Mediation Analysis : a Starting Guide for Epidemiologists

Benoit Lepage[1,2], Hélène Colineaux[1,2], Valérie Garès[3], Barbara Bodinier[4], Cyrille Delpierre[1,*], and Marc Chadeau-Hyam[4,*]

[1]INSERM CERPOP, UMR 1295, Toulouse University III Paul Sabatier, Team EQUITY
[2]CHU Toulouse, Epidemiology Department
[3]University of Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625
[4]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK
[*]Joint senior authors

March 29, 2024

Abstract

While causal thinking and mediation questions are widespread in Epidemiology, methods of causal and mediation analysis are still poorly implemented. The aim of this article is to present the main approaches and methods in a synthetic and non-technical way, to ensure that end-users such as epidemiologists could easily and quickly implement them in their practice.

We first introduce founding methods for mediation analysis (Baron and Kenny approach and path analysis approach), and subsequently describe the basis of structural causal model used for mediation analyses: introduction to the theory of causation; principles of counterfactual notation, directed acyclic graph and non parametric structural equations. We describe the various causal quantities of interest suggested to measure direct and indirect effects, their identification assumptions and the corresponding estimands expressed as parameters of the observed statistical distribution through the G-formula. We finally describe the main estimation methods. Using simulated data, R scripts are provided in supplementary online documents, giving examples for the various estimators of quantities of interest in mediation analyses.

Keywords: causality, mediation analysis, mediators, path analysis, g-computation, IPTW, total effect, direct effect, indirect effect

Table of Content

1 Introduction

1.1 Why a mediation analysis ?

The objectives of quantitative epidemiology can be classified into descriptive, predictive or etiological analyses [Hernán et al., 2019]. Causal and mediation analyses are part of the latter family and aim at assessing whether a factor is associated with an outcome by a *cause-and-effect* relationship, therefore informing on the mechanisms underlying these causal relationships. Etiological analyses also aim at predicting the risk of a certain outcome if risk factors were to be changed. This defines the scope of *counterfactual prediction* [Hernán et al., 2019].

Traditionally in Epidemiology, causality is studied using specific experimental designs such as randomized controlled trials (RCT). In non-experimental designs, causal relationships are investigated using sets of arguments such as Bradford Hill's criteria or triangulation approaches [Hill, 1965; Lawlor et al., 2016]. Various models have been suggested to define formal frameworks and tools for the assessment of causal effects from observational data: sufficient cause and component causes, potential outcomes, counterfactuals, non-parametric structural equation models, etc [Rothman et al., 2008; Pearl, 2009a].

The definition of causation given by [Pearl and Mackenzie, 2018] is: "a variable A is a cause of Y if Y "listens" to A and determines its value in response to what it hears". This could be interpreted as "changing the value of A results in a change in the values of Y" (all others things being unchanged) and on the contrary, values of A would not change had the values of Y been changed. In epidemiology, the corresponding question can be formulated as: "Does the exposure A *causes* the outcome Y?", or "if one changes (intervenes on) the exposure, will the outcome change?".

From observational studies one can assess if "the probability of having the outcome Y (or the value of the outcome Y) is different between those who have been exposed to $A = 1$ compared to those who have not been exposed ($A = 0$), all other things being equal?". In other word we assess the association between the exposure and the outcome. The fact that Y depends on A, i.e. that the variables are associated, is based on their joint distribution, and should not be over-interpreted as causation. In practice, when controlling for confounding, causal assumptions are being made [Pearl, 2010b]. According to [Reichenbach, 1991], there are two ways to explain an association between two variables: either one is the cause of the other, or a third variable is the cause of both (hypothesis of common cause). Hence, based on our background knowledge and temporal order regarding an exposure and an outcome, we could test the existence of independent and *causal* relationships between the exposure and the outcome by controlling for

potential common causes (i.e. confounders). Such an approch places us in a causal framework. Maria Glymour adds two other phenomena to the explanations given by Reichenbach that can explain spurious (non-causal) associations between two variables: random fluctuations (which could be handled using large samples and statistical tools), and conditioning on a third variable influenced by both variables (named "collider stratification bias" in the literature) [Glymour, 2017]. These different mechanisms will be detailed below.

So why mediation analyses? If a cause-and-effect relationship is established between the income and the probability of heart attack, a new explanatory objective can be to study if this effect *goes through* ("is mediated by") a higher risk of smoking, or a higher risk of being exposed to occupational toxic agents. To address this question, mediation analyses can be used to assess the role of intermediate variables (mediators) of interest. Mediation analyses have been applied for over 30 years in psychology and social sciences, mainly based on the work of authors such as Baron and Kenny, or MacKinnon [Baron and Kenny, 1986; MacKinnon, 2008]. These approaches were subsequently enriched by Rubin's counterfactual framework, and by Pearl's development on causality including directed acyclic graphs (DAG) and non-parametric structural equations [Rubin, 1991; Pearl, 2001].

In recent years, VanderWeele and other authors also formalised a number of methods to implement, defined new direct or indirect effects of interest, and dealt with potential difficulties encountered in practice, such as confounding, exposure-mediator interactions, multiple mediators, and estimation challenges (see for example [VanderWeele, 2016]).

These causal approaches are particularly important with the current explosion of medical or administrative data: it may not be ethical/feasible to implement expensive and risky RCT before exploring available "real-life" evidence from existing datasets covering thousands of people. In addition, observational study can explore non-randomisable exposures, multifactorial events, mediation effect, life-course causality chain, which are by construction overlooked in RCT. Epidemiologists have therefore to take up causality and mediation analysis methods to face contemporary challenges of the discipline. The slow uptake of these methods might be explained by the extensive, technical and diverse, bibliography describing causal and mediation methods. We hope that this report will present the different approaches in a synthetic and didactic way and will ultimately facilitate their implementation in practice by epidemiologists.

1.2 Notations and Examples

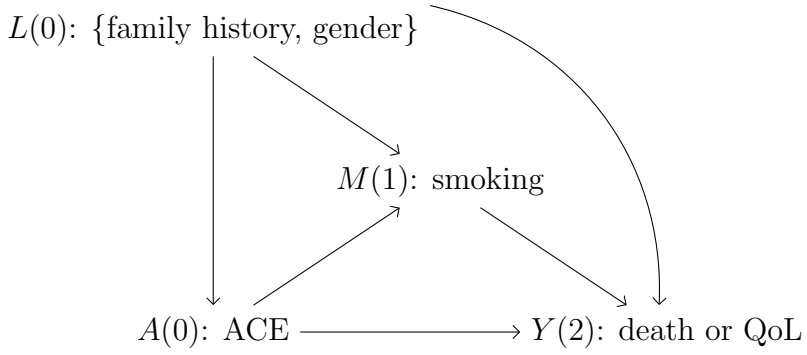
By convention, we note A an exposure of interest (also called "intervention" or "treatment") and Y the outcome. The mediator of interest will be denoted M. Temporal ordering assumptions are necessary in mediation analyses, and we let t denote the timing of a measurement. The

baseline measurement is therefore noted $A(0)$, the outcome $Y(2)$ and the mediator of interest is denoted $M(1)$. As $M(1)$ can also be considered as an exposure of interest, it might be denoted $A(1)$ in other contexts of repeated exposures. The set $L(t)$ is used to denote other covariates occurring at time t, assuming that $L(t)$ covariates precede the exposure of interest $A(t)$ or $M(t)$. $L(t)$ includes confounders or other relevant variables to consider in the underlying causal assumptions. A good practice is to include any variables (even unmeasured variables) that might affect at least two other variables in the causal model. Some authors might also include the outcome $Y(t)$ in the set $L(t)$ when describing very general causal structures with repeated measures of the exposure, mediator and outcome along time [Petersen et al., 2014; Zheng and van der Laan, 2017]. In this report, we will focus on more limited structures including only one time of measurement for the exposure A, the mediator M and the outcome Y.

As a working example, we consider the research question inspired by previous work on data from the NCDS58 birth cohort [Kelly-Irving et al., 2013]. We investigate if childhood adversity has an impact on adult health status, and if so, if it is mediated by intermediate behaviour such as smoking. In the following examples, we simulate data for baseline adverse childhood experience (ACE, $A(0)$) in relation to an health outcome of interest Y. We simulated quality of life (QoL, continuous outcome), or the death before 50 years of age (as a binary outcome). In our example, we want to assess if the link between ACE and health can be explained, at least partially, by the fact that people who experiment an adverse event during childhood are more likely to be exposed to unhealthy behaviour (such as smoking), so we used the variable "smoking at 23 years of age" as mediator, denoted $M(1)$. The set of baseline confounders will be denoted $L(0)$ (in our example, gender and parent's educational level). The R scripts used to simulate the data are available on a web site (under construction).

For example, assuming we did not forget any variable that might affect at least two of the variables in the causal model, the causal links between $L(0)$, $A(0)$, $M(1)$ and Y can be summarized in a directed acyclic graph (DAG) such as in the model of Figure 1.

Figure 1: Directed acyclic graph summarizing our example



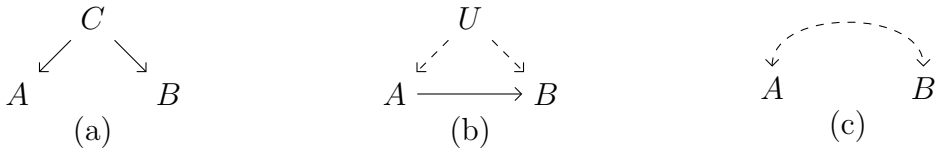
1.3 Directed Acyclic Graph

Causal relationships, which are directed from a cause to an effect, can not be formulated with equations alone, which by nature can only describe symmetrical relationships. [Wright, 1921] suggested to combine graphs with parametric equations in order to formulate the directionality of these relationships. These graphs are the "path diagrams", associated with the structural equations described below (equations (5) and Figure 8).

Beyond the "path diagram" framework, Directed Acyclic Graphs (DAGs) can be used to represent nonparametric structural equations [Pearl, 2009a; Tennant et al., 2020]. By definition, DAGs have the following principles:

- All the links are directed: every edge in a path is an arrow that points from one variable to the other. In the case of unobserved confounders, a two-way arrow (as in Figure 2(c)) represents the two unobserved arrows arising from a common unmeasured cause.
- The graph is acyclic: a DAG does not contain loops.

Figure 2: Examples of simple causal relationships represented using DAGs



For a DAG to be complete and correctly interpreted:

- Every arrow $V \rightarrow W$ is interpreted as a "possible" effect of V on W, the absence of an arrow from a variable V to another W is a strong and explicit statement that there is no direct effect of V on W. In other words, if one is uncertain about a possible effect of V on

W, it is better to include an arrow from V to W. For example, in Figure 2a, the explicit statement is "there is no direct effect from A to B, nor from B to A". On the contrary, Figure 2b indicates that a direct effect of A on B is possible.

- Every common cause of two variables represented in a DAG should also appear in the DAG, even if the common cause is unmeasured in the observed data. In the literature, such unmeasured common causes are usually represented with U variables (for *unknown*), with dashed arrows or dashed double arrows (assuming a common unknown cause is present between the double arrows, see Figure 2b and c). Figures 2a and 2c represent the same causal structure, the only difference is that the variable C is measured in Figure 2a and unmeasured in Figure 2c.

