.5  0  0  1 0.413
```

```
head(extdat2, 4)
```

```
    id    C A   M1   M2    Y a0 a1 a2    W2
1    1 -1.8 1 2.69 7.23 10.5  1  1  1 1.000
1.1  1 -1.8 1 2.69 7.23 10.5  0  1  1 1.000
1.2  1 -1.8 1 2.69 7.23 10.5  1  1  0 0.257
1.3  1 -1.8 1 2.69 7.23 10.5  0  1  0 0.257
```


5. Impute nested counterfactuals $Y(a, M_1(a'), M_2(a'', M_1(a')))$ by $\hat{E}\{Y|A = a, M_1, M_2, C\}$ using predict-functionality.

```
extdat1$Y <- predict(fitY, newdata = data.frame(A = extdat1$a0, extdat1),
                     type = "response")
extdat2$Y <- predict(fitY, newdata = data.frame(A = extdat2$a0, extdat2),
                     type = "response")
```

6. Fit a natural effect model for $E\{Y(a, M_1(a'), M_2(a'', M_1(a')))|C\}$ by regressing imputed counterfactuals $\hat{E}\{Y|A = a, M_1, M_2, C\}$ on a, a', a'' and C, upon weighting for either W_1 or W_2.

```
fitNEM1 <- glm(Y ~ a0 * a1 * a2 + C,
               family = gaussian("identity"), data = extdat1, weights = W1)
fitNEM2 <- glm(Y ~ a0 * a1 * a2 + C,
               family = gaussian("identity"), data = extdat2, weights = W2)

# obtain parameter estimates
fitNEM1
```

```
Call:  glm(formula = Y ~ a0 * a1 * a2 + C, family = gaussian("identity"),
    data = extdat1, weights = W1)

Coefficients:
(Intercept)           a0           a1           a2            C
    12.5898       0.4440       3.6483       2.1909       5.1121
      a0:a1        a0:a2        a1:a2     a0:a1:a2
     0.1576      -0.0717       0.1437      -0.0027

Degrees of Freedom: 3999 Total (i.e. Null);  3991 Residual
Null Deviance:     159000
Residual Deviance: 34700  AIC: 22000
```



```
fitNEM2


Call:  glm(formula = Y ~ a0 * a1 * a2 + C, family = gaussian("identity"),
    data = extdat2, weights = W2)

Coefficients:
(Intercept)           a0           a1           a2            C
    12.5475       0.4440       4.0046       2.1579       5.2925
      a0:a1        a0:a2        a1:a2     a0:a1:a2
     0.1676      -0.0209      -0.0967      -0.0635

Degrees of Freedom: 3999 Total (i.e. Null);  3991 Residual
Null Deviance:     145000
Residual Deviance: 36500  AIC: 22900
```

Moreover, population-average component effects (rather than effects conditional on C) can be obtained upon multiplying weights by $1/\hat{P}(A = A_i|C_i)$ for each subject in the extended data set and fitting a population-average natural effect model for
$E\{Y(a, M_1(a'), M_2(a'', M_1(a')))\}$, as illustrated below.

First, a model for the probability of exposure needs to be fitted on the original data set.

```
fitA <- glm(A ~ C,
            family = binomial("logit"), data = dat)
```

Next, updated weights need to be calculated.

```
meanA1 <- predict(fitA, newdata = extdat1, type = "response")
meanA2 <- predict(fitA, newdata = extdat2, type = "response")

extdat1$W1 <- extdat1$W1 / dbinom(extdat1$A, size = 1, prob = meanA1)
extdat2$W2 <- extdat2$W2 / dbinom(extdat2$A, size = 1, prob = meanA2)
```

Finally, the population-average natural effect model can be fitted.

```
fitNEM1 <- glm(Y ~ a0 * a1 * a2,
               family = gaussian("identity"), data = extdat1, weights = W1)
fitNEM2 <- glm(Y ~ a0 * a1 * a2,
               family = gaussian("identity"), data = extdat2, weights = W2)

# obtain parameter estimates
fitNEM1


Call:  glm(formula = Y ~ a0 * a1 * a2, family = gaussian("identity"),
    data = extdat1, weights = W1)

Coefficients:
(Intercept)           a0           a1           a2        a0:a1
    12.5472       0.4189       4.2108       2.2961       0.1619
      a0:a2        a1:a2     a0:a1:a2
    -0.0235      -0.6336      -0.0206

Degrees of Freedom: 3999 Total (i.e. Null);  3992 Residual
Null Deviance:     321000
Residual Deviance: 281000  AIC: 27800

fitNEM2


Call:  glm(formula = Y ~ a0 * a1 * a2, family = gaussian("identity"),
    data = extdat2, weights = W2)

Coefficients:
(Intercept)           a0           a1           a2        a0:a1
    12.5472       0.4189       2.5063       2.2323       0.2010
```

```
      a0:a2        a1:a2     a0:a1:a2
   -0.0166       1.1347      -0.0666

Degrees of Freedom: 3999 Total (i.e. Null);  3992 Residual
Null Deviance:     305000
Residual Deviance: 269000  AIC: 28300
```

5.A.4.2 Obtaining bootstrap-based standard errors and confidence intervals

In this section, we demonstrate how to obtain bootstrap-based standard errors and confidence intervals using the boot library in R.

We basically need to wrap the above code in a function, say `bootFun`, which has arguments `data` and `index`. Within that function, we need to add a line that makes sure that the analysis is done on bootstrapped samples `dat` and request to return the estimated coefficients of the natural effect model. For simplicity, we restrict our presentation to the bootstrap for natural effect models weighted by the density of M_1, although the procedure can easily be adopted to obtain bootstrap-based inference for natural effect models weighted by the density of M_2.

```r
library(boot)

bootFun <- function(data, index) {

  dat <- data[index, ]

  # 1
  fitM1 <- glm(M1 ~ A + C,
               family = gaussian("identity"), data = dat)

  # 2
  fitY <- glm(Y ~ A * M1 * M2 + C,
              family = gaussian("identity"), data = dat)

  # 3
  extdat <- dat[rep(dat$id, each = 4), ]
  tmp <- lapply(dat$id, function(x) expand.grid(rep(list(c(dat$A[x], 1-dat$A[x])), 2)))
  tmp <- do.call(rbind, tmp)
```

```
  extdat <- data.frame(extdat, a0 = tmp$Var1, a1 = tmp$Var2, a2 = extdat$A)

  # 4
  meanM1a1 <- predict(fitM1, newdata = data.frame(A = extdat$a1, extdat),
                      type = "response")
  meanM1A <- predict(fitM1, newdata = data.frame(A = extdat$A, extdat),
                     type = "response")
  sdM1 <- sqrt(summary(fitM1)$dispersion)

  num <- dnorm(extdat$M1, mean = meanM1a1, sd = sdM1)
  denom <- dnorm(extdat$M1, mean = meanM1A, sd = sdM1)
  extdat$W1 <- num/denom

  # 5
  extdat$Y <- predict(fitY, newdata = data.frame(A = extdat$a0, extdat),
                      type = "response")

  # 6
  fitNEM <- glm(Y ~ a0 * a1 * a2 + C,
                family = gaussian("identity"), data = extdat, weights = W1)

  return(coef(fitNEM))
}
```

Once this function is defined, apply the `boot` function as illustrated below. In order to obtain an acceptable level of precision, use a sufficient number of bootstrap samples `R` (i.e. at least 1000).

```
bootSE <- boot(data = dat, statistic = bootFun, R = 10)
```

Bootstrap-based standard errors can then readily be obtained by printing `bootSE`

```
bootSE
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
```

```
boot(data = dat, statistic = bootFun, R = 10)
```

```
Bootstrap Statistics :
    original   bias    std. error
t1*  12.5898  0.00823      0.2536
t2*   0.4440 -0.00582      0.0700
t3*   3.6483 -0.07565      0.4368
t4*   2.1909  0.11941      0.4282
t5*   5.1121  0.08345      0.2668
t6*   0.1576 -0.01628      0.0722
t7*  -0.0717 -0.00805      0.0782
t8*   0.1437  0.07422      0.5816
t9*  -0.0027  0.01361      0.0382
```


Moreover, bootstrap samples of the parameter estimates are stored in `bootSE$t`

```
head(bootSE$t)
```

```
     [,1]  [,2] [,3] [,4] [,5]   [,6]     [,7]   [,8]     [,9]
[1,] 12.5 0.386 3.91 2.74 5.48 0.1517 -0.05853 -0.410  0.00709
[2,] 12.3 0.332 2.97 2.01 4.98 0.0963  0.01095  0.967  0.03485
[3,] 12.5 0.529 3.78 3.01 5.11 0.1586 -0.03573 -0.670 -0.03323
[4,] 12.5 0.450 3.68 1.86 4.79 0.0396 -0.17524  0.530  0.02201
[5,] 12.5 0.500 3.68 2.64 5.53 0.0289  0.00263  0.318  0.06715
[6,] 13.0 0.328 2.64 1.76 5.03 0.1614 -0.15113  1.228  0.03482
```

```
apply(bootSE$t, 2, sd)
```

```
[1] 0.2536 0.0700 0.4368 0.4282 0.2668 0.0722 0.0782 0.5816 0.0382
```

95% bootstrap-based normal approximation confidence intervals can be obtained by running the code below.

```
sapply(1:length(bootSE$t0), FUN = function(x)
  boot.ci(bootSE, conf = 0.95, type = "norm", index = x)$normal[2:3])
```

```
     [,1]  [,2] [,3] [,4] [,5]   [,6]    [,7]  [,8]    [,9]
[1,] 12.1 0.313 2.87 1.23 4.51 0.0323 -0.2169 -1.07 -0.0912
[2,] 13.1 0.587 4.58 2.91 5.55 0.3155  0.0896  1.21  0.0586
```

Obtaining confidence intervals for linear combinations of the parameter estimates of the natural effect models requires some additional tweaking. For this purpose, the function `linfunCI` can be used from the code below.

```
# function to calculate 95% confidence intervals
# for linear combination of parameter estimates
linfunCI <- function(boot.out, L, conf) {
  est <- sum(L %*% boot.out$t0)
  se <- diag(sqrt(t(L) %*% var(boot.out$t) %*% L))
  bias <- est - L %*% colMeans(boot.out$t)

  CI <- est + bias + c(-1,1) * qnorm(1-(1-conf)/2) * se

  return(c("LCL" = CI[1], "UCL" = CI[2]))
}

# specify contrast matrix
# e.g. for natural indirect effect wrt M1
# (refer to Table 5.1
# for parameterization)
L <- c(0, 0, 1, 0, 0, 1, 0, 1, 1)

# obtain linear combination of parameter estimates
# and corresponding 95% confidence interval
c(L %*% bootSE$t0, linfunCI(bootSE, L, 0.95))

      LCL  UCL
3.95 3.32 4.58
```

Finally, hypothesis testing can be done using a multivariate Wald-type test based on the bootstrap normal approximation, implemented in the `bootChisq` function. Below, this function is illustrated for testing differences between decompositions (i.e. testing whether the parameters capturing interactions between a, a' and a'' jointly differ from 0) by appropriately specifying the contrast matrix `L`.

```
# function for Wald-type Chisquare test
# based on the bootstrap covariance matrix
bootChisq <- function(boot.out, L) {
  chisq <- t(boot.out$t0) %*% t(L) %*% solve(L %*% var(boot.out$t)
                                          %*% t(L)) %*% L %*% boot.out$t0
```

```
  df <- dim(L)[1]
  p <- pchisq(q = chisq, df, lower.tail = FALSE)
  return(c("Chisq" = chisq, "df" = df, "p" = p))
}

L <- matrix(c(0, 0, 0, 0, 0, 1, 0, 0, 0,
              0, 0, 0, 0, 0, 0, 1, 0, 0,
              0, 0, 0, 0, 0, 0, 0, 1, 0,
              0, 0, 0, 0, 0, 0, 0, 0, 1),
            nrow = 4, byrow = TRUE)

bootChisq(bootSE, L)

Chisq    df     p
7.526 4.000 0.111
```


5.A.4.3 Adapted approach in the absence of interactions

Whenever differences between decompositions are ignorable (i.e. interactions between a, a' and a'' in a natural effect model are close to zero), one can opt to refit a natural effect model that excludes the corresponding interaction terms to the extended data (as was done in the application in the main text). Particularly in settings with a large number of mediators, this may yield a considerable gain in precision.

i	A_i	a	a'	a''	$Y_i(a, M_{1i}(a'), M_{2i}(a'', M_{1i}(a')))$	
1	1	1	1	1	$Y_1(1, M_{11}(1), M_{21}(1, M_{11}(1)))$	$= Y_1$
	1	**0**	1	1	$Y_1(\mathbf{0}, M_{11}(1), M_{21}(1, M_{11}(1)))$	$= Y_1(\mathbf{0}, M_{11}, M_{21})$
	1	**0**	1	**0**	$Y_1(\mathbf{0}, M_{11}(1), M_{21}(\mathbf{0}, M_{11}(1)))$	$= Y_1(\mathbf{0}, M_{11})$
2	0	0	0	0	$Y_2(0, M_{12}(0), M_{22}(0, M_{12}(0)))$	$= Y_2$
	0	**1**	0	0	$Y_2(\mathbf{1}, M_{12}(0), M_{22}(0, M_{12}(0)))$	$= Y_2(\mathbf{1}, M_{12}, M_{22})$
	0	**1**	0	**1**	$Y_2(\mathbf{1}, M_{12}(0), M_{22}(\mathbf{1}, M_{12}(0)))$	$= Y_2(\mathbf{1}, M_{12})$
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 5.8: Adapted data extension in the absence of interactions.

Alternatively, in the absence of interactions, one may resort to a completely different approach. That is, in this specific setting, main effects of a, a' and a'' can be estimated based on the nested counterfactuals listed in Table 5.8. That is, the main effect of a can be obtained by contrasting the first and second line within each subject. The main effect of a' can be obtained by contrasting the first (last) line of exposed subjects ($A = 1$) with the last (first) line of unexposed subjects ($A = 0$). Finally, the main effect of a'' can be obtained by contrasting the second and third line within each subject. However, again, these nested counterfactuals are only observed for subjects i if $A_i = a = a' = a''$. As mentioned in section 5.A.3, the first two nested counterfactuals for each subject can be predicted by $\hat{E}\{Y_i|A = a, M_{1i}, M_{2i}, C_i\}$ based on an imputation model for $E\{Y|A, M_1, M_2, C\}$ and hence requires no weights and thus no working model for the density of M_1 or M_2. Furthermore, relying on the composition assumption that $Y(a, M_1(a'), M_2(a, M_1(a'))) = Y(a, M_1(a'))$, the third nested counterfactual can also be predicted by $\hat{E}\{Y_i|A = a, M_{1i}, C_i\}$ based on an imputation model for $E\{Y|A, M_1, C\}$.[7]

This notion leads to fitting strategies for (conditional) natural effect models that are exclusively based on imputation. Hence, in the absence of interactions between component effects, one may avoid modeling mediator densities at the expense of an additional (imputation) model for $E\{Y|A, M_1, C\}$. Such approach has already been suggested by VanderWeele et al. (2014) as an alternative to fully parametric estimation of component effects via a sequential decomposition approach. Note, however, that for non-linear models, this may induce additional concerns for incongeniality between models for $E\{Y|A, M_1, M_2, C\}$ and $E\{Y|A, M_1, C\}$.

Below, we illustrate how this approach can be implemented in R (on the simulated data set), given some minor modifications of the code presented above.

1. Instead of fitting a model for the probability (density) of (one of) the mediators, fit a suitable model for the outcome mean conditional on A, M_1 and C.