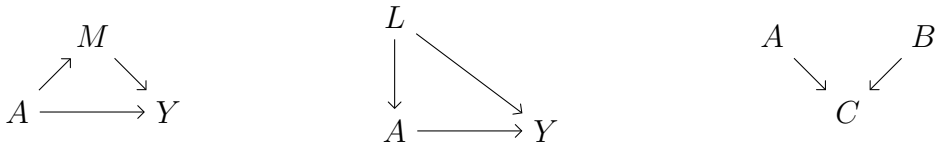
Kinship terminology is generally used to describe the relationships between the variables in a DAG. For example in the DAG of Figure 1, $Y(2)$ is a *child* of $M(1)$ and $M(1)$ is a *parent* of $Y(2)$. $A(0)$ and $M(1)$ are *ancestors* of $Y(2)$, and $M(1)$ and $Y(2)$ are *descendents* of $A(0)$.

Interpreting DAGs as *causal* networks relies on one additional strong assumption according to which each parent-child relationships in the DAG represents a stable (over time and circumstances) and autonomous physical mechanism [Pearl, 2009a, 2010b]. This implies that changing one relationship without changing the others is conceivable (a relationship is unaffected by possible changes in the form of other functions).

DAGs can be used as a convenient way to formulate and visualise possible causal relationships and independence assumptions. For example in Figure 3, we can represent:

- a direct effect of A on Y ($A \to Y$ in Figure 3a and 3b),
- an indirect effect of A on Y through M (M is a mediator of the effect of A on Y, $A \to M \to Y$ in Figure 3a),
- a back-door path between A and Y ($A \leftarrow L \to Y$ in Figure 3b),
- a collider C on the path between A and B (Figure 3c)

Figure 3: Direct effets, indirect effects, backdoor paths, and colliders



(a) (b) (c)

DAGs can also help to deduce graphically what are the expected independences and conditional independences using the *d-separation* criterion [Pearl, 1988, 1995]. For a DAG compatible with the data set under study, two variables V and W are said to be *d-separated* by a set of

variables Z if all the paths between them are blocked conditional on Z. Such d-separation by Z implies that V and W are independent given Z. Graphically, a path connecting two variables V and W is "blocked" conditional on Z if:

- there is a variable on the path which belongs to Z and which is not a collider,
- there is a *collider* on the path and the collider or any of its descendent does not belong to Z.

On the contrary,

- V and W are *d-connected* if there is an unblock path between them;
- V and W are *d-connected* conditional on Z if there is a collider-free path that does not contain any variable of the set Z.

.

Using the DAG in Figure 4, some definitions and examples of independences and conditional independences that can be deduced by the d-separation criterion are given below, .

Figure 4: Exemples of d-separation/d-connectedness:

From the Figure 4,

- Path: a sequence of unique nodes linked by arrows (regardless of their directionality)
- A path is unblocked if it does not contain any collider.
  - The path $B \rightarrow C \rightarrow E$ is unblocked,
  - the path $E \leftarrow C \leftarrow D$ is unblocked,
  - the paths $A \leftarrow B \rightarrow C$ and $A \leftarrow B \rightarrow C \rightarrow E$ are unblocked. Both are defined as "back-door paths" (paths starting with an arrow pointing at the first variable).
  - But the paths $A \leftarrow B \rightarrow C \leftarrow D$ and $B \rightarrow C \leftarrow D$ are blocked (C is a collider on those paths).
- Conditional on variables of a path which are not colliders:
  - Conditional on C, the three paths $B \rightarrow C \rightarrow E$, $E \leftarrow C \leftarrow D$ and $A \leftarrow B \rightarrow C \rightarrow E$ are blocked.
  - Conditional on B, both paths $A \leftarrow B \rightarrow C$ and $A \leftarrow B \rightarrow C \rightarrow E$ are blocked.
- Conditional on variables of a path which are colliders (or descendents of colliders):
  - Conditional on C, both paths $A \leftarrow B \rightarrow C \leftarrow D$ and $B \rightarrow C \leftarrow D$ are unblocked (because conditioning on the collider C "creates" a spurious correlation between B and D)

  - Similarly, conditional on E (a descendent of the collider C), both paths $A \leftarrow B \rightarrow C \leftarrow D$ and $B \rightarrow C \leftarrow D$ are unblocked.
- Two nodes are d-separated if there is no unblocked path between them.
  - A and D are d-separated,
  - B and D are d-separated.
- Conditional on B:
  - A and C are d-separated,
  - A and E are d-separated,
  - and A and D are d-separated.
- Conditional on C:
  - A and E are d-separated,
  - B and E are d-separated,
  - and E and D are d-separated.
  - B and D are not d-separated (conditioning on the collider C "creates" a spurious correlation between B and D),
  - A and D are not d-separated.
- Conditional on E (a descendent of the collider C):
  - B and D are not d-separated,
  - A and D are not d-separated.
- Conditional on $\{B, C\}$:
  - A and D are d-separated (the spurious correlation between B and D resulting from conditioning on C is blocked eventually by conditioning on B).
- Conditional on $\{B, E\}$:
  - A and D are d-separated (the spurious correlation between B and D resulting from conditioning on E is blocked eventually by conditioning on B).
- In the previous examples, two variables are d-connected if they are not d-separated.
  - For example, A and C are d-connected.
  - Conditional on C (or E), A and D are d-connected.

It is possible to test the compatibility between a DAG and a data-set using the d-separation criterion (several DAGs can usually be compatible with a single data-set in terms of independence and conditional independence). For example, the R package DAGitty can be used to evaluate the testable implications associated with a given DAG and assess if the DAG is consistent with a data-set [Textor et al., 2016].

Many useful applications have been described using the d-separation criterion to explain and give new structural definitions of classical biases in epidemiology [Glymour, 2006]: confounding bias [Greenland et al., 1999; Glymour et al., 2005; Lepage et al., 2015], selection bias [Hernàn et al., 2004; Daniel et al., 2012], measurement bias [Hernàn and Cole, 2009]. They have also

been used to clarify the definition of interaction and effect modification in a causal context [VanderWeele and Robins, 2007; VanderWeele, 2009b]. For example, selection bias is often described by stratification on a collider. Structural definitions of measurement bias imply various measurement models with latent nodes or unmeasured nodes.

Graphs can help to evaluate if a causal effect of interest is identifiable from observational data, under the assumptions depicted in the DAG. For example, the *backdoor criterion* can be used to select covariate adjustment sets. This criterion is formulated as: "Z is a sufficient adjustment set in order to test and estimate the effect of A on Y if : (i) no variable in the set Z is a descendant of A and (ii) each backdoor path between A and Y is blocked, where *backdoor paths* are defined as the path between A and Y with an arrow pointing to A (they correspond to the paths that would confound the A to Y relationship if they were not blocked). A step-by-step method has been described to apply this criterion [Greenland et al., 1999]. For example, assuming one of the DAGs in Figure 5 represent the "true" causal model (data generating model), in order to estimate the effect of A on Y:

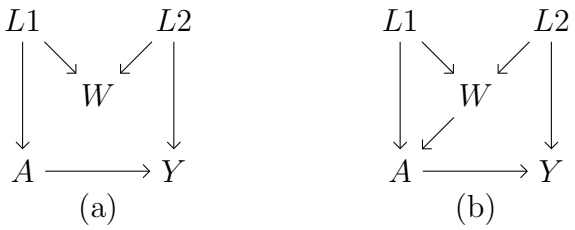
- if the causal model in Figure 5(a) is correct:
  - the empty set $Z = \{\varnothing\}$ is sufficient (because there is no open backdoor path between A and Y if we do not condition on any variable);
  - the set $Z = \{W\}$ is not sufficient, because conditioning only on W would open an unblocked backdoor path between A and Y;
  - $Z = \{L1, W\}$, $Z = \{L2, W\}$ and $Z = \{L1, L2, W\}$ are different sufficient adjustment sets;
- if the causal model in Figure 5(b) is correct:
  - the empty set $Z = \{\varnothing\}$ is not sufficient because there is an open backdoor path between A and Y ($A \leftarrow W \leftarrow L2 \rightarrow Y$);
  - $Z = \{L1, W\}$, $Z = \{L2, W\}$ and $Z = \{L1, L2, W\}$ are sufficient adjustment sets.

In these examples, we can distinguish:

- adjustment sets that add more bias than they reduce ($Z = \{W\}$ in Figure 5(a) and 5(b)),
- different sets, all valid to identify the effect of A on Y ($Z = \{L1, W\}$, $Z = \{L2, W\}$ and $Z = \{L1, L2, W\}$ in Figure 5(a) and 5(b)),
- minimally sufficient adjustment sets ($Z = \{\varnothing\}$ in Figure 5(a) ; $Z = \{L1, W\}$ or $Z = \{L2, W\}$ in Figure 5(b)).

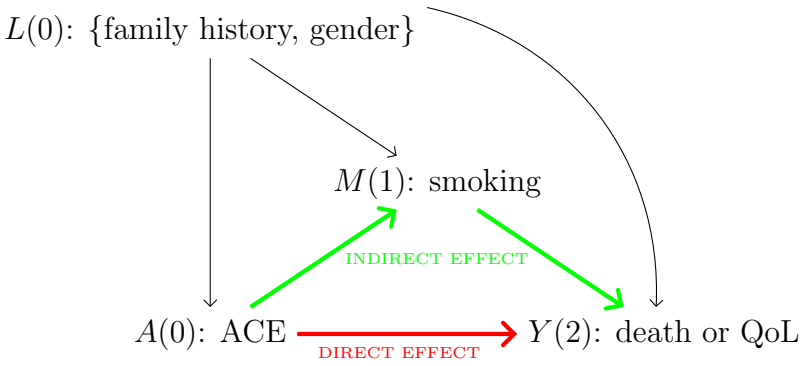
Interestingly in these examples, the notion of "confounder" as a variable does not operate, rather there are different sets of covariates to adjust for, which can be sufficient or not to remove confounding bias and identify the causal effect of interest.

Figure 5: Example for the backdoor criterion



Back to our first mediation example represented in Figure 1: the DAG formulates the following explicit assumptions: $A(0)$ might cause $Y(2)$ directly or indirectly through $M(1)$ (cf. Figure 6), showing a direct and an indirect paths between $A(0)$ and $Y(2)$.

Figure 6: Directed acyclic graph summarizing our example



1.4 Non parametric structural equations - Markov factorisation

The DAG presented in Figure 6 can be formulated using the following set of non parametric structural equations:

$$\begin{cases} L(0) = f_{L(0)}\left(U_{L(0)}\right) \\ A(0) = f_{A(0)}\left(L(0), U_{A(0)}\right) \\ M(1) = f_{M(1)}\left(A(0), L(0), U_{M(1)}\right) \\ Y(2) = f_{Y(2)}\left(M(1), A(0), L(0), U_{Y(2)}\right) \end{cases} \tag{1}$$

where all residual terms U are assumed to be independent from each other. Note that these residuals can be considered as unmeasured exogenous variables affecting each of the endogenous variables. As we assumed their mutual independence here, it is not necessary to represent them in the DAG.

The joint probability of variables represented as nodes in the DAG can be expressed as a product of (conditional) probabilities:

$$\begin{aligned} P\left[l(0), a(0), m(1), y(2)\right] = & P\left[l(0)\right] \\ & \times P\left[a(0) \mid l(0)\right] \\ & \times P\left[m(1) \mid l(0), a(0)\right] \\ & \times P\left[y(2) \mid l(0), a(0), m(1)\right] \end{aligned}$$

The (conditional) independence structure between variables encoded in this factorisation is equivalent to a representation using a DAG. It should be compatible with the data set under study through the d-separation criterion.

2 Classical methods in mediation analyses

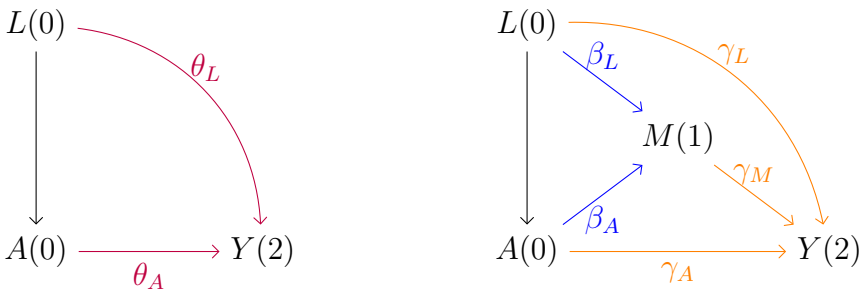
In this section, we introduce as founding methods for mediation analysis, the Baron and Kenny and the path analysis approaches.

2.1 Baron and Kenny approach

2.1.1 Principles

The Baron and Kenny approach is based on the sequential and step-wise estimation of linear regression models to model simple causal structures (Figure 7) [Baron and Kenny, 1986; Valeri, 2013].

Figure 7: Models used in the Baron and Kenny approach



(a) (b)

Confounders ($L(0)$) and mediators ($M(1)$) can be both associated with the exposures ($A(0)$) and the outcome ($Y(2)$). In a causal framework, confounders are formally defined as "common *causes* of exposure and outcome", while mediators are seen as intermediate variables along the path linking the exposure to the outcome (i.e. effect of the exposure and cause of the outcome). As such, both mediators and confounders need to be distinguished *a priori* based on conceptual considerations (as in Figure 7(b)) and irrespective of marginal statistical associations. In that setting, the Baron and Kenny approach would rely on the following steps:

1. Before investigating the mediation of the effect $A(0)$ on $Y(2)$ through $M(1)$, it is of interest to check if $A(0)$ has an effect on $Y(2)$. This *total effect* θ_A of $A(0)$ on $Y(2)$ (null hypothesis $\mathcal{H}_0 : \{\theta_A = 0\}$) can be tested using the following linear model:

$$\mathbb{E}(Y(2)) = \theta_0 + \theta_A A(0) + \theta_L L(0), \tag{2}$$

where the effect θ_A is the linear regression coefficient of $A(0)$. Based on the causal structures in Figure 7(a), we can see that adjusting for the confounders ($L(0)$) is sufficient to

identify the total effect of $A(0)$ on $Y(2)$. The variable $M(1)$ should not be used in the present model for the total effect.

2. Testing if $A(0)$ has a significant effect β_A on the intermediate variable $M(1)$ (null hypothesis $\mathcal{H}_0 : \{\beta_A = 0\}$). For $M(1)$ to qualify as an intermediate variable between $A(0)$ and $Y(2)$ one first needs to check if it *passes* from $A(0)$ to $M(1)$ using the following linear regression:

$$\mathbb{E}(M(1)) = \beta_0 + \beta_A A(0) + \beta_L L(0), \tag{3}$$

where β_A measures the effect of the exposure $A(0)$ on the mediator $M(1)$, adjusting for confounders $L(0)$.

3. Testing if the intermediate variable $M(1)$ has a significant effect γ_M on the outcome $Y(2)$ (null hypothesis $\mathcal{H}_0 : \{\gamma_M = 0\}$), independently of $A(0)$ and $L(0)$ (which are confounders of the relation $M(1) \to Y(2)$), using the linear regression:

$$\mathbb{E}(Y(2)) = \gamma_0 + \gamma_A A(0) + \gamma_M M(1) + \gamma_L L(0), \tag{4}$$

where γ_M is the effect of the intermediate variable $M(1)$ on the outcome $Y(2)$, as adjusting for $A(0)$ and $L(0)$ is sufficient to identify the effect of $M(1)$ on $Y(2)$.

From those three models, $M(1)$ is considered a mediator if the regression coefficients of the links $A(0) \to M(1)$ (Equation (3)) and $M(1) \to Y(2)$ (Equation (4)) are found statistically significant.

In equation (4), γ_A is interpreted as the "direct effect" of $A(0) \to Y(2)$, which does not pass through $M(1)$. If the null hypothesis $\mathcal{H}_0 : \{\gamma_A = 0\}$ cannot be rejected, the effect of $A(0)$ on $Y(2)$ is considered to be totally mediated by the intermediate variable $M(1)$; If the null is rejected and $\gamma_A < \theta_A$, the effect of $A(0)$ on $Y(2)$ is considered to be partially mediated by the intermediate variable $M(1)$.

2.1.2 Limitations

While this approach is intuitive, it is restricted to simple causal structures (such as the one described in the Figure 7(b)), limited to a small set of variables.

Several additional limitations have been highlighted, in particular:

- The first step does not seem to be necessary to prove the existence of a mediated effect, especially because a direct and an indirect pathway, or several mediation pathways, could cancel each other's, resulting in the non-rejection of the null hypothesis [MacKinnon, 2008].
- The method does not natively quantify the relative importance of the indirect effect ($A \to M \to Y$) in relation to the direct effect ($A \to Y$).

2.1.3 Extensions

Extensions to address the latter limitation have been proposed, in order to quantify direct and indirect effects [MacKinnon et al., 2002; Iacobucci, 2008; MacKinnon, 2008]. These extensions are also known as the "product method" or the "difference method" [Valeri, 2013].

From the models described above:

- the *total effect* of A on Y is estimated by the regression coefficient θ_A (Model (2)),
- the *direct effect* of A on Y is estimated by the regression coefficient γ_A (Model (4)),
- and the *indirect effect* of A on Y (corresponding to the path $A \to M \to Y$) is estimated by either :
  - the "difference in coefficients" $\theta_A - \gamma_A$ using the models (2) and (4)
  - or the "product of coefficients" $\beta_A \times \gamma_M$ using the models (3) and (4).

  In the particular situation where only linear least square regressions are involved with quantitatives $M(1)$ and $Y(2)$ variables, the two methods give the same estimation of the indirect effect: $\theta_A - \gamma_A = \beta_A \times \gamma_M$ [MacKinnon, 2008; Iacobucci, 2008]. In situations mixing qualitative mediators $M(1)$ and quantitative outcomes $Y(2)$, the coefficients are usually estimated on different scales so that the "product method" can not be applied, however it is still possible to use the "difference method".

Because the indirect effect is derived from two different equations, there is no straightforward way to compute standard errors or 95% confidence intervals. [MacKinnon et al., 2002] reviewed several solutions to compute the confidence intervals. Confidence intervals of the indirect effect can also be estimated using bootstrap approaches [MacKinnon, 2008].

2.2 Path analysis and structural equation modeling

2.2.1 Principles

Development of path analysis started in the early 20's, and continued in the 60's [Wright, 1921, 1960]. Whereas the Baron and Kenny approach presented above does not necessarily require a graphic representation of the causal structure, the path analysis method is more explicitly based on the combination of a graphical representation and a set of linear models and assumptions about the covariance structures of random residuals and latent variables. Path analyses have mainly been used and developed in the fields of econometric, social sciences and psychology [Tarka, 2018].

The graphical representation, called "path diagrams" depicts causal relationships among the variables of interest. Graphical rules defined by Wright have been summarized by [Loehlin, 2004]:

- A causal relationship is represented by a one-headed arrow, a cause being defined as the fact that "a change in the variable at the tail of the arrow will result in a change in the variable at the head, all else being equal". All causal connections between source and downstream variables should be included as straight arrows.
- A non-causal correlation is represented by a two-headed arrow, representing especially the situation where two variables have unmeasured common causes.
- The diagram is said complete if all the variables which does not receive causal inputs from any other variable in the diagram (called exogenous variable) are connected by double arrows, unless the assumption is explicitly made that two variables are uncorrelated because they do not have any common causal factor.
- On the contrary, variables which receive causal inputs from any other variable (called endogenous variable) are usually not connected by two-headed arrows. If there is unknown causes of these variables, these causes should rather be represented by residual arrows (unlabelled one-headed arrows). In complete diagrams, residual arrows should be attached to every downstream variables. Correlations among downstream variables caused by other variables can be represented by connecting residual arrows.
- The diagram must be complete, but also parsimonious.

Thoses graphical rules are similar to the rules applied with DAGs, however, path diagrams could also include loops and additional nodes representing some transformation of variables useful to model non-linearities (such as $L(0)^2$ or interaction terms $A(0) * M(1)$, next to the original $L(0)$, $A(0)$ and $M(1)$ nodes) [Loehlin, 2004; Iacobucci, 2008]. On the contrary, in DAGs (as a graphical representation of a joint probability model of a non parametric structural

causal model), nodes cannot include variables that are collinear to other variables already present in the DAG: the desirable equivalence between the graphical representation and the Markov factorization does not allow to represent collinear variables on a DAG.

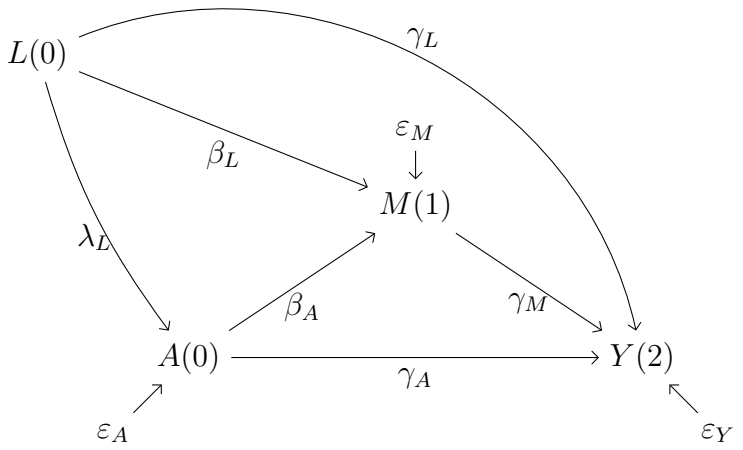
For example, the mediation structure depicted in Figure 1 can be represented by the path diagram of Figure 8 and the following set of structural equations (equations 5) completed by the assumptions about the covariance structure between the random residuals $\sigma(\varepsilon_{V_i}, \varepsilon_{V_j})$. In the present example, three endogenous variables are identified, $A(0)$, $M(1)$ and $Y(2)$. Each are modelled using specific linear models setting there direct causes as explanatory variables, and we assume that the random residuals ε_A, ε_M and ε_Y are independent from one another:

$$\begin{cases} A(0) = \lambda_0 + \lambda_L L(0) + \varepsilon_A \\ M(1) = \beta_0 + \beta_A A(0) + \beta_L L(0) + \varepsilon_M \\ Y(2) = \gamma_0 + \gamma_A A(0) + \gamma_M M(1) + \gamma_L L(0) + \varepsilon_Y \end{cases} \tag{5}$$

$$\sigma(\varepsilon_A, \varepsilon_M) = \sigma(\varepsilon_A, \varepsilon_Y) = \sigma(\varepsilon_M, \varepsilon_Y) = 0$$

The coefficients $\{\lambda_L, \beta_A, \beta_L, \gamma_A, \gamma_M, \gamma_L\}$ are called *path coefficients* and measure the direct effect of a cause on its target covariate. For example, the *path coefficient* γ_M quantifies the direct causal effect of exposure $M(1)$ on the outcome $Y(2)$. Path analyses have been developed using parametric linear models (with quantitative outcomes), so that one unit increase of $A(0)$ *causes* an increase of γ_A units of the variable Y.