```
fitY1 <- glm(Y ~ A * M1 + C,
             family = gaussian("identity"), data = dat)
```

[7] Note that assumptions (i')-(vi') enable obtaining all six possible three-way decompositions, including the one that involves the natural indirect effect with respect to M_1, i.e. $E_{A\to M_1 Y}(1, 1)$. As identification of this effect relies on identification of $E\{Y(a, M_1(a'))|C\}$, this expected nested counterfactual is then also identified. It can moreover be estimated by averaging predicted outcomes $\hat{E}\{Y|A = a, M_1, C\}$ in each stratum of C.

2. Fit a suitable model for the outcome mean conditional on A, M_1, M_2 and C.

```
fitY2 <- glm(Y ~ A * M1 * M2 + C,
             family = gaussian("identity"), data = dat)
```

3. Construct an extended data set *as if* one were to select a model for the density of M_2 as a working model, but for each subject leaving out the observation for which $a'' \neq a = a'$ (see Figure 5.4).

```
extdat <- dat[rep(dat$id, each = 4), ]
tmp <- lapply(dat$id, function(x) expand.grid(rep(list(c(dat$A[x], 1-dat$A[x])), 2)))
tmp <- do.call(rbind, tmp)
extdat <- data.frame(extdat, a0 = tmp$Var1, a1 = extdat$A, a2 = tmp$Var2)
extdat <- extdat[extdat$a2 == extdat$a0 | extdat$a2 == extdat$a1, ]
head(extdat)
```

```
    id      C A   M1   M2    Y a0 a1 a2
1    1 -1.802 1 2.69 7.23 10.5  1  1  1
1.1  1 -1.802 1 2.69 7.23 10.5  0  1  1
1.3  1 -1.802 1 2.69 7.23 10.5  0  1  0
2    2  0.316 0 3.94 9.38 15.4  0  0  0
2.1  2  0.316 0 3.94 9.38 15.4  1  0  0
2.3  2  0.316 0 3.94 9.38 15.4  1  0  1
```

4. Skip step 4.

5. Impute nested counterfactuals $Y(a, M_1(a'), M_2(a'', M_1(a')))$ for which $a' = a''$ by $\hat{E}\{Y|A = a, M_1, M_2, C\}$ (`fitY2`) and nested counterfactuals for which $a' \neq a''$ by $\hat{E}\{Y|A = a, M_1, C\}$ (`fitY1`).

```
ind <- which(extdat$a1 != extdat$a2)

extdat$Y <- predict(fitY2, newdata = data.frame(A = extdat$a0, extdat),
                    type = "response")
extdat$Y[ind] <- predict(fitY1, newdata = data.frame(A = extdat$a0, extdat)[ind, ],
                         type = "response")
```

6. Fit a natural effect model for $E\{Y(a, M_1(a'), M_2(a'', M_1(a')))|C\}$ by means of an unweighted regression of imputed outcomes on a, a', a'' and C, *excluding* all interactions between a, a' and a''.

```
fitNEM <- glm(Y ~ a0 + a1 + a2 + C,
              family = gaussian("identity"), data = extdat)
fitNEM
```

```
Call:  glm(formula = Y ~ a0 + a1 + a2 + C, family = gaussian("identity"),
    data = extdat)

Coefficients:
(Intercept)           a0           a1           a2            C
     12.594        0.491        3.977        1.994        4.995

Degrees of Freedom: 2999 Total (i.e. Null);  2995 Residual
Null Deviance:      103000
Residual Deviance: 26500  AIC: 15100
```

5.A.4.4 Continuous exposure

Vansteelandt et al. (2012b) suggested a modification of their imputation algorithm for continuous exposures in the single mediator case. Applying this modification to the procedure described in section 5.A.4.1 (for three-way decomposition in the presence of two sequential mediators) mainly entails some slight change to the data extension step (step 3 in section 5.3). Below, we illustrate how to implement this modified procedure in R, using a new artificial data set simulated from a linear SEM. The data-generating SEM is identical to the one used to simulate the data set in section 5.A.4.1, except that the exposure A is now drawn from a normal distribution

$$f(A|C) = N(\alpha'_0 + \alpha'_1 C, \sigma^2_{\alpha'}), \tag{5.23}$$

with $\alpha' = (0.25, 0.5)$, $\sigma^2_{\alpha'} = 1$ and C standard normal.

```
C  <- rnorm(n)
A  <- rnorm(n, mean = 0.25 + 0.5*C, sd = 1)
M1 <- rnorm(n, mean = 3 + 1.2*A + 0.8*C, sd = 1)
```

```
M2 <- rnorm(n, mean = 2 + 1.6*A + 2*M1 + 0.9*C, sd = 1)
Y  <- rnorm(n, mean = 1.6 + 0.4*A + 0.6*M1 + 1.2*M2 + 1.4*C, sd = 1)

dat <- data.frame(id = 1:n, C, A, M1, M2, Y)
head(dat)
```

```
  id      C      A     M1      M2     Y
1  1 -1.224 -1.331  0.718 -0.0809  1.32
2  2 -0.117 -1.279 -0.724 -1.5177 -2.97
3  3  1.215  1.222  5.449 14.6282 26.09
4  4 -0.524 -0.431  2.215  6.5245 10.70
5  5 -2.018 -2.282 -0.449 -4.5822 -7.28
6  6 -0.312 -0.493  1.660  2.7780  5.07
```


Steps 1 and 2 remain unchanged. For simplicity, we restrict our presentation to natural effect models weighted by the density of M_1, although the procedure can easily be adopted to natural effect models weighted by the density of M_2.

```
fitM1 <- glm(M1 ~ A + C,
             family = gaussian("identity"), data = dat)
fitY <- glm(Y ~ A * M1 * M2 + C,
            family = gaussian("identity"), data = dat)
```

In addition, fit a model for the density of the exposure A, given C.

```
fitA <- glm(A ~ C,
            family = gaussian("identity"), data = dat)
```

In step 3, construct an extended data set by sequentially replicating the original data set J (instead of 2) times. Again, three auxiliary variables a, a' and a'' need to be created. For the first-stage replications, let a take on the observed exposure level A for the first replication and different quantiles from the conditional exposure distribution $f(A|C)$ for the remaining $J-1$ replications. Ideally, this should be equally spaced quantiles extending the whole range of the distribution, e.g. if J is chosen to be 3, select the 10% and 90% percentiles. Quantiles from the normal distribution for each subject can be obtained by using the `qnorm` function, as illustrated in the code below. Each line i of q contains the 10% and 90% percentiles from $f(A|C)$ for subject i (with $C = C_i$).

```
meanA <- predict(fitA, type = "response")
sdA <- sqrt(summary(fitA)$dispersion)

q <- sapply(c(0.1, 0.9), FUN = qnorm, mean = meanA, sd = sdA)

# check result
head(q)
```

```
    [,1]  [,2]
1 -1.610 0.888
2 -1.053 1.445
3 -0.383 2.116
4 -1.258 1.241
5 -2.010 0.489
6 -1.151 1.347
```


Moreover, let both a' and a'' take on the observed exposure level A. For the second-stage replications (i.e. replications of the resulting replicated data set from the first-stage replication), let either a' or a'' take on a different quantile from $f(A|C)$ for each of the additional $J-1$ replications, depending on whether a model for the density of M_1 or M_2 is selected as a working model.

Although this replication procedure may seem daunting to apply, as the R code below illustrates, one can essentially use the same code for data extension as in section 5.A.4.1. If J is chosen to be 3, we will obtain in total $3^2 = 9$ replicates (instead of $2^2 = 4$ for a dichotomous exposure) when replicating the original data twice by a factor $J = 3$. More generally, in order to obtain a $(k+1)$-way decomposition in the presence of k sequential mediators, we need to replicate the original data k times by a factor J (with J the total number of exposure levels in case of dichotomous or polytomous exposures), leading to a total of J^k replicates of the original data set.

```
extdat <- dat[rep(dat$id, each = 9), ]

tmp <- lapply(dat$id, function(x) expand.grid(rep(list(c(dat$A[x], q[x, ])), 2)))
tmp <- do.call(rbind, tmp)

extdat <- data.frame(extdat, a0 = tmp$Var1, a1 = tmp$Var2, a2 = extdat$A)

# check result
head(extdat, 9)
```

```
    id     C     A    M1      M2    Y     a0     a1    a2
1    1 -1.22 -1.33 0.718 -0.0809 1.32 -1.331 -1.331 -1.33
1.1  1 -1.22 -1.33 0.718 -0.0809 1.32 -1.610 -1.331 -1.33
1.2  1 -1.22 -1.33 0.718 -0.0809 1.32  0.888 -1.331 -1.33
1.3  1 -1.22 -1.33 0.718 -0.0809 1.32 -1.331 -1.610 -1.33
1.4  1 -1.22 -1.33 0.718 -0.0809 1.32 -1.610 -1.610 -1.33
1.5  1 -1.22 -1.33 0.718 -0.0809 1.32  0.888 -1.610 -1.33
1.6  1 -1.22 -1.33 0.718 -0.0809 1.32 -1.331  0.888 -1.33
1.7  1 -1.22 -1.33 0.718 -0.0809 1.32 -1.610  0.888 -1.33
1.8  1 -1.22 -1.33 0.718 -0.0809 1.32  0.888  0.888 -1.33
```

The remaining steps (4-6) again remain unchanged.


```
meanM1a1 <- predict(fitM1, newdata = data.frame(A = extdat$a1, extdat),
                    type = "response")
meanM1A <- predict(fitM1, newdata = data.frame(A = extdat$A, extdat),
                   type = "response")
sdM1 <- sqrt(summary(fitM1)$dispersion)

num <- dnorm(extdat$M1, mean = meanM1a1, sd = sdM1)
denom <- dnorm(extdat$M1, mean = meanM1A, sd = sdM1)
extdat$W1 <- num/denom

extdat$Y <- predict(fitY, newdata = data.frame(A = extdat$a0, extdat),
                    type = "response")

fitNEM <- glm(Y ~ a0 * a1 * a2 + C,
              family = gaussian("identity"), data = extdat, weights = W1)

# check result
fitNEM


Call:  glm(formula = Y ~ a0 * a1 * a2 + C, family = gaussian("identity"),
    data = extdat, weights = W1)

Coefficients:
(Intercept)           a0           a1           a2            C
   12.81429      0.40228      3.12903      2.31652      5.15891
      a0:a1        a0:a2        a1:a2     a0:a1:a2
   -0.05298      0.12327     -0.00664     -0.03416
```