Figure 8: Path diagram



The main difference between the set of equation (5) and the set of non parametric struc-

tural equations (1) is that non parametric structural equations make no assumptions about the functional form of the equations [Pearl, 2010a]. Each of the f function determines the value of the output-variables from the value of the input-variables, and these can take any form.

2.2.2 Estimation of direct and indirect paths

According to Wright, the correlation between two variables is explained by the set of all paths which link these variables. He proposed some graphical rules in order to decompose this correlation between two variables according to path coefficients and correlation between residuals. Such an analysis is called a *path analysis*.

We denote:

- σ_V^2, the variance of a variable V
- $\sigma_{V,W}$ the covariance between V and W
- $\rho_{V,W}$, the Pearson correlation between V and W

Coefficients in lower-case letter are the original (unstandardised) path coefficients in raw-score units. For example from model 5: γ_M is the unstandardised path coefficient corresponding to the *direct effect* of $M \to Y$.

Coefficients in upper-case letter are the standardised path coefficients. For example from model 5, Γ_M is the standardised path coefficient defined by $\Gamma_M = \gamma_M \frac{\sigma_M}{\sigma_Y}$:

- Using the original (unstandardised) path coefficient, an increase of 1 unit of $M(1)$ is associated with an increase of γ_M units of $Y(2)$, all other things being equal.
- Using the standardised path coefficient, an increase of 1 standard deviation (σ_M) above the mean in $M(1)$ is associated with an increase of Γ_M standard deviation (σ_Y) above the mean of $Y(2)$, other things equal.

Calculations can be done using either standardised or unstandardised coefficients.

Assuming the underlying parametric hypotheses are true (uncorrelatedness of residuals, linearity and additivity), the correlation between two variables can be computed as the sum of paths with the following characteristics:

1. the paths cannot contain loops,
2. the paths cannot include colliders (such as $U \to V \leftarrow W$),
3. the paths cannot contain more than one two-headed arrow.

In other words, the correlation between two variables is equal to the sum of all the direct effects, indirect effects, and "joint effects", where joint effects correspond to backward (backdoor) paths including exactly one double-headed arrow.

Each path contribution to the association or correlation between the two variables can be

obtained by the product of the path coefficients.

As an example from the graph in Figure (8), the Pearson correlation between $A(0)$ and $Y(2)$ can be expressed as the sum of four paths connecting $A(0)$ and $Y(2)$, each path being quantified by the standardised paths coefficients (for single arrows between $A(0)$ and $Y(2)$) or by the product of standardised paths coefficients (for paths composed of a sequence of arrows):

- Path $A \to Y$: the "direct effect" of A on Y , quantified by Γ_A
- Path $A \to M \to Y$: the "indirect effect" of A on Y, quantified by $B_A\Gamma_M$
- Path $A \leftarrow L \to M \to Y$, quantified by $\Lambda_L B_L \Gamma_M$
- Path $A \leftarrow L \to Y$, quantified by $\Lambda_L \Gamma_L$

Path analysis approach in mediation analyses is very similar to the "product of coefficients" approach described above, with the difference that unstandardised coefficients were used in the "product of coefficients" approach: $\beta_A \gamma_M$.

The sum of the first two paths correspond to the total effect of interest of $A \to Y$ (the direct effect + the indirect effect). The two other paths correspond to confounding and the correlation between A and Y, can be decomposed as:

$$\overbrace{\Gamma_A + B_A\Gamma_M}^{\text{Total effect}} + \overbrace{\Lambda_L B_L \Gamma_M + \Lambda_L \Gamma_L}^{\text{Confounding by } L}$$

Our structural assumptions and the rules of path analysis imply correlations, which can be expressed as a combination of parameters to estimate. Those parameters are estimated by matching the implied correlations with the observed correlations. In practice, maximum likelihood and variants of generalized least squares are the most common estimation methods implemented in structural equation modeling sofware [Bollen, 1989].

These approaches can be applied in the same way to causal structures including latent variables (i.e. unobserved constructs), generally known as Structural Equation Modelling (SEM). In such a representation, observed measures are related to latent constructs as in factor analyses, characterizing a measurement model. These measurement models are then combined in the path analysis framework. By convention, endogenous latent variables are usually represented by ξ letters and exogenous variables by η letters. Corresponding path diagrams include both observed (rectangle) and unobserved variables (circles).

A general limitation for the estimation of path coefficients in both path analyses and SEM is when the model is underdetermined (or underidentified). A model is underidentified if at least one of the parameter cannot be identified from the observed correlations. On the contrary, a model is identified or overidentified if each parameter is identified (or overidentified).

Underidentification can occur when there are too few indicators for one or more latent variable of the model, or in the presence of too many reciprocal paths, feedback loops or correlated residuals [Loehlin, 2004]. For models including latent variables, it is advisable to use at least 3 or 4 indicators par latent variables in order to avoid underidentification difficulties [Loehlin, 2004].

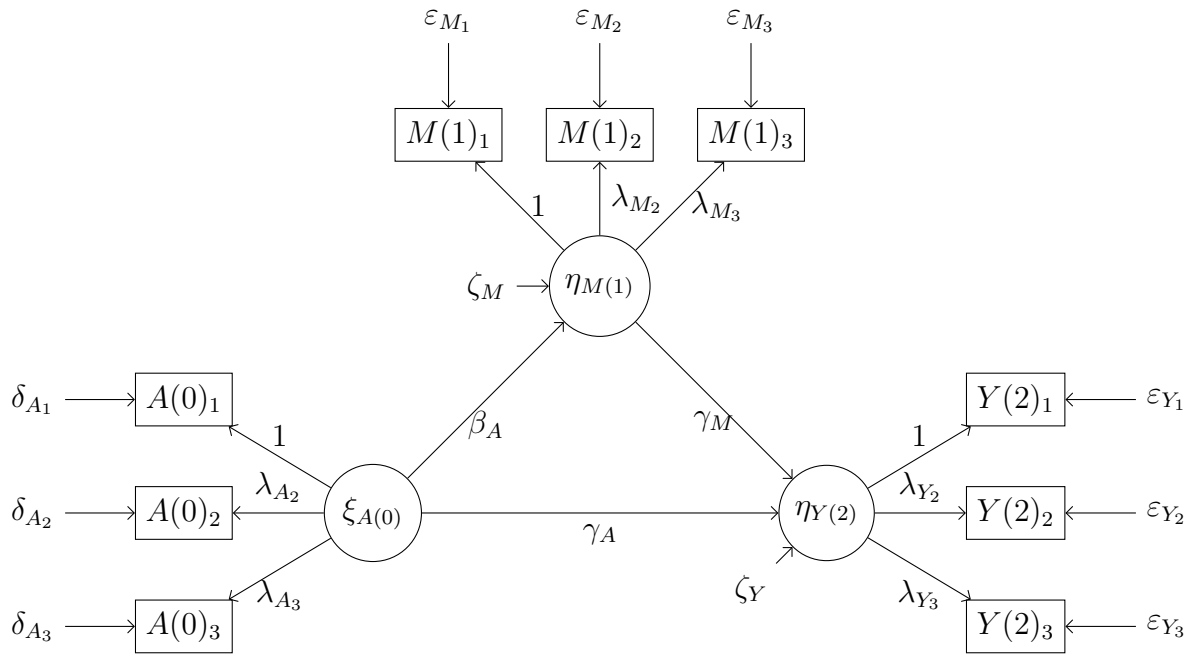
An example of a mediation model with latent variables given by MacKinnon is shown in Figure 9 [MacKinnon, 2008]. In this example, each latent variable is constructed using three observed variables.

The structural equations for the latent variables in the example are:

$$\begin{cases} \eta_{M(1)} = \beta_A \xi_{A(0)} + \zeta_M \\ \eta_{Y(2)} = \gamma_A \xi_{A(0)} + \gamma_M \eta_{M(1)} + \zeta_Y \end{cases} \tag{6}$$

and we assume a zero covariance between random residuals and latent variables (in the graph, there are no arrows connecting those variables).

Figure 9: Example of mediation model with latent variables [MacKinnon, 2008]



Using measurement models and latent variables in SEMs can be used to correct measurement errors. The corresponding models are presented as extensions of linear models and factorial analyses [Dumas et al., 2014].

2.2.3 Solutions for dealing with categorical variables or $(A * M)$ interactions

In case of binary outcomes and/or categorical mediators. Classical methods can be extended to accommodate binary mediator and/or outcome variables [Iacobucci, 2008].

For a binary outcome Y, the odds that $Y = 1$ versus $Y = 0$ given a combination of $A(0)$ and $M(1)$ can be calculated using a logistic regression:

$$\text{logit}(\mathbb{P}(Y = 1 \mid A(0), M(1), L(0))) = \gamma_Y + \gamma_A A(0) + \gamma_M M(1) + \gamma_L L(0).$$

When both the outcome $Y(2)$ and the mediator $M(1)$ are binary, the model for mediator can also be fit using a logistic regression:

$$\text{logit}(\mathbb{P}(M = 1 \mid A(0), L(0))) = \beta_M + \beta_A A(0) + \beta_L L(0)$$

However, path coefficients estimated from those two multiplicative models are difficult to integrate in an indirect effect as with coefficients from linear models, because the decomposition of a total effect in a sum of direct and indirect effect is usually defined under additive model assumptions.

[Kupek, 2005] described four main strategies to deal with categorical variables in SEM:

- asymptotic distribution-free estimators [Browne, 1984],
- robust maximum likelihood estimation [Browne and Shapiro, 1988] or using resampling techniques (jacknife or bootstrap) to obtain standard error of SEM parameters [Bollen and Stine, 1993],
- calculating biserial, polyserial, tetrachoric, or polychoric correlations between variables (assuming they have an underlying latent continuous scale) [Muthén, 1984],
- using logit or probit regression for binary variables at a first level, then proceeding with SEM as the second level [Muthén, 1993].

[Kupek, 2006] also suggested to transform binary variables, using the Yule's transformation of odds ratios into Q-metric $\left(Q = \frac{\text{OR}-1}{\text{OR}+1}\right)$ in order to approximate Pearson's correlation coefficient between binary variables. Another possibility was to use the inverse of the covariance matrix of the log-linear model parameters estimates ([Kupek, 2005]).

In case of interaction between the exposure $A(0)$ and the mediator $M(1)$. Procedures have been developed to study interactions in Structural Equation Modeling: interactive effects can be estimated between observed or latent variables [Kenny and Judd, 1984; Bollen and Paxton, 1998; Jöreskog and Yang, 1996].

In the context of mediation analyses, interaction effects have been studied in two situations

[Iacobucci, 2008]:

- Mediated moderation:
  In a mediation model involving an interaction term between two exposures A_1 and A_2, the objective is to assess if the moderation of A_1 by A_2 (or vice versa) is mediated by M, i.e. to decompose the total effect of the interaction on the outcome Y into the sum of:
  - a direct effect on the outcome, $(A_1 * A_2) \rightarrow Y$,
  - and an indirect effect, $(A_1 * A_2) \rightarrow M \rightarrow Y$
- Moderated mediation:
  The objective is to assess if the mediation relationship between an exposure A and an outcome Y through a mediator M, is moderated by another variable $L(0)$ (for example, the relationship A and Y might be mediated in the subgroup $L(0) = 0$, but not in the subgroup where $L(0) = 1$.

More recently, and in line with Pearl's causal definitions of a direct and indirect effect (see Chapter 4.2), the classical "product of coefficients" method or "difference in coefficients" method have been adapted to take into account interactions between the exposure and the mediator [Valeri and VanderWeele, 2013; Hayes, 2015; Preacher et al., 2007]. It should be noted that this work only dealt with observed variables, not with latent variables.

The outcome regression was reformulated as:

$$Y(2) = \gamma_Y + \gamma_A A(0) + \gamma_M M(1) + \gamma_{AM}(A(0) * M(1)) + \gamma_L L(0) + \epsilon_Y \tag{7}$$

Then, conditional on $L(0)$, and for a change in exposure from $A(0) = a^*$ to $A(0) = a$, we can use the parameters estimated from models (3) and (7) [Valeri and VanderWeele, 2013]:

- the *direct effect* of $A(0)$ on $Y(2)$ is:

  $$[\gamma_A + \gamma_{AM} \times (\beta_0 + \beta_A a^* + \beta'_L L(0))] (a - a^*)$$

  with $a = 1$ and $a^* = 0$, the expression is simplified to $\gamma_A + \gamma_{AM}(\beta_0 + \beta'_L L(0))$.
- the *indirect effect* of $A(0)$ on $Y(2)$ through $M(1)$ is

  $$[\beta_A \times (\gamma_M + \gamma_{AM} a)] (a - a^*)$$

  simplified to $\beta_A \times (\gamma_M + \gamma_{AM})$ with with $a = 1$ and $a^* = 0$.

2.2.4 Positioning of SEMs in mediation analyses

SEMs are of real interest as an approach that can explicitly represent hypotheses regarding the underlying structural model. Thoses hypotheses can be interpreted as a causal model according to principles similar to the use of DAGs presented in chapter 1.3 and chapter 3.

In a confirmatory analysis approach, we can use SEMs in order to [Pearl, 2009b]:

- Represent our *a priori* causal assumptions with a formal graph.
- Assess a set of statistically testable assumptions implied by the structural assumptions (missing arrows correspond to zero partial correlations). It is also possible to discuss "observationally equivalent models": models that are structurally different (holding contradictory causal assumptions), but have the same statistically testable implications.
- Assess, from *a priori* causal assumptions, if a path coefficient can be identified and estimated with causal interpretation from observational data, applying graphical criteria similar to the d-separation criterion presented in chapter 1.3.
- Have a causal interpretation of path coefficients, if the causal and statistical assumptions hold. Under those assumptions, it is possible to predict how the outcome Y would change if an external intervention was applied to change the values of other variables.
- Estimate direct and indirect effects (corresponding to direct and indirect paths), which are analogous to the Natural Direct Effet and Natural Indirect Effect defined by Pearl or to Randomized Direct and Indirect effects (detailed in chapter 4).

Moreover, compared to other methods presented in chapter 3 and chapter 5, SEMs can be used to:

- explicitly integrate latent variables, describing the conceptual structure between the latent and the observed variables;
- directly represent and integrate measurement models, that can be used to decrease measurement bias.

The conventional estimation method for SEMs, which consists in estimating a large set of parameters by iterative maximization of a fitness measure, might not be optimal in confirmatory analysis approaches:

- Intuitively, the estimation of a large number of statistical parameters associated with the model is a much more ambitious approach than the objective of correctly estimating a small number of parameters, focusing only on the scientific objective of a confirmatory analysis. More focused confirmatory approaches are described in the next chapters 4 and 5. Some estimators such as doubly robust efficient estimators aim to estimate a specific causal quantity of interest, based on a minimal set of causal and statistical assumptions.

- Some parameters of a SEM might not be identifiable when estimated by maximising a global fitness measure, but identifiable through a more local analysis focused only on the parameter of interest [Pearl, 2009b].
- As mentioned with DAGs, the absence of an arrow is a very strong assumption. Comparing alternative structural models with statistical tests or indicators is frequently done when using SEM methodology in order to get more parsimonious models. For example, nested structural equation models can be compared using χ^2 tests [Loehlin, 2004; Bollen, 1989]. However, deciding the presence or absence of direct effects between two nodes based on a statistical procedure might not be appropriate: the results usually depend on sample size and power. Null hypothesis testing is not a robust procedure to accept the null, and identifiability of the tested effects is generally overlooked (because of the large number of "no residual confounding" assumptions implied). [Bollen, 1989] considers that the respecification of an initial model corresponds rather to an exploratory analysis approach and advises to be guided first by expert knowledge before the use of empirical statistical tests and fit measures.
- In case of an intra-individual interaction between the exposure $A(0)$ and the mediator $M(1)$ affecting the outcome $Y(2)$, Kaufman *et al.* showed that the "product of coefficients" or the "difference in coefficient" methods might not be reliable for decomposing a total effect into the sum of a direct and an indirect effect [Kaufman et al., 2004].

Some authors consider that the main interest of structural equation modelling or path analysis is to explore new research hypotheses [VanderWeele, 2012]. For confirmatory objectives, other methods for mediation analysis have been developed. These alternative methods, based on counterfactual framework and DAGs, do not mix causal assumptions (non-parametric structure of the system) with statistical assumptions (about the shape of the relationships between variables), make interactions, confounding and sensitivity analyses easier to manage [VanderWeele, 2012].

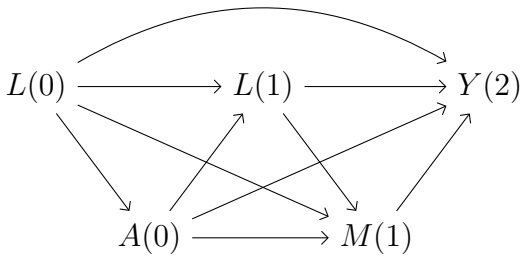
3 Non parametric Causal Models

3.1 Introduction

Mediation analyses using either Baron and Kenny approach, path analysis or structural equation models are by construction useful to explore mechanisms involved in exposure-disease relationships. However, they suffer from reported limitations described above.

One of the important limitations is their inability to address intermediate confounding factors, which is particularly important for correctly identifying indirect effects of interest. The effect of $M \rightarrow Y$ needs to be quantified to estimate indirect effects, and for this reason, the mediator-outcome confounding factors must be taken into account. These confounding factors can be baseline confounders $L(0)$ but they can also occur after the exposure of interest $A(0)$ and so also be affected by this exposure (see Figure 10), which might be common in longitudinal analyses or in lifecourse approaches. These mediator-outcome confounding factors $L(1)$ are called "intermediate confounders" or "time-varying covariates". When they are affected by the exposure $A(0)$, the use of simple regression models adjusting for mediators (like classical "difference in coefficients" or "product of coefficinents" approaches) can give biased estimation of direct effects [Robins and Greenland, 1992; Cole and Hernàn, 2002; Kaufman et al., 2004]. To overcome this issue, developments in mediation analyses have incorporated elements from the causal inference literature [Pearl, 2010a].

Figure 10: Example of DAG showing intermediate confounders $L(1)$ affected by the exposure $A(0)$



In the following paragraphs, we will present Pearl's structural causal model; the counterfactual concepts and notations that can be used to translate our scientific question; a causal roadmap with various specific definitions of direct and indirect effects; the identifiability assumptions needed to get estimands of the causal quantities of interest; and several estimation methods described in the literature.

3.2 Pearl's structural causal model

In order to clarify causal questions and propose tools, Pearl integrated three complementary elements in order to describe a structural causal model which combines "*features of the SEM [...], the potential outcome framework of Neyman and Rubin [Rubin, 1974], and the graphical models developed for probabilistic reasoning and causal analysis*" (i.e. non-parametric structural equation models associated with DAGs) [Pearl, 2010a].

3.3 Counterfactual concepts and notations

[Pearl and Mackenzie, 2018] described three levels of causation analysis:

1. *Seeing - Observation*: based on the observed association between one (past) exposure (e.g. smoking) and the occurrence of an subsequent event (e.g. myocardial infarction).
2. *Doing - Intervention*: based on the observed or the estimated consequence (on the outcome) of an intervention on the causal structure. For instance, assessing the reduction in myocardial infarction after referring patients to a tobaccologist. Such an estimation can be obtained using experimental designs, or observational designs in which the causal effect of interest is identifiable.
3. *Imagining - Speculation*: based on the prior modelling of the effect of a possible intervention on the causal structure, in order to assess the impact of this possible intervention on the disease outcome. This approach intends to address the question of what would have happened had the past been different (if we were able to change the past). For instance, "among patients who had an infarction, what would have been the risk of infarction had they all been referred to a tobaccologist before?" This approach relies on the concept of "counterfactuals" [Pearl and Mackenzie, 2018]. The use of counterfactuals even allows inferences to be made about situations that cannot be observed in the real world. Examples will be given in the definition of some direct and indirect effects.

In level 1 of causation analysis, the use of conditional probabilities allows the assessment of observational associations. For example, the probability of death by myocardial infarction ($Y(2) = 1$) knowing (i.e. seeing) that patients are smokers ($M(1) = 1$), $P(Y(2) = 1|M(1) = 1)$ can be compared to that in non-smokers $P\left(Y(2) = 1 \mid M(1) = 0\right)$ in order to assess their equivalence. Due to the directionality driving observations and predictors, conditional probabilities are of limited use in an interventional framework. Numerically, one can assess the independence between $Y(2)$ and $M(1)$ by testing if $P\left(Y(2) = 1 \mid M(1) = 1\right) = P\left(Y(2) = 1 \mid M(1) = 0\right)$, as well as testing if $P(M(1) = 1 \mid Y(2) = 1) = P\left(M(1) = 1 \mid Y(2) = 0\right)$.

In the case where $Y(2)$ and $M(1)$ are not independent, our observation of $M(1)$, changes our prediction of $Y(2)$ ($P(Y(2) = 1|M(1) = 1) \neq P(Y(2) = 1)$), as our observation of $Y(2)$,

changes our prediction of $M(1)$ ($P(M(1) = 1|Y(2) = 1) \neq P(M(1) = 1)$). For example, if a patient smokes, he/she has an excess risk of death by infarction, and if the patient died by infarction he/she was more likely to be a smoker.

In levels 2 and 3 of causation analysis which relie on an interventional framework and a structural causal model, the additional causal assumptions underlying the model will help us to better account for the directionality of causal effects in predicting the expected outcome of a possible intervention. Changing the level of $M(1)$, for example by stopping smoking ($P(M(1)) = 0$), may modify $P(Y(2) = 1)$, the risk of the health outcome; but changing $Y(2)$ (*e.g.* imagining an unethical intervention causing an infarction, $P(Y(2) = 1)$ would not modify $P(M(1) = 1)$, the probability of smoking history.

Donald Rubin and Judea Pearl suggested notations to carry-out levels 2 and 3 causation analyses.

- Rubin's *potential outcome* notations indicate values attached to counterfactual events as defined by events that did not happened but could have. The notation $Y_{A=a}$ or Y_a corresponds to the value the (potential) outcome Y would attain had the exposure A been at level a. For example, the probability of death by infarction if no one was smoking in the population is noted $P(Y(2)_{M(1)=0} = 1)$. Conversely the risk of death by infarction if everyone was smoking is noted $P(Y(2)_{M(1)=1} = 1)$. For an individual i, the causal effect of a binary variable A on Y could then be written using the contrast between the two potential outcomes, for example $Y_{A=1}(i) - Y_{A=0}(i)$.
- Pearl uses a "do()" notation in order to denote imaginary interventions:
  $P(Y = y|\text{do}(M(1) = 0))$ denotes the probability that the outcome Y would attain the value y in an imaginary scenario in which one would be able to prevent every participant from smoking.