```
Degrees of Freedom: 8999 Total (i.e. Null);  8991 Residual
Null Deviance:     856000
Residual Deviance: 76000  AIC: 55400
```

Note that misspecification of the sampling model for $f(A|C)$ does not induce bias in the estimated coefficients and standard errors of the natural effect model. Moreover, we recommend to use a minimum of $J = 3$ draws. Although finite sample bias and sampling variability can be reduced to some extent by choosing a larger number of draws, simulations have shown this gain to be ignorable when choosing more than $J = 5$ draws.

Moreover, the above R code can easily be adopted to also accommodate polytomous exposures (with $J > 2$ exposure levels).

5.B Empirical analysis

5.B.1 Data set and baseline covariates

Data were collected in the winter and spring of 2002/2003 in a large scale cross-sectional survey (WHO's Large Analysis and Review of European Housing and Health Status project) involving 5,882 adult respondents from 2,983 households in 8 European cities (Angers, France; Bonn, Germany; Bratislava, Slovakia; Budapest, Hungary; Ferreira do Alentejo, Portugal; Forli, Italy; Geneva, Switzerland; and Vilnius, Lithuania). Baseline measurements C were available on both respondent characteristics C_r (age, gender, marital status, education level, employment, smoking and environmental tobacco smoke at home) and household characteristics C_h (ownership, size, tenure, crowding, ventilation, natural light, heating and city of residence). This data set is described in greater detail in Shenassa et al. (2007).

5.B.2 Working models

Model selection for each of the working models was done using a backward elimination procedure (except for the exposure model). Each minimal model was constrained to include all baseline covariates (C) and, where applicable, exposure to dampness and mold (A) and mediators physical illness (M_1) and perception of control (M_2) as predictor terms. The minimal set of predictor terms of the imputation model for the outcome depressive symptoms (Y) also included second- and third-order interactions of A, M_1 and M_2 in order to ensure that different decompositions resulting from the final natural effect model appropriately reflected differences dictated by the data. Likewise, the minimal working model for M_2 included an interaction term between A and M_1. Maximal working models additionally included interactions between each of the baseline covariates (except for city of residence)[8] and each of the remaining predictor terms in the minimal model (that is, where applicable, A, M_1, M_2 and their second-order interaction terms).

In order to account for clustering by household, each of the working models was estimated by generalized estimating equations (GEE), assuming an independent working correlation structure, and backward elimination of the interaction terms involving baseline covariates was done via the QIC criterion.[9]

[8] Due to sparseness of the data, it was not feasible to include any higher-order interactions with city of residence in any of the working models, especially when refitting these models to the bootstrapped data sets.

[9] Note that clustering complicates identification assumptions. Although we do not

As levels of exposure to dampness and mold ($A = 1$ for exposed, $A = 0$ for non-exposed) were constant within households, corresponding weights resulting from the working model for exposure probability should be constructed so as to mimick a cluster-randomized trial. Therefore, only household characteristics C_h (but no respondent characteristics C_r) were included as predictors in this working model. Moreover, in contrast to the other working models, it was not subjected to model selection as it only stratified on a set of baseline covariates. The resulting model

$$\text{logit}P(A = 1|C_h) = \alpha_0 + \alpha_1^\top C_h, \tag{5.24}$$

was fitted to the original data set to calculate inverse probability of exposure weights

$$W_{0i} = \frac{1}{\hat{P}(A = A_i|C_{hi})}.$$

The final model for probability of physical illness ($M_1 = 1$ in the presence of at least one physical condition known to be related to mold exposure or 0 otherwise)

$$\text{logit}P(M_1 = 1|A, C) = \beta_0 + \beta_1 A + \beta_2^\top C, \tag{5.25}$$

was fitted to the original data set to calculate ratio-of-mediator-probability weights

$$W_{1i,a'} = \frac{\hat{P}(M_1 = M_{1i}|A = a', C_i)}{\hat{P}(M_1 = M_{1i}|A = a'', C_i)} = \frac{\hat{P}(M_1 = M_{1i}|A = a', C_i)}{\hat{P}(M_1 = M_{1i}|A = A_i, C_i)}$$

for each row in an extended data set as constructed as described in step 3 in section 5.3, henceforth referred to as extended data set 1.

Likewise, the final model for the density of perceived control (M_2, as measured on a 5-point Likert scale (reverse coded), and for convenience assumed to be normally distributed)

$$f(M_2|A, M_1, C) = N\Big(\gamma_0 + \gamma_1 A + \gamma_2 M_1 + \gamma_3 A M_1 + \gamma_4^\top C, \sigma^2\Big), \tag{5.26}$$

elaborate on implications for identification in multilevel settings, we would like to refer the interested reader to Talloen et al. (2016).

was fitted to the original data set to calculate ratio-of-mediator-probability weights

$$W_{2i,a''} = \frac{\hat{f}(M_2 = M_{2i}|A = a'', M_{1i}, C_i)}{\hat{f}(M_2 = M_{2i}|A = a', M_{1i}, C_i)} = \frac{\hat{f}(M_2 = M_{2i}|A = a'', M_{1i}, C_i)}{\hat{f}(M_2 = M_{2i}|A = A_i, M_{1i}, C_i)}$$

for each row in an extended data set as constructed as described in step 3 in section 5.3, henceforth referred to as extended data set 2.

Finally, the final model for probability of depressive symptoms ($Y = 1$ in the presence of at least 3 (out of 4) self-reported depressive symptoms or $Y = 0$ otherwise)

$$\begin{aligned}\text{logit}P(Y = 1|A, M_1, M_2, C) = {} & \delta_0 + \delta_1 A + \delta_2 M_1 + \delta_3 M_2 + \delta_4 A M_1 + \delta_5 A M_2 + \delta_6 M_1 M_2 \\ & + \delta_7 A M_1 M_2 + \delta_8^\top C + \delta_9^\top A C_1 + \delta_{10}^\top M_1 C_2 + \delta_{11}^\top M_2 C_3 \\ & + \delta_{12}^\top A M_1 C_4 + \delta_{13}^\top A M_2 C_5 + \delta_{14}^\top M_1 M_2 C_6 \qquad (5.27)\end{aligned}$$

with C_1 = (age, marital status, environmental tobacco smoke at home, crowding, ventilation), C_2 = (ownership, size, ventilation), C_3 = (marital status, education level, ownership, size, tenure, heating), C_4 = (ventilation), C_5 = (marital status) and C_6 = (ownership, size) was used to impute nested counterfactuals $Y_i(a, M_{1i}(a'), M_{2i}(a'', M_{1i}(a')))$ by fitted values

$$\hat{P}(Y_i = 1|A = a, M_{1i}, M_{2i}, C_i)$$

for each row in extended data set 1 or 2.

5.B.3 Conditional logistic natural effect model

No effect modification by covariates For simplicity of exposition, a simple logistic natural effect regression model excluding interaction or polynomial terms involving baseline covariates C

$$\begin{aligned}&\text{logit}P\{Y(a, M_1(a'), M_2(a'', M_1(a'))) = 1|C\} \\ &= \eta_0 + \eta_1 a + \eta_2 a' + \eta_3 a'' + \eta_4 a a' + \eta_5 a a'' + \eta_6 a' a'' + \eta_7 a a' a'' + \eta_8^\top C \qquad (5.28)\end{aligned}$$

was fitted either to extended data set 1, weighting by $W_{1i,a'}$, or to extended data set 2, weighting by $W_{2i,a''}$. The resulting decomposition of the total effect is described in the main text and corresponding estimates and 95% confidence intervals for each of the component effects for each of the six possible three-way decompositions

are displayed in Figure 5.4.[10] Moreover, since different decompositions did not differ significantly at the 5% significance level, an alternative natural effect model excluding interaction terms between a, a' and a''

$$\begin{aligned}&\text{logit}P\{Y(a, M_1(a'), M_2(a'', M_1(a'))) = 1|C\}\\&= \zeta_0 + \zeta_1 a + \zeta_2 a' + \zeta_3 a'' + \zeta_4'^{\top} C\end{aligned} \tag{5.29}$$

was again fitted either to extended data set 1, weighting by $W_{1i,a'}$, or to extended data set 2, weighting by $W_{2i,a''}$. Corresponding component effect estimates and 95% confidence intervals are listed in Table 5.5.

Allowing for effect modification by covariates In addition, a more elaborate logistic natural effect model focusing on effect modification by baseline covariates C

$$\begin{aligned}&\text{logit}P\{Y(a, M_1(a'), M_2(a'', M_1(a'))) = 1|C\}\\&= \omega_0 + \omega_1 a + \omega_2 a' + \omega_3 a'' + \omega_4 a a' + \omega_5 a a'' + \omega_6 a' a'' + \omega_7 a a' a'' + \omega_8^{\top} C\\&\quad + \omega_9^{\top} a C + \omega_{10}^{\top} a' C + \omega_{11}^{\top} a'' C\end{aligned} \tag{5.30}$$

was again fitted either to extended data set 1, weighting by $W_{1i,a'}$, or to extended data set 2, weighting by $W_{2i,a''}$. Note that this model still implies some constraints on how differences in decompositions may vary between strata defined by baseline covariates, since it excludes interactions between baseline covariates and second- and higher-order interaction terms involving a, a' and a''.

The joint natural direct effect of the presence of dampness or mold exposure on the odds of depression, $\exp\big(\hat{E}_{A\to Y}(0,0)\big|C\big)$, was significantly different (at the 5% significance level) in less crowded homes ($<$ 0.5 residents/room), medium crowded homes ($0.51 - 1$ residents/room) and crowded homes ($>$ 1 residents/room) ($\chi^2 = 6.24$, $P = 0.04$), according to model (5.30) fitted to extended data set 1 weighted by $W_{1i,a'}$. Corresponding effect estimates are listed in Table 5.9. Fitting model (5.30) to extended data set 2, weighting by $W_{2i,a''}$, led to the same conclusion ($\chi^2 = 6.31$, $P = 0.04$) and yielded nearly identical estimates (see Table 5.9).

[10] Confidence intervals and inference for the natural effect models were based on the bootstrap covariance matrix of 1000 bootstrap samples. To account for clustering by households, the data was resampled at the household level instead of individual respondent level.

	weighting by $W_{1i,a'}$		weighting by $W_{2i,a''}$	
	Estimate	95% CI	Estimate	95% CI
< 0.5 residents/room	1.97	1.03, 3.72	1.97	1.02, 3.72
$0.51 - 1$ residents/room	0.99	0.61, 1.67	0.99	0.61, 1.65
> 1 residents/room	1.28	0.72, 2.23	1.27	0.72, 2.22

Table 5.9: Estimates and 95% confidence intervals of the joint natural direct effect odds ratio, $\exp(E_{A\to Y}(0,0)|C)$ for different levels of crowding. Estimates are based on model (5.30) fitted either to extended data set 1, weighted by $W_{1i,a'}$ (left column) or to extended data set 2, weighted by $W_{2i,a''}$ (right column).[11]

	weighting by $W_{1i,a'}$		weighting by $W_{2i,a''}$	
	Estimate	95% CI	Estimate	95% CI
absence of env. tobacco smoke	0.99	0.61, 1.67	0.99	0.61, 1.65
presence of env. tobacco smoke	0.65	0.37, 1.15	0.65	0.37, 1.14

Table 5.10: Estimates and 95% confidence intervals of the joint natural direct effect odds ratio, $\exp(E_{A\to Y}(0,0)|C)$ for different levels of environmental tobacco smoke at the home. Estimates are based on model (5.30) fitted either to extended data set 1, weighted by $W_{1i,a'}$ (left column) or to extended data set 2, weighted by $W_{2i,a''}$ (right column).[12]

Moreover, the joint natural direct effect, $\exp\big(\hat{E}_{A\to Y}(0,0)\big|C\big)$, in homes without environmental tobacco smoke significantly differed (at the 5% significance level) from the joint natural direct effect in homes with environmental tobacco smoke ($\chi^2 = 3.99, P < 0.05$) according to model (5.30) fitted to extended data set 1 weighted by $W_{1i,a'}$. Corresponding effect estimates are listed in Table 5.10. Again, fitting model (5.30) to extended data set 2, weighting by $W_{2i,a''}$, led to the same conclusion and yielded nearly identical estimates (see Table 5.10).

[11] Estimates were obtained for a reference group of married, non-smoking men, aged 46 (mean age), living in Bonn, employed outside the home and that own a home sized 50-99 m^2 (with ventilation, enough natural light, heating in all rooms and no environmental tobacco smoke) in which they have lived for 17 years (mean tenure).

[12] Estimates were obtained for a reference group of married, non-smoking men, aged 46 (mean age), living in Bonn, employed outside the home and that own a home sized 50-99 m^2 (with ventilation, enough natural light, heating in all rooms and $0.51 - 1$ residents/room) in which they have lived for 17 years (mean tenure).

	weighting by $W_{1i,a'}$		weighting by $W_{2i,a''}$	
	Estimate	95% CI	Estimate	95% CI
male	1.05	1.01, 1.10	0.89	0.69, 1.15
female	1.05	1.01, 1.10	0.99	0.77, 1.26

Table 5.11: Estimates and 95% confidence intervals of the natural indirect effect odds ratio with respect to physical illness, $\exp(E_{A\to M_1Y}(1,1)|C)$ for men and women. Estimates are based on model (5.30) fitted either to extended data set 1, weighted by $W_{1i,a'}$ (left column) or to extended data set 2, weighted by $W_{2i,a''}$ (right column).[13]

According to model (5.30) fitted to extended data set 2, weighted by $W_{2i,a''}$, the natural indirect effect odds ratio with respect to physical illness, $\exp(E_{A\to M_1Y}(1,1)|C)$, differed significanly between men and women ($\chi^2 = 5.06, P = 0.02$). However, effect modification of this effect by gender could not be established upon fitting model (5.30) to extended data set 1 weighted by $W_{1i,a'}$ ($\chi^2 = 0.05, P = 0.83$). Corresponding effect estimates are listed in Table 5.11. These contrasting findings may be due to the fact that working model (5.26), used for calculating weights $W_{2i,a''}$, is more prone to model misspecification because it makes additional parametric assumptions on the distribution of M_2 and the association between M_1 and M_2. We may therefore avoid reliance on working model (5.26), and instead use weights $W_{1i,a'}$ (as derived from working model (5.25)), as also highlighted in the main text.

5.B.4 Marginal logistic natural effect model

Finally, a marginal logistic natural effect model

$$\begin{aligned}&\text{logit}P\{Y(a, M_1(a'), M_2(a'', M_1(a'))) = 1\}\\&= \theta_0 + \theta_1 a + \theta_2 a' + \theta_3 a'' + \theta_4 aa' + \theta_5 aa'' + \theta_6 a'a'' + \theta_7 aa'a''\end{aligned} \tag{5.31}$$

was fitted either to extended data set 1, weighting by $W_{0i} \times W_{1i,a'}$, or to extended data set 2, weighting by $W_{0i} \times W_{2i,a''}$. Resulting estimates for the marginal odds

[13] Estimates were obtained for a reference group of married, non-smoking respondents, aged 46 (mean age), living in Bonn, employed outside the home and that own a home sized 50-99 m^2 (with ventilation, enough natural light, heating in all rooms, $0.51 - 1$ residents/room and no environmental tobacco smoke) in which they have lived for 17 years (mean tenure).

ratios (and corresponding 95% confidence intervals) are displayed in the left panels of Figure 5.10. These odds ratios could easily be translated into risk ratios, as displayed in the right panels of Figure 5.10, along with their 95% confidence intervals.[14] Since odds ratio estimates are relatively close to the null, they can be seen to approximate associated risk ratio estimates quite well.

[14]Note that such translation is less straightforward when dealing with non-saturated models (e.g., when dealing with continuous exposures or models that condition on continuous or high-dimensional covariates) as risk ratios may differ according to reference exposure or covariate levels even when associated odds ratios are not modelled in such a way. For this reason, effects from the aforementioned conditional natural effect models were only expressed in terms of odds ratios. Alternatively, one could resort to log-linear instead of logistic natural effect models. However, because of ensuing extrapolation in the presence of even minor misspecification of non-saturated log-linear models (for risk ratios), these models tend to yield fitted values beyond the range of [0; 1], despite the outcome being binary. This can be especially problematic when fitting log-linear outcome models for imputing nested counterfactuals.