More generally, counterfactual interventions or counterfactual scenarios can be defined using various types of scenarios, where the imaginary interventions can be:

- *Static interventions*: the whole population is exposed to a given value of the exposure variable. For example, $P(Y(2) = y|\text{do}(A(0) = a)) = P(Y(2)_{A(0)=a} = y)$ is the probability the outcome $Y(2)$ would attain the value y, had the whole population been exposed to the value $A(0) = a$.
- *Dynamic interventions*: usually used to describe dynamic regimes in which the imaginary intervention is a joint modification of several exposures of interest, where the level of exposure depends on the values of previous time-varying covariates. As an example in Figure 10, it is possible to define a joint exposure on $\{A(0), M(1)\}$ setting the values of $A(0)$ and $M(1)$ as a function of the previous time-varying covariates $d_t(L(0), \ldots, L(t))$. For instance, one could use the following rules d to decide the value to be given to $A(0)$

and $M(1)$:

$$\begin{cases} \text{set } A(0) = 1 \text{ if } L(0) > s_0, & \text{otherwise set } A(0) = 0 \\ \text{set } M(1) = 1 \text{ if } L(1)_d > s_1, & \text{otherwise set } M(1) = 0 \end{cases}$$

where s_1 and s_2 are threshold chosen by the analyst. Different dynamic regimes can be defined based on alternative rules. These rules are useful to define treatments according to monitoring variables, for example "change insulin therapy if blood glucose exceeds a given threshold".

This approach has been generalized with "modified treatment policies" defined as hypothetical interventions where the post-intervention value of treatment can depend on the actual observed treatment level and the unit's history [Díaz et al., 2021b]. For example, "increase physical activity at time t if the expected physical activity at time t is not optimal".

- *Stochastic interventions*: The hypothetical intervention corresponds to a random draw in a distribution specified by the analyst. For example, set the value of $A(0)$ as a random draw from a Bernoulli distribution of parameter π (setting $A(0) \sim \mathcal{B}(\pi)$). In mediation analyses, some direct and indirect effects can be defined based on hypothetical random draws of the mediator distribution, for example setting $M(1) \sim \Gamma_{M_a}$, where Γ_{M_a} is the expected distribution of the mediator under the counterfactual intervention setting $A(0) = a$.

3.4 Causal roadmap

Based on DAGs and new notations, the following steps have been suggested as a causal roadmap to investigate a causal question [Petersen and van der Laan, 2014]:

1. *Specify knowledge about the system to be studied using a causal model.* Background knowledge can be represented using DAGs.
2. *Specify the observed data and their link to the causal model.* This step might help to clarify if some variables are unmeasured and how these missing variables might result in bias. It is also possible to test the compatibility between the data and the causal model using packages like dagitty for R.
3. *Specify a target causal quantity.* The aim is here to translate the scientific question of interest using counterfactual notations. Regarding mediation analysis, several quantities of interest have been defined in this way and are detailed below.
4. *Assess identifiability.* Discuss the assumptions that make possible to represent the causal quantity of interest as a parameter of the observed data distribution (i.e. an estimand). The assumptions include "no residual confounding assumptions", consistency and positivity assumptions.

5. *State the statistical estimation problem and estimate.* From the estimand and the assumed statistical model, choose an estimator to approximate the causal quantity of interest. Several estimators have been described in the literature: estimation based on regression methods, g-computation, inverse-probability of treatment weighting estimators, double robust efficient methods like targeted maximum likelihood estimation.
6. *Interpret the results.* Results have to be interpreted by assessing the discrepancy between the data available for our analysis and the causal and statistical assumptions as well as the methodology employed. Sensitivity analyses can help discussing measurement error or "no residual confounding assumptions".

4 Causal quantities of interest in mediation analysis

Based on the concepts developed in the causal inference literature, several causal quantities of interest were defined to explore the role of mediation variables.

These quantities, listed in Table 1, correspond to 2-way, 3-way, or 4-way decompositions of a total effect of the exposure A on the outcome Y.

Table 1: Synthesis of causal quantities of interest in mediation analyses

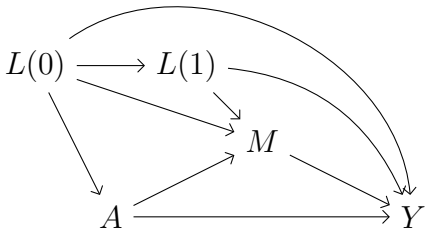
Parameters		**Definition**
Total effects		
Average Total Effect	ATE	$\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*})$
Overall Effect	OE	$\mathbb{E}\left(Y_{a,G_{a\|L(0)}}\right) - \mathbb{E}\left(Y_{a^*,G_{a^*\|L(0)}}\right)$
2-Way decomposition (1)		
Controlled Direct Effect	CDE_m	$\mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m})$
Eliminated Effect	EE	$[\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*})] - [\mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m})]$
2-Way decomposition (2)		
Pure Natural Direct Effect	PNDE	$\mathbb{E}\left(Y_{a,M_{a^*}}\right) - \mathbb{E}\left(Y_{a^*,M_{a^*}}\right)$
Total Natural Indirect Effect	TNIE	$\mathbb{E}\left(Y_{a,M_a}\right) - \mathbb{E}\left(Y_{a,M_{a^*}}\right)$
2-Way decomposition (3)		
Total Natural Direct Effect	TNDE	$\mathbb{E}\left(Y_{a,M_a}\right) - \mathbb{E}\left(Y_{a^*,M_a}\right)$
Pure Natural Indirect Effect	PNIE	$\mathbb{E}\left(Y_{a^*,M_a}\right) - \mathbb{E}\left(Y_{a^*,M_{a^*}}\right)$
2-Way decomposition (4)†		
Marginal Randomised Direct Effect	MRDE	$\mathbb{E}\left(Y_{a,G_{a^*\|L(0)}}\right) - \mathbb{E}\left(Y_{a^*,G_{a^*\|L(0)}}\right)$
Marginal Randomised Indirect Effect	MRIE	$\mathbb{E}\left(Y_{a,G_{a\|L(0)}}\right) - \mathbb{E}\left(Y_{a,G_{a^*\|L(0)}}\right)$
2-Way decomposition (5)		
Conditional Randomised Direct Effect	CRDE	$\mathbb{E}\left(Y_{a,\Gamma_{a^*\|L(0),L(1)}}\right) - \mathbb{E}\left(Y_{a^*,\Gamma_{a^*\|L(0),L(1)}}\right)$
Conditional Randomised Indirect Effect	CRIE	$\mathbb{E}\left(Y_{a,\Gamma_{a\|L(0),L(1)}}\right) - \mathbb{E}\left(Y_{a,\Gamma_{a^*\|L(0),L(1)}}\right)$
3-Way decomposition, with $a = 1$, $a^* = 0$ and binary M		
Pure Natural Direct Effect	PNDE	$\mathbb{E}\left(Y_{1,M_0}\right) - \mathbb{E}\left(Y_{0,M_0}\right)$
Mediated Interactive Effect	MIE	$\mathbb{E}\left((Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}) \times (M_1 - M_0)\right)$
Pure Natural Indirect Effect	PNIE	$\mathbb{E}\left(Y_{0,M_1}\right) - \mathbb{E}\left(Y_{0,M_0}\right)$
4-Way decomposition, with $a = 1$, $a^* = 0$ and binary M		
Controlled Direct Effect	CDE_0	$\mathbb{E}\left(Y_{1,0}\right) - \mathbb{E}\left(Y_{0,0}\right)$
Mediated Interaction Effect	MIE	$\mathbb{E}\left((Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}) \times (M_1 - M_0)\right)$
Reference Interaction Effect	RIE	$\mathbb{E}\left((Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}) \times M_0\right)$
Pure Natural Indirect Effect	PNIE	$\mathbb{E}\left(Y_{0,M_1}\right) - \mathbb{E}\left(Y_{0,M_0}\right) = \mathbb{E}\left((Y_{0,1} - Y_{0,0}) \times (M_1 - M_0)\right)$

† *the sum is equal to the Overall Effect*

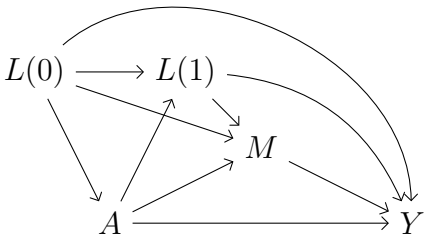
(table adapted from [VanderWeele, 2014; Wang and Arah, 2015])

In order to present these quantities in a practical way, we will present the theoretical values expected from 4 different data generating mechanisms (4 causal models, denoted $\mathcal{M}_1$, $\mathcal{M}_{1*}$, $\mathcal{M}_2$, $\mathcal{M}_{2*}$), represented in the Figure 11 with Y as the outcome, A the main exposure, M the mediator of interest, and $L(0)$ and $L(1)$ respectively baseline and intermediate confounders. Models $\mathcal{M}_1$ and $\mathcal{M}_{1*}$ are represented by the DAG (a) on the left of Figure 11, models $\mathcal{M}_2$ and $\mathcal{M}_{2*}$ are represented by the DAG (b) on the right side of the Figure. The difference between them is the presence of intermediate confounding of the mediator-outcome relationshtip, affected by the exposure A on the right side of Figure 11 for models $\mathcal{M}_2$ and $\mathcal{M}_{2*}$. On the left side, while $L(1)$ is a confounder of the $M \rightarrow Y$ relationship, we make the (strong) assumption that it is not affected by A. The statistical models $\mathcal{M}_1$ and $\mathcal{M}_{1*}$, or $\mathcal{M}_2$ and $\mathcal{M}_{2*}$ were the same except for an additional interaction term between A and M, which was added in models $\mathcal{M}_{1*}$ and $\mathcal{M}_{2*}$ to simulate the expectation $\mathbb{E}(Y|L(0), A, L(1), M)$. Because DAGs are non-parametric representation of causal models, such interaction terms are not apparent in Figure 11. Compared to the "classical" approaches described earlier, interactions between the exposure A and the mediator M can be studied in a more specific way using the causal inference approaches. The causal quantities of interest will be expressed using potential outcome notations.

Figure 11: DAGs representing data-generating mechanisms for the distribution of $\{L(0), A, L(1), M, Y\}$



(a) DAG of models $\mathcal{M}_1$ and $\mathcal{M}_{1*}$



(b) DAG of models $\mathcal{M}_2$ and $\mathcal{M}_{2*}$

The quantitative results expected under the 4 data-generating mechanisms are presented in Table 2 (for a binary outcome) and 3 (for a continuous outcome). R scripts are associated with the present document to calculate those theoretical results and to simulate corresponding data sets in order to implement and test the estimators of the causal quantities of interest in a didactic way.

These quantitative results can be used as an illustration of the quantities of interest defined below.