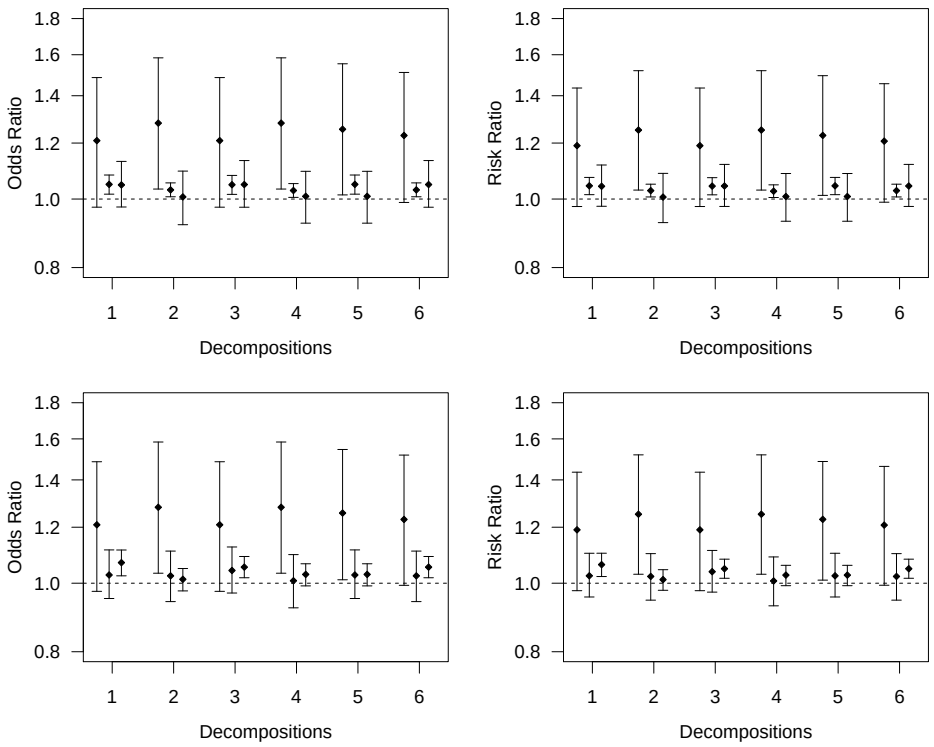


Figure 5.10: The left panels display marginal odds ratio estimates and 95% confidence intervals for each of the component effects listed in Table 5.2 (on the log odds ratio scale). Components are grouped per decomposition and displayed in the same order as in Table 5.2. The upper left panel displays estimates and confidence intervals as obtained from fitting model (5.31) to extended data set 1, weighting by $W_{0i} \times W_{1i,a'}$, whereas the lower left panel displays those obtained from fitting model (5.31) to extended data set 2, weighting by $W_{0i} \times W_{2i,a''}$. The right panels display associated risk ratio estimates and 95% confidence intervals as calculated from the fitted probabilities from model (5.31).

Chapter 6

Discussion

To further reflect on the previous chapters and sections in this thesis, in this last chapter, we will, in turn, discuss identifying assumptions for the causal effects of interest and methods for modeling and estimating these effects. In addition, we briefly summarize further challenges within the field of causal mediation analysis.

6.1 Identifying assumptions

6.1.1 Why non-parametric identification?

Because causal inference is often framed as inherently being a missing data problem (Holland, 1986) – in which unobserved counterfactual outcomes could be viewed as missing outcomes in a different hypothetical world – a main concern is whether the observed data carries sufficient information to infer (and possibly estimate) causal effects; that is, whether causal effects are *identifiable* from the data at hand. If the answer is positive, causal effects can be expressed as some functional of the observed data distribution.

However, in most, if not all, research settings, we need to make certain causal assumptions in order to make progress. Clearly articulating such assumptions is of utmost importance in order to be able to assess to what extent causal claims that follow from these assumptions may or may not be deemed credible. That is, to strengthen such claims, ideally, we wish to impose as few assumptions as possible. In particular, one should strive

to articulate assumptions that are sufficient for identifying the causal effect of interest without additionally imposing any restrictions on the full data distribution by means of parametric modeling assumptions. Failing to do so, may make results become very sensitive to the correctness of the assumed parametric models. We have therefore focused on *non-parametric identification* of natural and path-specific effects, in order to ensure that, if identification is obtained, it is solely based on sufficient causal assumptions, without the potential influence or interference of certain parametric restrictions, which may compensate for the lack of information in the observed data under insufficient causal assumptions.

Despite the importance of untangling causal from statistical modeling assumptions, inevitably, in a later stage – i.e. once non-parametric identification is obtained – additional modeling assumptions will typically need to be imposed to deal with the curse of dimensionality and/or facilitate interpretable results that are easy to communicate.

6.1.2 Identification via the adjustment criterion

In this thesis, we have mainly focused on non-parametric identification of natural effects (chapter 4) and, more generally, path-specific effects (chapter 5) by a generalization of the adjustment criterion (Shpitser et al., 2010) for mediation analysis (Shpitser and VanderWeele, 2011; Steen et al., 2016b). This generalized adjustment criterion provides a graphical rule that can be used as a guide to determine whether natural or path-specific effects are identified in a given graph (interpreted as a NPSEM) by adjustment for a common set of baseline covariates.

A major appeal of non-parametric identification via the adjustment criterion, a criterion that is implicitly prescribed in most applied papers, is that it leads to a standard form of identification result, which, in turn, allows for general estimation and modeling strategies that can easily be incorporated into the natural effect modeling framework, as presented in chapter 4 and chapter 5.

In single mediator settings or settings that aim for two-way effect decomposition, the identification result has been referred to as the mediation

formula (Pearl, 2001) or the adjustment formula for mediation analysis (Shpitser and VanderWeele, 2011). Generalizations of this formula that enable a more fine-grained effect decomposition along certain path-specific effects involving multiple mediators readily follow from repeated application of the adjustment criterion, as pointed out in chapter 5.

Sequential application of the adjustment criterion and its shortcomings

Under Markovian NPSEMs – i.e. in the absence of unmeasured confounding – the recanting witness criterion is both a sufficient and necessary graphical criterion for non-parametric identification of path-specific effects (Avin et al., 2005). From this criterion, it follows that, in settings with k sequential mediators, the finest identifiable decomposition is characterized in terms of $k+1$ distinct path-specific effects. Because of causal sufficiency under Markovian NPSEMs, each of these $k+1$ path-specific effects are identifiable by means of covariate adjustment. Sequential application of the adjustment criterion – henceforth referred to as the *sequential adjustment criterion* – can thus be conceived as an aid in finding a (minimal) common set of covariates that satisfies sufficient ignorability conditions under NPSEMs in order to obtain this most fine-grained $(k+1)$-way decomposition by means of simultaneous identification of its $k+1$ component path-specific effects. It turns out that, under Markovian NPSEMs, there is always an available set of covariates that satisfies the adjustment criterion for each of the consecutive identification steps described in section 5.A.2.

Under semi-Markovian NPSEMs – i.e. in the presence of unmeasured confounding – on the other hand, the recanting witness criterion, yet still a necessary criterion, is no longer sufficient for identification of path-specific effects. As a result, there is no guarantee that the targeted $(k+1)$-way decompositions are identifiable. Moreover, even if the $k+1$ component path-specific effects turn out to be identifiable, they might not all be identified by adjustment for a common set of covariates, as in Markovian NPSEMs. We may therefore need to settle for less fine-grained decompositions that might necessitate grouping certain path-specific effects in order to recover identification. The sequential adjustment criterion indicates that such coarser

decompositions can often be identified by means of adjustment for a common set of covariates either by ignoring certain mediators that occur later in the causal chain or by grouping mediators into a joint mediator, as suggested in section 5.A.2 (also see e.g. VanderWeele and Vansteelandt, 2013).

For instance, in Figure 3.3B, three-way effect decomposition is compromised because of unmeasured $M - Y$ confounding, which hampers separation of $E_{A\to Y}$, i.e. the joint natural direct effect (through neither L nor M), from $E_{A\to M\to Y}$, i.e. the partial indirect effect with respect to M. Indeed, the sequential adjustment criterion indicates that $L(a) \perp\!\!\!\perp A$ and $M(a,l) \perp\!\!\!\perp \{A, L\}$ but $Y(a,l,m) \not\perp\!\!\!\perp \{A, L, M\}$, because the latter independence would only hold conditional on unmeasured confounder U. Nonetheless, a coarser decomposition can be obtained by simply ignoring M. That is, decomposition into natural effects with respect to mediator L can be obtained (by means of adjustment for the empty set), since $L(a) \perp\!\!\!\perp A$ and $Y(a,l) \perp\!\!\!\perp \{A, L\}$.

Fine-grained three-way effect decomposition is also hampered in Figure 3.3C because unmeasured $L - M$ confounding does not permit to disentangle $E_{A\to LY}$, i.e. the natural indirect effect through L, and $E_{A\to M\to Y}$. In this case, the sequential adjustment criterion indicates that again, $L(a) \perp\!\!\!\perp A$, but that $M(a,l) \not\perp\!\!\!\perp \{A, L\}$ if data on U is unavailable. Treating $\{L, M\}$ as a joint mediator leads to a coarser decomposition into natural effects with respect to $\{L, M\}$. This two-way decomposition is identified by means of covariate adjustment (for the empty set), since $\{L(a), M(a)\} \perp\!\!\!\perp A$ and $Y(a,l,m) \perp\!\!\!\perp \{A, L, M\}$.

Finally, in Figure 3.3D, path-specific effects $E_{A\to LY}$ and $E_{A\to Y}$ cannot be separated because of unmeasured $L - Y$ confounding, which thus once more hinders three-way decomposition. Again, non-identification of this three-way decomposition is indicated by the sequential adjustment criterion, since $L(a) \perp\!\!\!\perp A$ and $M(a,l) \perp\!\!\!\perp \{A, L\}$, but $Y(a,l,m) \not\perp\!\!\!\perp \{A, L, M\}$ if data on U is unavailable. Nonetheless, by the recanting district criterion, $E_{A\to M\to Y}$ – and thus a coarser two-way decomposition – remains identifiable, without the need to rely on more general identification strategies (beyond adjustment for a common set of covariates) (see e.g. Miles et al., 2014). Interestingly, the sequential adjustment criterion does not seem to

be able to signal identification of this coarser decomposition by means of covariate adjustment. This can be appreciated by the fact that the sequential adjustment criterion adheres to a certain hierarchy of identification. That is, due to its inherently sequential nature, it prioritizes identification of natural effects over more generally defined path-specific effects, such as $E_{A\to M\to Y}$. As such, identification of certain generally defined path-specific effects could be conceived as merely a by-product of identification of specific natural effects, the contrast of which happens to correspond to these path-specific effects of interest.

For instance, whereas the partial indirect effect $E_{A\to M\to Y}$ corresponds to the contrast of two natural indirect effects[1] – i.e. the contrast between the joint natural indirect effect with respect to $\{L, M\}$ and the natural indirect effect with respect to L – in all causal diagrams in Figure 3.3, it is identified by the sequential adjustment criterion in Figure 3.3A but not in Figure 3.3D, because only in the former does the sequential adjustment criterion lead to identification of these two natural indirect effects.

It thus seems that the sequential adjustment criterion falls short of giving a complete characterization of decompositions that are identified by adjustment for a common set of covariates under semi-Markovian NPSEMs. In other words, despite the fact that it can be considered a sufficient criterion for identification of path-specific effects by means of adjustment for a common set of covariates, it is not a necessary criterion. Moreover, it is quite an indirect criterion for identification of path-specific effects, in that it primarily focuses on identification of decompositions. Identification of a certain path-specific effect of interest can thus only indirectly be assessed by verifying whether this path-specific effect corresponds to one of the components of an identified decomposition. Given these limitations and the fact that empirical studies may rarely aspire to obtain the finest possible decomposition in settings with multiple sequential mediators, further development of a complete graphical criterion for (direct) identification of any given path-specific effect by means of adjustment for a common set of

[1] Alternatively, $E_{A\to M\to Y}$ also corresponds to the contrast of two natural direct effects, i.e. the contrast between the natural direct effect with respect to L and the natural direct effect with respect to $\{L, M\}$.

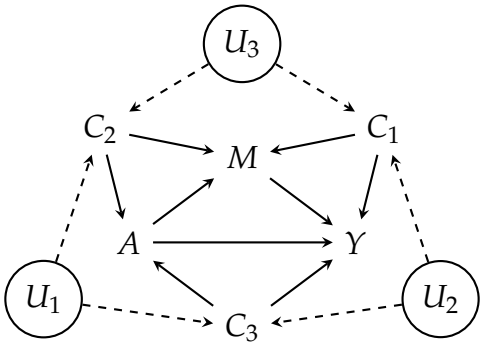


Figure 6.1: Causal diagram which allows to exploit certain exclusion restrictions pertaining to some of the confounders.

covariates may be warranted.

6.1.3 Beyond the adjustment criterion

Furthermore, in the light of more general – both sufficient and necessary – conditions for non-parametric identification that were outlined in chapter 3, identification by means of adjustment for a common set of covariates has been criticised for being too stringent, placing serious limits on identification power (e.g. Pearl, 2014). The question therefore naturally arises as to what extent more general identification results can be embedded in the natural effect modeling framework that we have presented.

In certain settings where natural effects are identified by the adjustment criterion, identification can also be obtained under a weaker set of conditions associated with an identification strategy sometimes referred to as 'piecemeal deconfounding', which involves adjustment for separate sets of baseline covariates (e.g. Pearl, 2014). For instance, since $\{C_1, C_2, C_3\}$ satisfies the adjustment criterion for natural effects in the causal diagram in Figure 6.1, we obtain the standard identification result

$$\begin{aligned}\sum_{c_1,c_2,c_3,m} & E(Y|A=a, M=m, C_1=c_1, C_2=c_2, C_3=c_3) \\ & \times P(M=m|A=a, C_1=c_1, C_2=c_2, C_3=c_3) \\ & \times P(C_1=c_1, C_2=c_2, C_3=c_3).\end{aligned}$$

However, because certain exclusion restrictions encoded in this diagram

imply that $Y \perp\!\!\!\perp C_2 | A, M, C_1, C_3$ and $M \perp\!\!\!\perp C_3 | A, C_1, C_2$, this result can be simplified as

$$\sum_{c_1, c_2, c_3, m} E(Y|A=a, M=m, C_1=c_1, C_3=c_3) \times P(M=m|A=a, C_1=c_1, C_2=c_2) P(C_1=c_1, C_2=c_2, C_3=c_3),$$

which corresponds to the result one would obtain via piecemeal deconfounding (as discussed in section 3.4.2). It thus follows that modeling demands can sometimes be significantly reduced when exploiting exclusion restrictions encoded in a causal diagram. In this particular case, one may refrain either from modeling the association between the outcome and C_2 – i.e. when relying on imputation-based estimation – or from modeling the association between the mediator and C_3, i.e. when relying on weighting-based estimation. As exemplified above, most often, simplified results will correspond to those obtained by a piecemeal deconfounding identification approach.

Interestingly, in settings with unmeasured mediator-outcome confounding, where identification of natural effects can be obtained by relying on mediating instruments, such as in the causal diagrams in Figure 3.5A and Figure 3.5B, the identification result can be considered to be closely related to Pearl's mediation formula. More specifically, since L may substitute for the mediator of interest M in Figure 3.5A, the natural direct (indirect) effect with respect to L and the natural direct (indirect) effect with respect to M coincide. Not surprisingly, $E\{Y(a, M(a'))\}$ is therefore identified by

$$\sum_l E(Y|A=a, L=l) P(L=l|A=a').$$

Similarly, in Figure 3.5B, Z transmits the direct effect with respect to the mediator of interest M, such that the natural direct (indirect) effect with respect to M corresponds to the natural indirect (direct) effect with respect to Z. Accordingly, $E\{Y(a, M(a'))\}$ is easily shown to be identified by

$$\sum_z E(Y|A=a', Z=z) P(Z=z|A=a).$$

Because of their standard form, functionals based on identification by mediating instruments seem, at least to some extent, to be amenable to natural effect modeling. Further work is needed to further investigate to what extent identification by mediating instruments can be integrated within the natural effect models framework.

However, whenever the adjustment criterion does not enable identification, identification results cannot generally be reduced to simple and standard expressions such as Pearl's mediation formula or generalizations thereof, thereby preventing a generic and elegant approach for estimation. It can be shown that, often, such identification results involve additional factors, expressing the need for additional working models that either lead to more involved calculation of weights or necessitate some form of Monte Carlo integration. For instance, although, under the NPSEM representation of the causal diagram $\mathcal{G}$ in Figure 3.4, identification of natural effects can be obtained by piecemeal deconfounding, the result (expression (3.8)) cannot easily be translated into an imputation- or weighting-based estimator. That is, at least not without reliance on additional working models.

6.1.4 Dealing with uncertainty about causal structure

It turns out that increased identification power in graphs with hidden variables (relative to identification by the adjustment criterion) mostly seems to rely on additional exclusion restrictions, either pertaining to observed confounders or to mediating instruments (see chapter 2 and chapter 3).

The question then arises whether such exclusion restrictions on observed variables, in some settings, may perhaps even be considered stronger assumptions than the usual assumption of causal sufficiency, i.e. that of no unmeasured common causes of any two variables in the graph. There might indeed be reason to question the extent to which subject matter knowledge or expert judgment may lead to impose certain exclusion restrictions. Imai et al. (2014) recently argued that 'in many substantive research settings, scholars are unlikely to possess such precise knowledge about the structure of confounding. ... In most observational research, however, researchers measure a large number of covariates, and the exact structure between these

covariates and unobservables is usually highly uncertain.'

Testable implications

However, as argued by the same authors, the data itself may often provide further guidance in uncovering aspects of the causal structure. As discussed in chapter 2, conditional independencies encoded in a graph can be considered to constitute a set of implications that can be tested from observed data, thus enabling to partially validate a given graphical model. In other words, – at least certain – exclusion assumptions may thus in theory be refuted. From this perspective, one could argue that, ideally, subject matter knowledge and expert judgment should be combined with additional information that can be extracted from the data itself by model testing procedures and/or causal induction algorithms.

Imai et al. (2014), however, further point to the potential limitations of such testing procedures, which are often characterized by inflated Type I and Type II errors due to multiple testing and small sample size, respectively. Moreover, even if one were to focus on testing a limited set of crucial exclusion restrictions in a sufficiently large dataset, the potential presence of hidden variables may significantly reduce the number of testable implications[2] such that targeted exclusion restrictions may turn out te be untestable.

Nonetheless, it has been shown that exclusion restrictions encoded in such latent variable causal models (such as semi-Markovian models) may impose different types of non-parametric constraints on the observed data distribution that cannot be expressed in terms of conditional independencies. These include so-called *Verma constraints* (Robins, 1986; Shpitser and Pearl, 2008b; Verma and Pearl, 1991) – another type of equality constraints that can be conceived of as conditional independencies that arise in interventional distributions – and inequality constraints (Evans, 2012; Pearl, 1995b). Future research will likely be able to further assess whether and to what extent such additional constraints can aid in uncovering aspects

[2]That is, since each conditional independence that involves an unmeasured variable is logically excluded from the set of testable implications

of the causal structure in the presence of hidden variables (Shpitser et al., 2009) in order to possibly strengthen exclusion assumptions which may be crucial for non-parametric identification of natural effects or, more generally, path-specific effects of interest.

In addition, two alternative strategies have been proposed for dealing with uncertainty about causal structure when aiming to learn about causal effects.

Non-parametric bounds

A first, non-parametric, strategy could be considered to invert the problem by constructing empirical *bounds* for the causal parameter of interest, under constraints that are solely imposed by the observed data distribution. As such bounds tend to be uninformative when relaxing all identifying assumptions – even about the direction of a potential causal effect – more tight bounds can often be obtained by relaxing only those identifying assumptions that are subject to high uncertainty and/or by relying on certain monotonicity assumptions.

In the context of mediation analysis, sharp bounds for natural direct and indirect effects have been derived under assumptions that can be enforced experimentally, either by single intervention experiments that only randomize treatment (Chiba and Taguri, 2013; Kaufman et al., 2009; Sjölander, 2009) – and thus ensure identification of treatment effects – or by parallel designs that augment the single intervention with a joint intervention that randomizes both treatment and mediator – thereby additionally eliminating mediator-outcome confounding (Imai et al., 2013; Naimi, 2015; Robins and Richardson, 2010). Corresponding bounds can thus be considered informative as to the importance of Pearl's cross-world independence assumption (Imai et al., 2013; Robins and Richardson, 2010).

Recently, bounds have also been derived under the assumption of no unmeasured mediator-outcome confounding, accommodating confounders to be possibly affected by treatment (Miles et al., 2015; Tchetgen Tchetgen and Phiri, 2014). These bounding methods have proven to give quite informative results, often allowing to identify the sign of the natural effects of

interest (at least when unmeasured mediator-outcome confounding can be excluded).

Nonetheless, bounds have been criticised for still being too uninformative since they only consider the most extreme possible scenarios, which are often more extreme than one would usually be willing to consider plausible (Jiang and VanderWeele, 2015). Furthermore, their applicability is rather limited as available bounding formulae are restricted to binary – or possibly multicategorical (Chiba and Taguri, 2013; Miles et al., 2015) – variables.

Sensitivity analysis

A second strategy considers the extent to which certain key identifying assumptions would need to be violated in order to considerably change one's conclusions. Corresponding *sensitivity analysis* techniques thus aim to assess the robustness of causal claims in function of the degree of violation of the identifying assumptions, as captured by certain sensitivity parameters. Such strategy has been claimed to be more informative than bounds because, as opposed to bounds, sensitivity analyses allow to consider scenarios that reflect realistic degrees of violations of the identifying assumptions (Jiang and VanderWeele, 2015).

Most available sensitivity analyses for mediation analysis (Albert and Nelson, 2011; Albert and Wang, 2015; Daniel et al., 2015; Imai et al., 2010b; Imai and Yamamoto, 2013; le Cessie, 2016; Vansteelandt and VanderWeele, 2012; VanderWeele, 2010) are, however, embedded within strictly parametric frameworks and are therefore often limited in the number of settings in which they can be applied (although see Ding and VanderWeele, 2016; Hafeman, 2011; Imai et al., 2010b; Tchetgen Tchetgen and Shpitser, 2012; VanderWeele and Chiba, 2014, for exceptions).

Unlike bounding methods, there are currently no sensitivity analysis methods available that quantify the robustness of empirical findings against violations of cross-world independence in isolation (Jiang and VanderWeele, 2015). Moreover, the lack of a generic framework for sensitivity analysis has long prohibited its integration in the natural effect model framework. Nonetheless, some promising recent developments, such as the

non-parametric framework of Ding and VanderWeele (2016), may pave the way for such integration.

6.1.5 Cross-world contemplations

As pointed out in detail in chapter 3, both the cross-world nature of natural and path-specific effects and the required cross-world assumptions for their identification have been the subject of an ongoing debate (see e.g. Robins and Richardson, 2010; Naimi et al., 2014a), roughly dividing the field into NPSEM 'skeptics' and 'advocates'. In chapter 3, we have tried to shed some light on this controversy, and illustrated the important role of mediating instruments and deterministic expanded graphs (Robins and Richardson, 2010) in elucidating and bridging this conceptual and ontological divide.

In addition, as pointed out by Petersen et al. (2006), even in the absence of any reference to cross-world quantities or restrictions, the identification result of the (conditional) natural direct effect (by the mediation formula) still carries an empirically meaningful interpretation (also see Didelez et al., 2006; Geneletti, 2007; van der Laan and Petersen, 2008). More specifically, since, under all remaining assumptions,

$$\sum_m \{E(Y|A=a,M=m,C) - E(Y|A=a',M=m,C)\}P(M=m|A=a',C)$$
$$= \sum_m E\{Y(a,m) - Y(a',m)|C\}P(M(a')=m|C),$$

this result can be interpreted as a *standardized* direct effect; that is, a weighted average of the controlled direct effect at each possible level m of the mediator weighted with respect to the distribution of $M(a')$, i.e. the counterfactual intermediate outcome under treatment $A = a'$.

This interpretation has given rise to the more formal definition of so-called randomized intervention analogs of natural effects (VanderWeele et al., 2014), which conceive of fixing the mediator at some level that is randomly assigned from the conditional counterfactual mediator distribution $P(M(a') = m|C)$ rather than at the individual counterfactual level (see Lok, 2016; Naimi et al., 2014b, for related approaches). Importantly, because their definitions do not employ cross-world counterfactuals, strong and

untestable assumptions, such as cross-world independence, may be thus avoided. Such causal quantities may therefore be of particular interest in settings that typically suffer from issues of intermediate confounding, i.e. settings with multiple mediators (Vansteelandt and Daniel, 2016) and/or longitudinal measurements (Vanderweele and Tchetgen Tchetgen, 2016). Moreover, they tend to correspond more closely to relevant policy measures that can be estimated from actual interventions.

6.2 Flexible modeling using natural effect models

Along with direct application of Pearl's mediation formula, the estimators discussed in chapter 4 can be considered within Tchetgen Tchetgen and Shpitser (2012)'s more general semi-parametric framework for mediation analysis. In particular, Tchetgen Tchetgen and Shpitser (2012) showed that estimation of population-averaged natural effects requires postulating a correct statistical model for any two of the following quantities:

(i) the expected outcome Y, given mediator, treatment and a sufficient adjustment set of baseline covariates C,

(ii) the distribution of the mediator M, given treatment and baseline covariates C

(iii) the distribution of treatment A, given baseline covariates C.

Because of the curse of dimensionality and potential presence of continuous variables, non-parametric modeling is typically not feasible, so one needs to rely on possibly misspecified (semi-)parametric working models.

Whereas direct application of the mediation formula (e.g. Imai et al., 2010a; VanderWeele and Vansteelandt, 2009, 2010) typically relies on (adequate specification of) an outcome model (i) and a mediator model (ii), weighting- and imputation-based estimators substitute a propensity score model (iii) for either a model for the outcome (i) or for the mediator distribution (ii), respectively. In addition, when treatment is randomized[3] or

[3] In this case, the propensity score model (iii) is known exactly.

interest lies in conditional natural effects given C one does not require a model for (iii), such that a correct model for either (i) or (ii) suffices.

Similar to marginal structural models offering a modeling framework for semi-parametric imputation- and weighting-based estimators for total causal effects, natural effect models offer such a generic and flexible framework for related semi-parametric estimators in the context of mediation analysis. Moreover, this framework enables to both circumvent potentially computer-intensive Monte-Carlo integration methods (Imai et al., 2010a) and simplify results and hypothesis testing because of reliance on a parsimonious model structure for the natural effects of interest (also see van der Laan and Petersen, 2008).

Such parsimonious model structure may prove to be especially useful for estimation of stratum-specific natural effects when either model (i) or (ii) involves a non-linear link function, in which case, direct application of the mediation formula may yield complex results, even when using simple parametric models for (i) and (ii).[4] For instance, marginalization of a logistic regression model for the outcome with respect to the mediator distribution only results in a logistic model – and thus simple expressions for conditional natural effect parameters – if the mediator follows a so-called bridge distribution (Tchetgen Tchetgen, 2014). Alternatively, simple expressions can be obtained when the mediator is normally distributed with constant error variance and, additionally, the outcome is rare, since odds ratios are then known to approximate risk ratios, which are 'collapsible' (VanderWeele and Vansteelandt, 2010).

6.2.1 Strengths and weaknesses of the proposed estimators

Strengths and weaknesses of each of these estimators have been discussed in detail in chapter 4. In sum, the weighting-based estimator may suffer from weight instability in the presence of strong associations between the mediator and either (or both) treatment or baseline covariates or when dealing with continuous mediators. In particular, besides misspecification of

[4] These issues may arise for any setting that results in a non-saturated natural effect model. In this sense, similar issues may arise for population-average natural effects for a continuous treatment.

the model for the mediator distribution, extreme weights may also indicate potential violations of the *positivity assumption*

$$P(M = m|A = a, C = c) > 0, \qquad \text{for all levels of } a, m, c.$$

This assumption basically states that within each treatment (or exposure) group and within each stratum defined by baseline covariates C, there is a nonzero probability of finding subjects with any given mediator level m along the support of the marginal mediator distribution. Although this assumption is crucial for identification (e.g. Hong, 2010; Imai et al., 2010b), it is often, as in this thesis, made implicitly.[5]