Table 2: "True" causal quantities expected from the simulated data generating mechanisms (binary outcome)

	Without time varying confounders								With time varying confounders			
	$\mathcal{M}_1$ without $A*M$ interaction				$\mathcal{M}_{1*}$ with $A*M$ interaction				$\mathcal{M}_2$ without $A*M$ interaction		$\mathcal{M}_{2*}$ with $A*M$ interaction	
Binary Y												
Total effect	0.058				0.06955				0.0752		0.089282	
Controlled direct effects	CDE_m	EE			CDE_m	EE			CDE_m	EE	CDE_m	EE
- setting $M=0$	0.05	0.008			0.05	0.0195			0.064	0.0112	0.064	0.025282
- setting $M=1$	0.05	0.008			0.08	-0.01045			0.064	0.0112	0.094	-0.004718
Pure NDE and Total NIE	PNDE	TNIE			PNDE	TNIE			non identifiable		non identifiable	
	0.05	0.008			0.05855	0.011			-	-	-	-
Total NDE and Pure NIE	TNDE	PNIE			TNDE	PNIE			non identifiable		non identifiable	
	0.05	0.008			0.06155	0.008			-	-	-	-
3-way decomposition	PNDE	PNIE	MIE		PNDE	PNIE	MIE		non identifiable		non identifiable	
	0.05	0.008	0		0.05855	0.008	0.003		-	-	-	-
4-way decomposition	CDE_0	PNIE	MIE	RIE	CDE_0	PNIE	MIE	RIE	non identifiable		non identifiable	
	0.05	0.008	0	0	0.05	0.008	0.003	0.00855	-	-	-	-
Randomised effects												
Marginal randomised effects	MRDE	MRIE			MRDE	MRIE			MRDE	MRIE	MRDE	MRIE
	0.05	0.008			0.05855	0.011			0.064	0.0112	0.073882	0.0154
Conditional randomised effects	CRDE	CRIE			CRDE	CRIE			CRDE	CRIE	CRDE	CRIE
	0.05	0.008			0.05855	0.011			0.0672	0.008	0.078282	0.011

CDE=controlled direct effect; EE=eliminated effect;

PNDE=pure natural direct effect; TNIE=total natural direct effect;TNDE=total natural direct effect; PNIE=pure natural indirect effect;

MIE=mediated interactive effect; RIE=reference interaction;

MRDE=marginal randomized direct effect; MRIE=marginal randomized indirect effect; CRDE=conditional randomized direct effect; CRIE=conditional randomized indirect effect

Table 3: "True" causal quantities expected from the simulated data generating mechanisms (quantitative outcome)

	Without time varying confounders								With time varying confounders			
	$\mathcal{M}_1$ without $A*M$ interaction				$\mathcal{M}_{1*}$ with $A*M$ interaction				$\mathcal{M}_2$ without $A*M$ interaction		$\mathcal{M}_{2*}$ with $A*M$ interaction	
Continuous Y												
Total effect	-4.9				-6.825				-6.26		-8.607	
Controlled direct effects	CDE_m	EE			CDE_m	EE			CDE_m	EE	CDE_m	EE
- setting $M=0$	-4	-0.9			-4	-2.825			-5	-1.26	-5	-3.607
- setting $M=1$	-4	-0.9			-9	+2.175			-5	-1.26	-10	+1.393
Pure NDE and Total NIE	PNDE	TNIE			PNDE	TNIE			non identifiable		non identifiable	
	-4	-0.9			-5.425	-1.4			-	-	-	-
Total NDE and Pure NIE	TNDE	PNIE			TNDE	PNIE			non identifiable		non identifiable	
	-4	-0.9			-5.925	-0.9			-	-	-	-
3-way decomposition	PNDE	PIE	MIE		PNDE	PIE	MIE		non identifiable		non identifiable	
	-4	-0.9	0		-5.425	-0.9	-0.5		-	-	-	-
4-way decomposition	CDE_0	PNIE	MIE	RIE	CDE_0	PNIE	MIE	RIE	non identifiable		non identifiable	
	-4	-0.9	0	0	-4	-0.9	-0.5	-1.425	-	-	-	-
Randomised effects												
Marginal randomised effects	MRDE	MRIE			MRDE	MRIE			MRDE	MRIE	MRDE	MRIE
	-4	-0.9			-5.425	-1.4			-5	-1.26	-6.647	-1.96
Conditional randomised effects	CRDE	CRIE			CRDE	CRIE			CRDE	CRIE	CRDE	CRIE
	-4	-0.9			-5.425	-1.4			-5.36	-0.9	-7.207	-1.4

CDE=controlled direct effect; EE=eliminated effect;

PNDE=pure natural direct effect; TNIE=total natural direct effect;TNDE=total natural direct effect; PNIE=pure natural indirect effect;

MIE=mediated interactive effect; RIE=reference interaction effect;

MRDE=marginal randomized direct effect; MRIE=marginal randomized indirect effect; CRDE=conditional randomized direct effect; CRIE=conditional randomized indirect effect

Note: In the following definitions, we will contrast two levels of exposure for A: $\text{do}(A = a)$ versus the reference level $\text{do}(A = a^*)$. If A is a binary exposure, we have $a = 1$ and $a^* = 0$.

4.1 Average total effect

The aim of mediation analyses is to decompose a total effect, so the first step is to define a total effect of interest. The most common total effect studied in causal analyses is the *average total effect* (ATE), defined as *the difference between the average outcome in the population had everyone been exposed to $A = a$ and the average outcome had everyone been exposed to $A = a^*$*. Using counterfactual notation, the ATE is (see table 1):

$$\text{ATE} = \mathbb{E}(Y_{A=a}) - \mathbb{E}(Y_{A=a^*}) \tag{8}$$

4.1.1 Identification assumptions

Under Identification assumptions, it is possible to express this unobserved causal quantity of interest as a parameter of the observed data. The following assumptions are necessary:

Randomisation assumption. According to the DAGs in Figures 11, applying the backdoor criterion shows that adjusting for all the baseline confounders $L(0)$ is sufficient to identify the causal effect of A on Y. In other words, conditional on $L(0)$, there is no unmeasured confounding between A and Y, denoted

$$A \perp\!\!\!\perp Y_a \mid L(0) \tag{9}$$

Positivity assumption. Also named *experimental treatment assignment*, the positivity assumption states that within each observed stratum of $L(0)$, each treatment level of interest $A = a$ and $A = a^*$ occurs with some positive probability:

$$\text{If } P\left(L(0) = l(0)\right) \neq 0, \quad \text{then} \quad \forall a, P\left(A = a \mid L(0) = l(0)\right) > 0 \tag{10}$$

We can distinguish:

- "Theoretical" positivity violation. For a treatment $A = 1$ contraindicated beyond the age of 60, the probability of observing this treatment in an 80 year old subject is supposed to be zero $P(A = 1|age = 80) = 0)$. However in this case, a scientific objective comparing treatments $A = 1$ versus $A = 0$ in participants over 60 years of age would be irrelevant.
- "Practical" positivity violation. If participant profiles are characterized by a high dimensional set of $L(0)$ variables, there is a chance for some observed profiles not to be exposed to one of the exposure levels of interest $A = a$ or $A = a^*$, because of a limited sample size.

In case of positivity violation for the level of exposure $A = a$ in some strata $l(0)$, we would have $P(A = a \mid l(0)) = 0$ and the estimation of

$$\mathbb{E}\left[Y \mid A = a, L(0) = l(0)\right] = \sum_{y \in Y} y \times P\left[Y = y \mid A = a, L(0) = l(0)\right]$$

would not be supported by the data in the $l(0)$ strata.
In practice, positivity can be assessed:

- by describing the distribution of the exposures A according to $L(0)$ and the distribution of M according to $\{L(0), A, L(1)\}$ [Westreich and Cole, 2010];
- describing the distribution of propensity scores $P\left(A = 1 \mid L(0)\right)$ and $P\left(M = 1 \mid L(0), A, L(1)\right)$;
- using some specific tools to diagnose positivity violation [Petersen et al., 2012].

Consistency assumption. This assumption states that "an individual's potential outcome under a hypothetical condition that happened to materialised is precisely the outcome experienced by that individual" [Pearl, 2010b]. Under the consistency assumption, we can write:

$$P\left[Y_a = y \mid A = a, L(0) = l(0)\right] = P\left[Y = y \mid A = a, L(0) = l(0)\right] \tag{11}$$

This statement is used to express an unobserved counterfactual concept (with Y_a on the left hand side of the equation) with a parameter of the observed data distribution (Y on the right hand side of the equation). It is linked to Rubin's "stable unit treatment value assumption" (i.e. no hidden version of the treatment: "no matter how individual i received treatment $A = a$ the potential outcome that would be observed would be $Y_{A=a}(i)$") [Rubin, 2005]. Its definition and position have been debated in the literature, mainly around the notion of a "well defined intervention" which should be discussed transparently:

- According to [Holland, 1986], "causes are only those things that could, in principle, be treatment in experiments". We should not consider a variable as a possible cause if it is not possible to imagine a practical way to change its values without directly changing the values of other variables in the system.
- According to [Hernán, 2016], "a sufficiently well-defined intervention needs to specify the start and end of the intervention and the implementation of its different components over time". Such well defined interventions usually rely on expert consensus and judgement, especially if the interventions of interest are absent in the data.
- According to [Rehkopf et al., 2016], the assumption "requires that there are no two versions of treatment such that $A = 1$ under both versions but the outcome Y would be different under the alternative versions". For intervention that are difficult to define, such as intervening on social determinants, epidemiologist should be explicit about their version of exposure and identify potentially relevant variations, examine reduced range of such

versions, and test multiple definitions of an exposure.

- According to [Pearl, 2010b], it should be possible to encode the different ways of performing an action on A as well as its side effects on Y in the causal model. From such a complete causal model, consistency is more a theorem that can be derived in the logic of counterfactuals than an assumption used to preclude any side effects of the exposure on the outcome.
- According to [Schwartz et al., 2016], restriction to well-defined interventions may inhibit the possibility of conceptualising significant structural changes and radical interventions. They encourage to consider system thinking with multiple outcomes, including unintended and long term consequences. They claim that causal work outside of the policy space, with less "well-defined interventions", is still usefull to study mechanisms and understand the world.

4.1.2 Estimand

Under the identification assumptions, the ATE can be expressed using parameters of the observed data distribution (cf. Appendix 7.1 and table 5):

$$\Psi^{\text{ATE}} = \sum_{l(0)} [\mathbb{E}(Y \mid A = a, l(0)) - \mathbb{E}(Y \mid A = a^*, l(0))] \times P(L(0) = l(0)) \tag{12}$$

4.2 Two-way decompositions of the total effect

Several approaches have been described in the literature to decompose a total effect into two components, using counterfactual notations. Some of these quantities (Controlled Direct Effects, Marginal or Conditional Randomised Direct and Indirect effects) can be identified in both causal structures shown in Figures 11(a) and 11(b), but other quantities (Natural Direct and Indirect effects) can be identified only if confounders $L(1)$ of the $M \rightarrow Y$ relationship are not affected by the exposure A (as in Figure 11(a)).

4.2.1 Controlled direct effect and eliminated effect

The *controlled direct effect* is defined as the effect of a joint hypothetical intervention that would change the exposure A from a reference value $A = a^*$ to the value $A = a$, while keeping the mediator constant to a given value $M = m$ [Robins and Greenland, 1992; Pearl, 2001]:

$$\text{CDE}_m = \mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m}) \tag{13}$$

Controlled direct effects can be identified under the following sequential randomisation assumptions and the positivity assumptions indicated in Table 4 [Pearl, 2001]:

i . No unmeasured confounding between A and Y, given $L(0)$

ii . No unmeasured confounding between M and Y, given $L(0)$, A and $L(1)$

Under the identification assumptions, CDE_m can be expressed using parameters of the observed data distribution, as indicated in Table 4.

A controlled direct effect correspond to all the effect of A on Y that is not mediated by M, when setting the mediator to a specific value $M = m$. By analogy with path analyses, the CDE corresponds to the direct path $A \to Y$ in Figure 11(a) and to both paths $A \to Y$ and $A \to L(1) \to Y$ in Figure 11(b).

Moreover, CDE can unambiguously deal with $A * M$ interaction effects: in case of $A * M$ interaction, CDE_m will vary according to the value chosen for $M = m$. For example, in results from simulated data sets in Tables 2 and 3, CDE_m does not depend on m in $\mathcal{M}_1$ and $\mathcal{M}_2$ statistical models (which do not include $A * M$ interactions), but $\text{CDE}_0 \neq \text{CDE}_1$ in statistical models $\mathcal{M}_{1*}$ and $\mathcal{M}_{2*}$.