Violations of the positivity assumption due to strong exposure-mediator or confounder-mediator associations may go unnoticed when using the imputation-based estimator. This is because, typically, when information about the effect of the mediator on the outcome is sparse within certain strata defined by the exposure and covariates, this information is borrowed across strata, resulting in potential model extrapolation. Apart from the fact that routine analysis of the estimated weights – as part of model validation procedures for the weighting-based estimator – may signal potential violations of the positivity assumption, it has been argued that standard errors of the weighting-based estimator more honestly reflect extrapolation uncertainty in general, even under weak indication of violations of the positivity assumption (e.g. Rubin, 1997; Tan, 2007).

Another main concern with respect to the imputation-based estimator is the potential failure of the imputation model for the outcome to reflect the structure of the natural effect model – i.e. model uncongeniality – whenever the latter is not saturated. To the extent that this is justified, we have drawn some reassurance from missing data studies on multiple imputation which have found that uncongenial model specification in settings with missingness not only in the outcome, but possibly also in high-dimensional covariates, yields relatively modest bias (e.g. Van Buuren et al., 2006). To

[5]For population-averaged natural effect models, or natural effect models conditional on a subset of confounders, an additional positivity assumption regarding the exposure distribution is implicitly made, i.e. $P(A = a|C = c)$ for all levels of a, c.

the best of our knowledge, extensive studies on model congeniality have not been conducted in the context of mediation analysis, perhaps with the exception of the aforementioned work of Tchetgen Tchetgen (2014), in which it was shown that logistic imputation regression models are collapsible over the mediator – and hence congenial with a logistic natural effect model – only if the mediator follows a bridge distribution (Wang and Louis, 2003).[6] Even though uncongeniality may be less of a concern if the natural effect model is considered to provide a useful summary result, remaining concerns can be partially alleviated upon fitting a sufficiently rich imputation model by means of more advanced modeling methods such as generalized additive models or machine learning techniques. Crucially, a minimal imputation model should include exposure-mediator interactions in order not to attenuate potential interactions – i.e. mediated interaction – in the natural effect model. Nonetheless, more investigation is required to further assess the importance of issues surronding uncongeniality and to arrive at a more formal characterization of 'sufficiently rich' imputation models. Alternatively, one could rely on a doubly robust estimator that yields consistent estimates when either the imputation model (i) or both the mediator model (ii) and the propensity score model (iii) are correctly specified (Vansteelandt et al., 2012b).

6.2.2 Multiply robust estimators

More generally, in analogy with doubly robust estimators for total causal effects – which require either an outcome model or propensity score model to be adequately specified – a triply robust estimation approach has been developed that provides consistent estimators for both population-averaged (Tchetgen Tchetgen and Shpitser, 2012; Zheng and van der Laan, 2012) and stratum-specific natural effects (Tchetgen Tchetgen and Shpitser, 2014) when any two of the three aforementioned models are correctly specified.

Despite the theoretical appeal of such estimators, it has been argued that, as opposed to the imputation- and weighting-based estimator for

[6] This symmetric distribution has slightly heavier tails than the standard normal distribution when standardized to have unit variance. For more details, see Wang and Louis (2003).

natural effect models, their relative complexity may be a barrier to routine application (Vansteelandt et al., 2012b). Nonetheless, future research might further extend earlier attempts to integrate such multiply robust estimators in the natural effect model framework (Vansteelandt et al., 2012b).

6.2.3 Inverse odds weighting

Importantly, Zheng and van der Laan (2012) argued that reliance on the model for the mediator (ii) may in general be reduced in multiply robust estimators when, in addition to models for (i), (ii) and (iii), specifying a model for

(iv) the distribution of treatment A, given the mediator and baseline covariates C.

This additional model (iv) can be shown to give rise to another weighting-based estimator that deserves further consideration, as it can be shown to easily be incorporated in the natural effect model framework. This inverse-odds-weighted estimator, as originally suggested by Huber (2014) (although see Nguyen et al., 2015, 2016; Tchetgen Tchetgen, 2013, for a highly similar estimator) relies on working models for (iii) and (iv). It can be seen to arise naturally upon rewriting the population expectation of the ratio-of-mediator-probability-weighted estimator for stratum-specific natural effects by application of Bayes rule:

$$\begin{aligned} E\{Y(a,M(a'))|C\} &= E\left[Y\frac{P(M|A=a',C)}{P(M|A=a,C)}\middle|A=a,C\right] \\ &= E\left[Y\frac{P(M,A=a'|C)P(A=a|C)}{P(M,A=a|C)P(A=a'|C)}\middle|A=a,C\right] \\ &= E\left[Y\frac{P(A=a'|M,C)P(A=a|C)}{P(A=a|M,C)P(A=a'|C)}\middle|A=a,C\right]. \end{aligned}$$

Similarly, for population-average natural effects, we obtain

$$\begin{aligned} E\{Y(a,M(a'))\} &= E\left[\frac{YI(A=a)}{P(A=a|C)}\frac{P(M|A=a',C)}{P(M|A=a,C)}\right] \\ &= E\left[\frac{YI(A=a)}{P(A=a|C)}\frac{P(A=a'|M,C)P(A=a|C)}{P(A=a|M,C)P(A=a'|C)}\right] \end{aligned}$$

$$= E\left[\frac{YI(A=a)}{P(A=a'|C)}\frac{P(A=a'|M,C)}{P(A=a|M,C)}\right].$$

This weighting-based estimator may offer a promising alternative to both the imputation-based estimator and the ratio-of-mediator-probability-weighted estimator, as in many settings, it may combine the strength of both estimators. As the latter estimator, it circumvents certain incongeniality issues since it does not rely on a model for the outcome. However, unless treatment is randomized, issues of uncongeniality may still arise between the two working models for the probability of treatment. Moreover, it may lessen modeling demands and provide more stable weights than the other weighting-based estimator, especially when dealing with binary or multicategorical treatments. Furthermore, it may be the preferred estimator when interest lies in joint mediated effects along multiple mediators because, similar to the imputation-based estimator, it avoids reliance on a model for the joint mediator density (Nguyen et al., 2015). Given the strengths of this estimator, its implementation would be of added value to our medflex package.

6.2.4 Multiple sequential mediators

The practical appeal of natural effect models becomes even more apparent in settings with multiple mediators, in which researchers may aim to obtain a more fine-grained decomposition of the total causal effect into path-specific effects, as discussed in chapter 5.

In chapter 5, we have focused on an estimator that relies on a model for the outcome and a model for the distribution of either of two mediators, thereby combining the imputation- and weighting-based estimator of chapter 4. In practice, however, even in settings with only two sequential mediators M_1 and M_2, multiple combinations or subsets of working models are possible in order to construct estimators for the three component effects.

For instance, Lange et al. (2014)'s weighting-based approach for causally unrelated mediators, which requires a model for the distribution of each of the involved mediators, could easily be modified in order to also accommodate for sequential mediators. Instead of calculating weights based on the

joint mediator distribution

$$\frac{P(M_1|A=a,C)P(M_2|A=a,M_1,C)}{P(M_1|A=a,C)P(M_2|A=a,M_1,C)},$$

a weighting-based approach could alternatively rely on inverse odds weights

$$= \frac{P(A=a|C)P(A=a'|M_1,C)P(A=a''|M_1,M_2,C)}{P(A=a'|C)P(A=a''|M_1,C)P(A=a|M_1,M_2,C)},$$

which requires three different models for the treatment distribution. In general, if one aims to obtain the most fine-grained decomposition[7] that can be non-parametrically recovered in the presence of k sequential mediators – i.e. a $(k+1)$-way decomposition – the number of nuisance working models that can be combined to construct estimators for the component effects can be shown to equal $2k+2$. However, a general theory for multiply robust estimators in settings with multiple sequential mediators is currently lacking.

In one of the next releases of the medflex package, we hope to provide additional functionalities for conducting mediation analyses with multiple sequential mediators, as discussed in chapter 5.

6.2.5 Finite sample performance

Although finite sample performance of the weighting- and imputation-based estimators from chapter 4 has been studied by Vansteelandt et al. (2012b), so far no studies have further compared finite-sample properties of the inverse-odds-weighted estimator or related estimators for component effects in settings with multiple sequential mediators; that is, at least not within the realm of natural effect models (although see Huber et al., 2016, for a very recent and detailed comparison of estimators outside the realm of natural effect models).

[7] It can be argued, however, that, in settings with k multiple sequential mediators, one may rarely be interested in the finest possible $(k+1)$-way decomposition. Alternatively one may decide which coarser decomposition would be of most interest, and adopt identifying criteria and estimation accordingly.

6.2.6 Measures of precision

A general bootstrap approach can be applied to obtain standard errors or confidence intervals for any of the natural effect model parameters. Moreover, inference can be based on the bootstrap variance-covariance matrix of the natural effect model. Since the bootstrap can be computationally intensive, we have additionally derived robust sandwich estimators in chapter 4. Although derivations may become more involved in the presence of additional nuisance working models, similar sandwich estimators could be constructed for the variance of natural effect model parameter estimates in settings with sequential mediators, such as those obtained by weighted imputation, as discussed in chapter 5.

6.3 Further challenges

6.3.1 Mediation analysis with time-to-event outcomes

Although, in this thesis, we have only considered a class of generalized linear natural effect models, our proposed framework has already proven useful for mediation analysis in a survival context. However, suggested approaches for natural effect modeling of time-to-event outcomes have hitherto only focused on a weighting-based estimator that requires a correct model for the mediator distribution (Lange et al., 2012, 2014).

Ongoing research therefore aims to extend this work by additionally incorporating imputation-based and inverse-odds-weighted estimators (see Tchetgen Tchetgen, 2013; Nguyen et al., 2015, 2016, for applications of this estimator in a survival context). It merits attention that these different estimators may have different implications when it comes to dealing with censoring. That is, using an imputation-based estimator, censoring can be naturally dealt with via the imputation model, given that one can assume censoring to be non-informative conditional on treatment, mediator(s) and baseline covariates C (and possibly additional covariates associated with both censoring and outcome). More specifically, under this assumption, censored survival times can simply be predicted based on the imputation model, in order to eliminate potential selection bias. The weighting-based

estimators, on the other hand, require additional modeling of the censoring mechanism in order to construct inverse-probability-of-censoring weights to tackle potential selection bias via the mediator.

6.3.2 Mediation analysis with longitudinal measurements and latent constructs

It is known that questions of mediation pose additional challenges in longitudinal studies because the inherent time-varying nature of mediators and confounders (and possibly treatment) adds to the level of complexity. Apart from some notable exceptions (Bind et al., 2015; van der Laan and Petersen, 2008; VanderWeele, 2009; Vanderweele and Tchetgen Tchetgen, 2014), few studies have attempted to approach longitudinal mediation analysis from a formal perspective, such as the counterfactual framework, which gives clear definitions of target estimands and enables articulating sufficient identifying assumptions. When it comes to identifiability, it merits attention that Shpitser (2013)'s recanting district criterion may be a very helpful guide in assessing which potential pathways of interest may or may not be identifiable from the data at hand (under NPSEMs), possibly in the presence of unmeasured confounding between, for instance, repeated measurements of the same mediator.

Because intermediate confounding may be even more difficult to rule out in longitudinal studies than in point treatment studies, Vanderweele and Tchetgen Tchetgen (2016) have suggested switching focus to randomized intervention analogs of natural effects, since these estimands require weaker assumption – i.e. they do not demand the absence of intermediate confounding – while at the same time allowing meaningful interpretations.

Related, and perhaps more challenging complications arise in the field of psychology, where interest often lies in mediators or outcomes that are psychological constructs, such as attitudes or emotional distress (in chapter 4), which are only indirectly observable via certain indicator variables that are subject to measurement error. The (parametric) structural equation modeling tradition naturally deals with this by incorporating a measurement model for the latent construct(s). However, this topic has only recently

been subjected to formal analysis within the counterfactual framework (e.g. Albert et al., 2016; Loeys et al., 2014; Muthén and Asparouhov, 2015) (also see earlier, related work on measurement error in a mediation context, e.g. le Cessie et al., 2012; VanderWeele et al., 2012).

The modeling approaches discussed in this thesis may – at least to some extent – provide an ad-hoc answer to both of these challenges, insofar that simultaneously considering all repeated measurements of the same mediator – or all indicators of a latent variable – as a joint mediator enables to decompose total effects into a natural direct and a natural indirect effect with respect to this joint mediator. Indeed, the advantage of such an approach is that, in the likely presence of unmeasured confounding between multiple measurements of the same mediator, simply treating them as a joint mediator may render identifying assumptions of the corresponding component effects more plausible. Nonetheless, more fundamental solutions are needed.

Chapter 7

Samenvatting

Naast het opsporen van oorzakelijke verbanden, poogt men via empirisch onderzoek vaak tot meer diepgaand wetenschappelijk inzicht te komen door zich verder toe te spitsen op mogelijke onderliggende processen die dergelijke oorzaak-gevolg relaties kunnen verklaren. Hiertoe tracht men via statistische mediatie-analyse na te gaan in welke mate het effect van een bepaalde blootstelling of behandeling (zoals bv. psychotherapie) op een bepaalde uitkomst (bv. depressieve symptomen) is toe te schrijven aan de invloed van vermoedelijke tussenliggende of mediërende factoren (bv. verandering in attitudes) enerzijds, en/of aan niet nader omschreven alternatieve processen anderzijds. Stel dat men de onderliggende processen van een oorzakelijk effect beschouwt als een verzameling van verschillende mogelijke causale *paden* in een *oorzaak-gevolg ketting*. Het doel van mediatie-analyse bestaat er dan in om inzicht te verschaffen in *hoe* dit effect precies tot stand komt, meer bepaald door onrechtstreekse of *indirecte* paden via één of meerdere vermoedelijke mediatoren te kwantificeren en te onderscheiden van alle andere mogelijke paden, welke gemakshalve samengebracht worden onder de noemer van een *direct* effect.

Dankzij de ontwikkeling van een formele benadering binnen de causale inferentie literatuur, zijn er de laatste decennia enorme theoretische en methodologische bijdrages geleverd op het vlak van mediatie-analyse. In tegenstelling tot traditionele benaderingen, is deze benadering erin geslaagd om (i) welomlijnde, interpreteerbare definities van directe en indirecte ef-

fecten voort te brengen en (ii) de nodige – vaak onuitgesproken en mogelijk zwak onderbouwde – veronderstellingen waarop conclusies uit mediatie-analyses gefundeerd zijn, duidelijk in kaart te brengen en te onderwerpen aan kritische evaluatie.

De aannemelijkheid van dergelijke assumpties hangt veelal af van de mate waarin bijkomende cruciale variabelen, zoals mogelijke gemeenschappelijke oorzaken of *confounders* van de blootstelling (of behandeling) en de uitkomst (bv. de mate waarin men op eigen initiatief in therapie gaat), in rekening gebracht worden in de uiteindelijke statistische analyse. Om een onvertekende schatting van zogenaamde *natuurlijke* directe en indirecte effecten (Robins and Greenland, 1992; Pearl, 2001) te bekomen via mediatie-analyse dient men traditioneel niet enkel mogelijke confounders van de blootstelling en uitkomst in rekening te brengen, maar eveneens deze van de blootstelling en de vermoedelijke mediator, en deze van de mediator en de uitkomst (VanderWeele and Vansteelandt, 2009).

Bovendien bevat geobserveerde data geen informatie omtrent natuurlijke directe en indirect effecten indien confounders van de mediator en uitkomst zelf zijn beïnvloed door de blootstelling (VanderWeele and Vansteelandt, 2009). Gezien het feit dat dergelijke confounders tegelijk fungeren als mediator, zou statistische controle voor deze confounders immers deel van het beoogde natuurlijk indirect effect wegcijferen in de uiteindelijke schatting. Mediatie-analyse met meerdere mediatoren brengt op deze manier bijkomende uitdagingen met zich mee, welke in het verleden vaak onderbelicht zijn gebleven door een aantal vereenvoudigende, maar vaak zwak onderbouwde assumpties (zoals bijvoorbeeld de assumptie dat mediators elkaar niet onderling beïnvloeden of de assumpties van lineariteit en/of effect homogeniteit).

Voorts brengt mediatie-analyse ook een aantal uitdagingen met zich mee wat betreft statistisch modelleren. Confounders waarvoor statistische controle vereist is, zijn immers vaak hoog-dimensioneel en bestaan niet zelden uit een mix van discrete en continue variabelen. De resulterende *vloek van dimensionaliteit* die hiermee gepaard gaat, impliceert dat men, gezien de schaarscheid van data, bij het statistisch modelleren genoodzaakt is te vertrouwen op vereenvoudigende modelassumpties (voornamelijk

parametrische assumpties). Het risico op incorrecte modelassumpties, en dus op vertekende effectschattingen en conclusies, vergroot echter naarmate meer mediatoren en confounders in rekening worden genomen. Hoewel dit probleem niet uniek is voor mediatie-analyse, staat de ontwikkeling van semi-parametrische methoden, die het mogelijk maken om bepaalde modelassumpties te vermijden, nog in zijn kinderschoenen op het vlak van mediatie-analyse (Tchetgen Tchetgen and Shpitser, 2012, 2014; Zheng and van der Laan, 2012).

In het eerste luik van dit proefschrift trachten we een toegankelijk en volledig overzicht te bieden van minimale causale assumpties voor mediatie-analyse.

In **hoofdstuk 2** bespreken we eerst de theoretische achtergrond omtrent grafische modellen en gaan we dieper in op recent ontwikkelde grafische algoritmes uit de artificiële intelligentie literatuur (Huang and Valtorta, 2006; Shpitser and Pearl, 2006a,b, 2008a; Tian and Pearl, 2002, 2003). Het belang van deze algoritmes is dat ze een volledige beschrijving geven van grafische modellen waaronder causale effecten *geidentificeerd* zijn op basis van geobserveerde data en op deze manier dus zowel *noodzakelijke* als *voldoende* voorwaarden omlijnen voor non-parametrische identificatie. Deze algoritmes laten onder andere toe om, onder bepaalde assumpties, tot een onvertekende schatting te komen van causale effecten, zelfs indien men er niet in slaagt om alle confounders in kaart te brengen.

In **hoofdstuk 3** brengen we de lezer meer inzicht in de aard van causale assumpties waarop mediatie-analyse berust (Robins and Richardson, 2010). Aan de hand van een aantal uitgewerkte voorbeelden kaderen we de voorafgaande literatuur rond voldoende voorwaarden voor non-parametrische identificatie (Pearl, 2001) binnen de algemeenheid van een recent voorgesteld (voldoende en noodzakelijk) grafisch algoritme (Shpitser, 2013) dat voortbouwt op de inzichten en algoritmes uit hoofdstuk 2. In dit hoofdstuk leveren we ook een belangrijke bijdrage aan de huidige literatuur door stil te staan bij de specifieke implicaties van de algemeenheid van dit algoritme. Hierbij duiden we op nieuwe identificatie-strategieën die toelaten om onvertekende schattingen te bekomen van natuurlijke directe en indirecte (en meer algemeen gedefinieerde pad-specifieke) effecten onder ongemeten

confounding van mediator en uitkomst.

In het tweede luik bieden we praktische oplossingen voor het schatten van pad-specifieke effecten met behulp van flexibele statistische modellen – zogenaamde *natural effect models* (Lange et al., 2012; Loeys et al., 2013; Vansteelandt et al., 2012a) – en semi-parametrische methoden. De flexibiliteit van dergelijke modellen laat ons niet enkel toe om bepaalde parametrische modelassumpties te vermijden, maar ook om de interpreteerbaarheid van resultaten te verbeteren en het toetsen van hypotheses te vereenvoudigen.

In **hoofdstuk 4** bespreken we medflex, een open-source software pakket in de statistische programmeertaal R, dat we zelf hebben ontwikkeld om deze recente flexibele mediatie-analyse technieken toegankelijker te maken voor een breder toegepast publiek (Steen et al., 2016b). In dit hoofdstuk geven we een uitgebreid overzicht van de mogelijkheden van dit software pakket aan de hand van een uitgewerkte voorbeeldanalyse, en bespreken we ook de voordelen ten op zichte van alternatieve software voor mediatie-analyse.

In **hoofdstuk 5** breiden we dit flexibel modelleerkader uit om praktische oplossingen te bieden voor mediatie-analyse in meer complexe toepassingen waarbij men pad-specifieke effecten via meerdere mediatoren wenst te ontwarren (Steen et al., 2016a). We tonen aan dat deze methode onderzoekers beter in staat stelt mogelijke interacties tussen verschillende mechanismen in kaart te brengen dan bestaande analytische methodes (VanderWeele and Vansteelandt, 2013) en bovendien het risico op incorrecte modelassumpties reduceert ten opzichte van volledig parametrische alternatieven (Daniel et al., 2015).

In **hoofdstuk 6**, ten slotte, kaderen we de voorgaande hoofdstukken binnen recente ontwikkelingen in de mediatie-analyse literatuur, belichten we de voor- en nadelen van de voorgestelde modelleringsaanpak en bespreken we uitdagingen voor toekomstig onderzoek.

Bibliography

Albert, J. M. (2008). Mediation Analysis Via Potential Outcomes Models. *Statistics in Medicine*, 27:1282–1304.

Albert, J. M. (2012). Distribution-Free Mediation Analysis for Nonlinear Models with Confounding. *Epidemiology*, 23(6):879–88.

Albert, J. M., Geng, C., and Nelson, S. (2016). Causal mediation analysis with a latent mediator. *Biometrical Journal*, 58(3):535–548.

Albert, J. M. and Nelson, S. (2011). Generalized Causal Mediation Analysis. *Biometrics*, 67(3):1028–38.

Albert, J. M. and Wang, W. (2015). Sensitivity analyses for parametric causal mediation effect estimation. *Biostatistics*, 16(2):339–351.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of Path-Specific Effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 357–363, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Baron, R. M. and Kenny, D. A. (1986). The ModeratorMediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Berkson, J. (1946). Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin*, 2(3):47–53.

Bind, M.-a. C., VanderWeele, T. J., Coull, B. a., and Schwartz, J. D. (2015). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics*, page kxv029.

Bullock, J. G., Green, D. P., and Ha, S. E. (2010). Yes, But What's the Mechanism? (Don't Expect an Easy Answer). *Journal of Personality and Social Psychology*, 98(4):550–558.

Burns, B. D. and Wieth, M. (2004). The Collider Principle in Causal Reasoning: Why the Monty Hall Dilemma Is So Hard. *Journal of Experimental Psychology: General*, 133(3):434–449.

Canty, A. and Ripley, B. D. (2015). *boot: Bootstrap R (S-PLUS) Functions*. R package version 1.3-17.

Chiba, Y. and Taguri, M. (2013). Alternative Monotonicity Assumptions for Improving Bounds on Natural Direct Effects. *International Journal of Biostatistics*, 9(2):235–249.

Daniel, R. M., De Stavola, B. L., and Cousens, S. N. (2011). gformula: Estimating Causal Effects in the Presence of Time-Varying Confounding or Mediation Using the G-computation Formula. *Stata Journal*, 11(4):479–517.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal Mediation Analysis with Multiple Mediators. *Biometrics*, 71(1):1–14.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

De Smet, O., Loeys, T., and Buysse, A. (2012). Post-Breakup Unwanted Pursuit: A Refined Analysis of the Role of Romantic Relationship Characteristics. *Journal of Family Violence*, 27(5):437–452.

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., and Micali, N. (2014). Mediation Analysis With Intermediate Confounding: Structural Equation

Modeling Viewed Through the Causal Inference Lens. *American Journal of Epidemiology*, 181(1):64–80.

Didelez, V. (2013a). Basic concepts of causal mediation analysis and some extensions. Talk at Symposium on causal mediation analysis, Ghent.

Didelez, V. (2013b). Discussion on the paper by Imai, Tingley and Yamamoto. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 176(1):39.

Didelez, V., Dawid, A. P., and Geneletti, S. (2006). Direct and Indirect Effects of Sequential Treatments. In Dechter, R. and Richardson, T., editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 138–146, Arlington, Virginia. AUAI Press.

Ding, P. and VanderWeele, T. J. (2016). Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika*, (April):asw012.

Elwert, F. (2013). Graphical Causal Models. In Morgan, S. L., editor, *Handbook of Causal Analysis for Social Research*, Handbooks of Sociology and Social Research. Springer Netherlands, Dordrecht.

Emsley, R. and Liu, H. (2013). PARAMED: Stata Module to Perform Causal Mediation Analysis Using Parametric Regression Models.

Evans, R. J. (2012). Graphical methods for inequality constraints in marginalized DAGs. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69(2):199–215.

Ghent University and Catholic University of Louvain (2010). Interdisciplinary project for the optimisation of separation trajectories - divorce and separation in flanders.

Greenland, S. (2003). Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias. *Epidemiology*, 14(3):300–306.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Hafeman, D. M. (2011). Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. *American Journal of Epidemiology*, 174(6):710–7.

Hastie, T. (2015). *gam: Generalized Additive Models*. R package version 1.12.

Hayes, A. F. and Preacher, K. J. (2010). Quantifying and Testing Indirect Effects in Simple Mediation Models When the Constituent Paths Are Nonlinear. *Multivariate Behavioral Research*, 45(4):627–660.

Hayes, A. F. and Preacher, K. J. (2014). Statistical Mediation Analysis with a Multicategorical Independent Variable. *The British Journal of Mathematical and Statistical Psychology*, 67:451–470.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A Structural Approach to Selection Bias. *Epidemiology*, 15(5):615–625.

Hicks, R. and Tingley, D. (2011). Causal Mediation Analysis. *The Stata Journal*, 11(4):1–15.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hong, G. (2010). Ratio of Mediator Probability Weighting for Estimating Natural Direct and Indirect Effects. In *Proceedings of the American Statistical Association, Biometrics Section*, pages 2401–2415, Alexandria, VA. American Statistical Association.

Hong, G., Deutsch, J., and Hill, H. D. (2015). Ratio-of-Mediator-Probability Weighting for Causal Mediation Analysis in the Presence of Treatment-by-Mediator Interaction. *Journal of Educational and Behavioral Statistics*, 40(3):307–340.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363.

Huang, Y. and Valtorta, M. (2006). Identifiability in causal bayesian networks: A sound and complete algorithm. *Proceedings of the National Conference on Artificial Intelligence*, 21(2):1149.

Huber, M. (2014). Identifying Causal Mechanisms (Primarily) Based on Inverse Probability Weighting. *Journal of Applied Econometrics*, 29(6):920–943.

Huber, M., Lechner, M., and Mellace, G. (2016). The Finite Sample Performance of Estimators for Mediation Analysis Under Sequential Conditional Independence. *Journal of Business & Economic Statistics*, 34(1):139–160.

Iacobucci, D. (2012). Mediation Analysis and Categorical Variables: The Final Frontier. *Journal of Consumer Psychology*, 22(4):582–594.

IBM Corporation (2013). *IBM SPSS Statistics, Version 22.0*. IBM Corporation, Armonk, NY.

Imai, K., Keele, L., and Tingley, D. (2010a). A General Approach to Causal Mediation Analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2014). Comment on Pearl: Practical Implications of Theoretical Results for Causal Mediation Analysis. *Psychological Methods*, 19(4):482–487.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1):51–71.

Imai, K., Tingley, D., and Yamamoto, T. (2013). Experimental Designs for Identifying Causal Mechanisms. *Journal of the Royal Statistical Society A*, 176(1):5–51.

Imai, K. and Yamamoto, T. (2013). Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments. *Political Analysis*, 21(2):141–171.

Jiang, Z. and VanderWeele, T. J. (2015). Jiang and VanderWeele Respond to "Bounding Natural Direct and Indirect Effects". *American Journal of Epidemiology*, 182(2):115–117.

Josephy, H., Vansteelandt, S., Vanderhasselt, M.-A., and Loeys, T. (2015). Within-Subject Mediation Analysis in AB/BA Crossover Designs. *The International Journal of Biostatistics*, 11(1):1–22.

Judd, C. M. and Kenny, D. A. (1981). Process Analysis: Estimating Mediation in Treatment Evaluations. *Evaluation Review*, 5(5):602–619.

Kaufman, J. S. (2010). Invited commentary: Decomposing with a lot of supposing. *American Journal of Epidemiology*, 172(12):1349–51; discussion 1355–6.

Kaufman, S., Kaufman, J. S., and MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *Journal of Statistical Planning and Inference*, 139(10):3473–3487.

Kim, J. H. and Pearl, J. (1983). A Computational Model for Causal and Diagnostic Reasoning in Inference Systems. *Proceedings of the Eighth international joint conference on Artificial intelligence*, pages 190–193.

Lange, T., Rasmussen, M., and Thygesen, L. C. (2014). Assessing Natural Direct and Indirect Effects Through Multiple Pathways. *American Journal of Epidemiology*, 179(4):513–8.

Lange, T., Vansteelandt, S., and Bekaert, M. (2012). A Simple Unified Approach for Estimating Natural Direct and Indirect Effects. *American Journal of Epidemiology*, 176(3):190–195.

le Cessie, S. (2016). Bias Formulas for Estimating Direct and Indirect Effects When Unmeasured Confounding Is Present. *Epidemiology*, 27(1):125–132.

le Cessie, S., Debeij, J., Rosendaal, F. R., Cannegieter, S. C., and Vandenbroucke, J. P. (2012). Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*, 23(4):551–60.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1):13–22.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., Steen, J., and Vansteelandt, S. (2013). Flexible Mediation Analysis in the Presence of Nonlinear Relations: Beyond the Mediation Formula. *Multivariate Behavioral Research*, 48(6):871–894.

Loeys, T., Moerkerke, B., Raes, A., Rosseel, Y., and Vansteelandt, S. (2014). Estimation of Controlled Direct Effects in the Presence of Exposure-Induced Confounding and Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3):396–407.

Lok, J. J. (2016). Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. *Statistics in Medicine*, (October 2015).

Lumley, T. (2014). *mitools: Tools for multiple imputation of missing data*. R package version 2.3.

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

MacKinnon, D. P. and Dwyer, J. H. (1993). Estimating Mediated Effects in Prevention Studies. *Evaluation Review*, 17(2):144–158.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., and Hoffman, J. M. (2007). The Intermediate Endpoint Effect in Logistic and Probit Regression. *Clinical Trials*, 4:499–513.

Mayer, A., Thoemmes, F. J., Rose, N., Steyer, R., and West, S. G. (2014). Theory and Analysis of Total, Direct, and Indirect Causal Effects. *Multivariate Behavioral Research*, 49(5):425–442.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4):538–558.

Miles, C., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2015). On Partial Identification of the Pure Direct Effect. *Harvard University Biostatistics Working Paper Series*, page Paper 196.

Miles, C., Shpitser, I., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2014). Quantifying an Adherence Path-Specific Effect of Antiretroviral Therapy in the Nigeria PEPFAR Program. *Harvard University Biostatistics Working Paper Series*, pages 1–42.

Muller, D., Judd, C. M., and Yzerbyt, V. Y. (2005). When Moderation Is Mediated and Mediation Is Moderated. *Journal of Personality and Social Psychology*, 89(6):852–63.

Muthén, B. and Asparouhov, T. (2015). Causal Effects in Mediation Modeling: An Introduction with Applications to Latent Variables. *Structural Equation Modeling*, 22(1):12–23.

Muthén, L. K. and Muthén, B. O. (1998-2012). *Mplus User's Guide. Seventh Edition.* Muthén & Muthén, Los Angeles, CA.

Naimi, A. I. (2015). Invited Commentary: Boundless Science–Putting Natural Direct and Indirect Effects in a Clearer Empirical Context. *American Journal of Epidemiology*, 182(2):109–114.

Naimi, A. I., Kaufman, J. S., and MacLehose, R. F. (2014a). Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects. *International Journal of Epidemiology*, 43(5):1656–1661.

Naimi, A. I., Moodie, E. E. M., Auger, N., and Kaufman, J. S. (2014b). Stochastic Mediation Contrasts in Epidemiologic Research: Interpregnancy Interval and the Educational Disparity in Preterm Delivery. *American Journal of Epidemiology*, 180(4):436–445.

Nguyen, Q. C., Osypuk, T. L., Schmidt, N. M., Glymour, M. M., and Tchetgen Tchetgen, E. J. (2015). Practical Guidance for Conducting Mediation Analysis With Multiple Mediators Using Inverse Odds Ratio Weighting. *American Journal of Epidemiology*, 181(5):349–356.

Nguyen, T. T., Tchetgen Tchetgen, E. J., Kawachi, I., Gilman, S. E., Walter, S., and Glymour, M. (2016). Comparing alternative effect decomposition methods. *Epidemiology*, 27(5):1.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems.

Pearl, J. (1993). Comment: Graphical Models, Causality and Intervention. *Statistical Science*, 8(3):266–269.

Pearl, J. (1995a). Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–710.

Pearl, J. (1995b). On the Testability of Causal Models with Latent and Instrumental Variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 435–443, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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

Pearl, J. (2001). Causal inference in statistics: a gentle introduction. In *Computing Science and Statistics*, volume 33, pages 1–20.

Pearl, J. (2012). The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models. In C. Berzuini, Dawid, P., and Bernardinelli, L., editors, *Causality: Statistical Perspectives and Applications*, number October 2011, pages 151–179. John Wiley & Sons, Chichester, UK.

Pearl, J. (2014). Interpretation and Identification of Causal Mediation. *Psychological Methods*, 19(4):459–481.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2015). A Complete Generalized Adjustment Criterion. In Meila, M. and Heskes, T., editors, *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pages 682–691, Corvallis, Oregon. UAUI Press.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of Direct Causal Effects. *Epidemiology*, 17(3):276–84.

Polley, E. and van der Laan, M. (2014). *SuperLearner: Super Learner Prediction*. R package version 2.0-15.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions. *Multivariate Behavioral Research*, 42(1):185–227.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richardson, T. (2009). A factorization criterion for acyclic directed mixed graphs. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 462–470.

Richardson, T. S. (2003). Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1):145–157.

Richardson, T. S. and Robins, J. M. (2013). Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(912):1393–1512.

Robins, J. M. (1999). Association, Causation, And Marginal Structural Models. *Synthese*, 121(1):151–179.

Robins, J. M. (2003). Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects. In Green, P., Hjort, N., and Richardson, S.,

editors, *Highly Structured Stochastic Systems*, pages 70–81. Oxford University Press, New York.

Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, 3(2):143–155.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–60.

Robins, J. M. and Richardson, T. S. (2010). Alternative Graphical Causal Models and the Identification of Direct Effects. In Shrout, P., editor, *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, pages 103–158. Oxford University Press, Oxford, England.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine*, 127(8):757–763.

SAS Institute Inc. (2014). *SAS/STAT 13.2*. SAS Institute Inc., Cary, NC.

Shenassa, E. D., Daskalakis, C., Liebhaber, A., Braubach, M., and Brown, M. (2007). Dampness and mold in the home and depression: an examination of mold-related illness and perceived control of one's home as possible depression pathways. *American Journal of Public Health*, 97(10):1893–9.

Shipley, B. (2002). Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference. *Cambridge University Press*, 2:332.

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37(6):1011–35.

Shpitser, I. and Pearl, J. (2006a). Identification of Conditional Interventional Distributions. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444.

Shpitser, I. and Pearl, J. (2006b). Identification of Joint Interventional Distributions in Recursive semi-Markovian Causal Models. *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, pages 1219–1226.

Shpitser, I. and Pearl, J. (2008a). Complete identification methods for the causal hierarchy. *The Journal of Machine Learning Research*, 9:1941–1979.

Shpitser, I. and Pearl, J. (2008b). Dormant Independence. *Proceedings of the Twenty-Third Conference on Artificial Intelligence*, (April):1081–1087.

Shpitser, I., Richardson, T. S., and Robins, J. M. (2009). Testing edges by truncations. *IJCAI International Joint Conference on Artificial Intelligence*, 1:1957–1963.

Shpitser, I. and VanderWeele, T. J. (2011). A Complete Graphical Criterion for the Adjustment Formula in Mediation Analysis. *The International Journal of Biostatistics*, 7(1):1–24.

Shpitser, I., VanderWeele, T. J., and Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)* , pages 527–536.

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, 28(4):558–571.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–8.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

Splawa-Neyman, J., Dabrowska, D., and Speed, T. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472.

StataCorp (2013). *Stata Statistical Software: Release 13*. StataCorp LP, College Station, TX.

Steen, J., Loeys, T., Moerkerke, B., and Vansteelandt, S. (2015). *medflex: Flexible Mediation Analysis Using Natural Effect Models*. R package version 0.6-1.

Steen, J., Loeys, T., Moerkerke, B., and Vansteelandt, S. (2016a). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, in press.

Steen, J., Loeys, T., Moerkerke, B., and Vansteelandt, S. (2016b). Medflex: An R Package for Flexible Mediation Analysis Using Natural Effect Models. *Journal of Statistical Software*, in press.

Taguri, M., Featherstone, J., and Cheng, J. (2015). Causal mediation analysis with multiple causally non-ordered mediators. *Statistical Methods in Medical Research*, (January):1–14.

Talloen, W., Moerkerke, B., Loeys, T., De Naeghel, J., Van Keer, H., and Vansteelandt, S. (2016). Estimation of Indirect Effects in the Presence of Unmeasured Confounding for the Mediator-Outcome Relationship in a Multilevel 2-1-1 Mediation Model. *Journal of Educational and Behavioral Statistics*, 41(4):359–391.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22(4):560–568.

Taylor, A. B., MacKinnon, D. P., and Tein, J.-Y. (2007). Tests of the Three-Path Mediated Effect. *Organizational Research Methods*, 11(2):241–269.

Tchetgen Tchetgen, E. J. (2013). Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis. *Statistics in Medicine*, 32(26):4567–80.

Tchetgen Tchetgen, E. J. (2014). A Note on Formulae for Causal Mediation Analysis in an Odds Ratio Context. *Epidemiologic Methods*, 2(1):21–31.

Tchetgen Tchetgen, E. J. and Phiri, K. (2014). Bounds for pure direct effect. *Epidemiology*, 25(5):775–6.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2011). Semiparametric Estimation of Models for Natural Direct and Indirect Effects.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, 40(3):1816–1845.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2014). Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2014). Identification of Natural Direct Effects When a Confounder of the Mediator Is Directly Affected by Exposure. *Epidemiology*, 25(2):282–91.

Textor, J., Hardt, J., and Knüppel, S. (2011). DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology*, 22(5):745.

Tian, J. and Pearl, J. (2002). A General Identification Condition for Causal Effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI)*, pages 567–573, Menlo Park, CA. AAAI Press/The MIT Press.

Tian, J. and Pearl, J. (2003). On the identification of causal effects. Technical report, Department of Computer Science, University of California, Los Angeles.

Tian, J. and Shpitser, I. (2010). On identifying causal effects. *Heuristics, Probability and Causality: A Tribute to ...*.

Tikka, S. (2016). *causaleffect: Deriving Expressions of Joint Interventional Distributions and Transport Formulas in Causal Models*.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014a). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, 59(5).

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014b). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38.

Valeri, L. and VanderWeele, T. J. (2013). Mediation Analysis Allowing for Exposure-Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation with SAS and SPSS Macros. *Psychological Methods*, 18(2):137–50.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

van der Laan, M. J. and Petersen, M. L. (2008). Direct Effect Models. *The International Journal of Biostatistics*, 4(1):1–27.

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

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 21(4):540–51.

VanderWeele, T. J. (2011a). Causal Mediation Analysis with Survival Data. *Epidemiology*, 22(4):582–585.

VanderWeele, T. J. (2011b). Controlled Direct and Mediated Effects: Definition, Identification and Bounds. *Scandinavian Journal of Statistics*, 38(3):551–563.

VanderWeele, T. J. (2013). Policy-relevant proportions for direct effects. *Epidemiology*, 24(1):175–6.

VanderWeele, T. J. and Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, Biostatistics and Public Health*, 11(2):1–16.

Vanderweele, T. J. and Tchetgen Tchetgen, E. J. (2014). Mediation Analysis with Time-Varying Exposures and Mediators. *Harvard University Biostatistics Working Paper Series*, 168:1–22.

Vanderweele, T. J. and Tchetgen Tchetgen, E. J. (2016). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*.

VanderWeele, T. J., Valeri, L., and Ogburn, E. L. (2012). The role of measurement error and misclassification in mediation analysis: mediation and measurement error. *Epidemiology*, 23(4):561–4.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual Issues Concerning Mediation, Interventions and Composition. *Statistics and Its Interface*, 2(4):457–468.

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *American Journal of Epidemiology*, 172(12):1339–48.

VanderWeele, T. J. and Vansteelandt, S. (2013). Mediation Analysis with Multiple Mediators. *Epidemiologic Methods*, 2(1):95–115.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiology*, 25(2):300–306.

Vansteelandt, S. (2012). Understanding Counterfactual-Based Mediation Analysis Approaches and Their Differences. *Epidemiology*, 23(6):889–91.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012a). On Model Selection and Model Misspecification in Causal Inference. *Statistical Methods in Medical Research*, 21(1):7–30.

Vansteelandt, S., Bekaert, M., and Lange, T. (2012b). Imputation Strategies for the Estimation of Natural Direct and Indirect Effects. *Epidemiologic Methods*, 1(1):Article 7.

Vansteelandt, S. and Daniel, R. M. (2016). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, in press.

Vansteelandt, S. and VanderWeele, T. J. (2012). Natural Direct and Indirect Effects on the Exposed: Effect Decomposition Under Weaker Assumptions. *Biometrics*, 68(4):1019–1027.

Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models.

Vinokur, A. D., Price, R. H., and Schul, Y. (1995). Impact of the JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology*, 23(1):39–74.

Vinokur, A. D. and Schul, Y. (1997). Mastery and inoculation against setbacks as active ingredients in the JOBS intervention for the unemployed. *Journal of Consulting and Clinical Psychology*, 65(5):867–877.

Wang, Z. and Louis, T. A. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, 90(4):765–775.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer-Verlag, New York, USA.

Zheng, W. and van der Laan, M. J. (2012). Targeted maximum likelihood estimation of natural direct effects. *The international journal of biostatistics*, 8(1):Article 3.