Controlled direct effects are usually described as measures of interest in policy setting. In our simulated example, we could imagine a very efficient public health policy that would ban smoking, setting the smoking variable to $M = 0$ for every subject in the population. The controlled direct effect would be the remaining effect of ACE on mortality or quality of life (contrasting a hypothetical population completely exposed to ACE *versus* a population completely free of ACE), had we been able to completely ban smoking.

Because the causal quantity CDE_m is not directly compared to the average total effect, VanderWeele suggested to use the "eliminated effect" (EE) and "proportion eliminated" in order to express the controlled direct effect as a percentage of the average total effect [VanderWeele, 2013a]:

$$\text{EE} = \text{ATE} - \text{CDE}_m = [\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*})] - [\mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m})] \tag{14}$$

where the proportion of the average total effect eliminated by a hypothetical intervention setting $M = m$ in the whole population is :

$$\frac{\text{ATE} - \text{CDE}_m}{\text{ATE}} = \frac{[\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*})] - [\mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m})]}{\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*})}$$

In our simulated example (Table 2), under the statistical model $\mathcal{M}_{1*}$ and $\mathcal{M}_{2*}$, a hypothetical intervention banning smoking in the whole population ($M = 0$) would eliminate respectively $\frac{0.06955-0.05}{0.06955} = 28.1\%$ and $\frac{0.089282-0.064}{0.089282} = 28.3\%$ of the average total effect of ACE on mortality.

Of course, setting a hypothetical intervention requiring every participant to smoke ($M = 1$) would not make sense as a public health intervention; this would correspond to an increase in the Average total effect of ACE on mortality of 15% and 5.3% respectively.

4.2.2 *Natural Direct Effect* and a *Natural Indirect Effect*

Pure Natural Direct and Total Natural Indirect Effect (PNDE and TNIE). [Pearl, 2001] defined the *pure natural direct effects* and *total natural indirect effects*.

The *pure natural direct effect* (PNDE) is the effect on Y that would be realised under the hypothetical intervention of changing the value of A from a^* to a, while the mediator was kept constant at the individual counterfactual values $M_{A=a^*}$ that would be observed under the hypothetical intervention setting $A = a^*$:

$$\text{PNDE} = \mathbb{E}\left(Y_{a,M_{a^*}}\right) - \mathbb{E}\left(Y_{a^*,M_{a^*}}\right)$$

Based on the following composition assumption: $Y_a = Y_{a,M_a}$ (i.e. the potential outcome Y expected under the hypothetical intervention setting $A = a$ is equal to the potential outcome expected under the joint hypothetical intervention setting $A = a$ and M to the counterfactual value M_a expected had the exposure been $A = a$), it is possible to define the *total natural indirect effect* (TNIE) as the difference between the average total effect (ATE) and the pure natural direct effect:

$$\begin{aligned}\text{TNIE} = \text{ATE} - \text{PNDE} &= \left[\mathbb{E}\left(Y_{a,M_a}\right) - \mathbb{E}\left(Y_{a^*,M_{a^*}}\right)\right] - \left[\mathbb{E}\left(Y_{a,M_{a^*}}\right) - \mathbb{E}\left(Y_{a^*,M_{a^*}}\right)\right] \\ &= \mathbb{E}\left(Y_{a,M_a}\right) - \mathbb{E}\left(Y_{a,M_{a^*}}\right)\end{aligned}$$

The total natural indirect effect (TNIE) is interpreted as the effect on Y that would be realised under the hypothetical intervention of changing the value of the mediator M from the counterfactual value M_{a^*} (expected had A been set to $A = a^*$) to the individual counterfactual value M_a (had A been set to $A = a$), while the exposure to A was kept constant at $A = a$.

These two definitions allow to decompose the ATE into the sum of a direct and an indirect effect:

$$\text{ATE} = \text{PNDE} + \text{TNIE}$$

Natural direct and indirect effects might be informative regarding mediation mechanisms, but they are defined using unobservable cross-world counterfactuals which could make their intuitive interpretation difficult. PNDE can be interpreted as the effect of the exposure on the outcome, if the exposure had no effect on the mediator, i.e. when the mediator takes its value when the exposure is absent [Zheng and van der Laan, 2012]. In our example, the natural direct effect is the effect of changing the level of exposure to ACE in the population from "completely free" to "completely exposed", while setting smoking constant to the values which would naturally be observed under no exposure to ACE. The Total Natural Indirect Effect is the effect that would be observed had smoking changed from the values that would be observed under "no exposure to ACE" in the whole population to the values that would be observed had the whole

population been exposed to ACE, while setting a constant exposure to ACE in the whole population. A whole population exposed at the same time to ACE and to the smoking level that would be observed had they not been exposed to ACE cannot be observed in the real world. The corresponding potential outcome on mortality $\mathbb{E}(Y_{a,M_{a^*}})$ is an unobservable cross-world counterfactual. It is a good example of the third level of causation analysis (imagining-speculation) described in paragraph 3.3 [Pearl and Mackenzie, 2018].

Natural direct and indirect effects can be identified under the following assumptions (see table 4):

i . No unmeasured confounding between A and Y, given $L(0)$;
ii . No unmeasured confounding between M and Y, given $L(0)$, A and $L(1)$;
iii . No unmeasured confounding between A and M, given $L(0)$;
iv . No confounder $L(1)$ of the $M - Y$ relationship is affected by A.

Note that the assumption (iv) does not hold in the causal structure depicted in Figure 11(b). This is a significant limitation for this definition of natural direct and indirect effects.

Positivity assumptions necessary to identify natural direct and indirect effects are detailed in table 4 [Pearl, 2001; VanderWeele and Vansteelandt, 2009].

Under the identification assumptions, PNDE and TNIE can be expressed using parameters of the observed data distribution (see Table 5).

Total Natural Direct and Pure Natural Indirect Effect (TNDE and PNIE). Alternatively, it is possible to define a *Total Natural Direct Effect* (TNDE) where the values of the mediator which are kept constant are the counterfactual values $M_{A=a}$ had A been set to $A = a$ (instead of $A = a^*$ as in the definition of PNDE described above). On a difference scale:

$$\text{TNDE} = \mathbb{E}\left(Y_{a,M_a}\right) - \mathbb{E}\left(Y_{a^*,M_a}\right)$$

A *Pure Natural Indirect effect* (PNIE) can then be defined as:

$$\text{PNIE} = \text{ATE} - \text{TNDE} = \mathbb{E}\left(Y_{a^*,M_a}\right) - \mathbb{E}\left(Y_{a^*,M_{a^*}}\right)$$

Identification assumptions of TNDE and PNIE are the same as those described above for PNDE and TNIE. Under the identification assumptions, TNDE and PNIE can be expressed using parameters of the observed data distribution (see Table 5).

The difference between TNDE/PNIE and PNDE/TNIE definitions of direct and indirect effects is that in the presence of an $A * M$ interaction affecting Y, the interaction effect appears in the "total" component of the direct or indirect effect (while the "pure" component is free of this interaction effect).

4.2.3 *Interventional (or randomised) Natural Direct and Indirect Effect*

The natural direct (PNDE or TNDE) and indirect effects (TNIE or PNIE) described above have some limits: they are defined using unobservable cross-world concepts and they are not identifiable in causal structures with intermediate confounder of the $M-Y$ relationship affected by the exposure A (as in Figure 11(b)).

Using the notion of stochastic counterfactual interventions, two other types of natural direct and indirect effects have been defined, allowing to overcome these limits [Didelez et al., 2006; Nguyen et al., 2020]:

- Marginal interventional (or randomised) natural direct and indirect effects [Vanderweele et al., 2014; VanderWeele, 2017; Vansteelandt and Daniel, 2017; Rudolph et al., 2017, 2018, 2019; Díaz et al., 2021a],
- Conditional interventional (or randomised) natural direct and indirect effects [Zheng and van der Laan, 2012, 2017].

With interventional (or randomised) direct and indirect effects, the hypothetical intervention on the mediator implies a random draw in the counterfactual distribution of M had the exposure been set to a given level, instead of setting the value of M to the individual values had the exposure been set to a given level.

The approach is similar to the description of natural direct effects as a weighted average of the controlled direct effects, given by Petersen [Petersen et al., 2006]:

$$\text{PNDE} = \mathbb{E}_{L(0),L(1)} \sum_m \left[\mathbb{E}(Y_{a,m} \mid L(0), L(1)) - \mathbb{E}(Y_{a^*,m} \mid L(0), L(1))\right] P(M_{a^*} = m \mid L(0), L(1))$$

Interventional (randomised) direct and indirect effects are identifiable in causal structures with intermediate confounders of the $M - Y$ relationship affected by the exposure A (as in Figure 11(b)). Moreover, because they involve distributions rather than unknown individual values, they are considered more policy relevant [Vansteelandt and Daniel, 2017].

Marginal interventional natural direct and indirect effects. [Vanderweele et al., 2014] defined the *marginal randomised (or interventional) natural direct effect* and *marginal randomised (or interventional) natural indirect effect*.

The *marginal randomised natural direct effect* (MRDE) is the effect on Y that would be observed under the hypothetical intervention of changing the value of A from a^* to a, while the mediator is set to a random draw for each subject from the distribution of M_{a^*} (the counterfactual distribution of M had the exposure been set to $A = a^*$), conditional on $L(0)$. Such counterfactual distribution of the mediator is denoted $G_{a^*|L(0)}$:

$$\text{MRDE} = \mathbb{E}\left(Y_{a,G_{a^*|L(0)}}\right) - \mathbb{E}\left(Y_{a^*,G_{a^*|L(0)}}\right)$$

The *marginal randomised natural indirect effect* (MRIE) is the effect on Y that would be observed under the hypothetical intervention of setting the value of the exposure to $A = a$, while shifting the values of M:

- from a random draw for each subject from the counterfactual distribution of the mediator (conditional on $L(0)$) had the exposure been set to $A = a^*$ ($M \sim G_{a^*|L(0)}$),
- to a random draw from the counterfactual distribution of the mediator had the exposure been set to $A = a$ ($M \sim G_{a|L(0)}$).

$$\text{MRIE} = \mathbb{E}\left(Y_{a,G_{a|L(0)}}\right) - \mathbb{E}\left(Y_{a,G_{a^*|L(0)}}\right)$$

The sum of the marginal randomised natural direct and indirect effects gives an *Overall Effect*:

$$\text{OE} = \mathbb{E}\left(Y_{a,G_{a|L(0)}}\right) - \mathbb{E}\left(Y_{a^*,G_{a^*|L(0)}}\right)$$

The Overall Effect can be interpreted as a total effect, however because it is defined using random draws from counterfactual distributions of the mediator (conditional on $L(0)$) rather than individual counterfactual values, the Overall Effect may differ from the Average Total Effect defined in equation (8), especially in case of non-linear models and $L(0) * L(1)$ interaction effects affecting the mediator M [Vansteelandt and Daniel, 2017].

Another limit described for MRDE and MRIE is that the counterfactual variables $Y_{a,G_{a'|L(0)}}$ might not be well-defined in survival settings where time-to-event outcomes can occur before the mediator: a participant still alive under $A = a$ would be allowed to draw the mediator value of a participant who has died under $A = a'$ [Zheng and van der Laan, 2017].

Marginal randomised natural direct and indirect effects can be identified under the following assumptions (see table 4):

i . No unmeasured confounding between A and Y, given $L(0)$
ii . No unmeasured confounding between M and Y, given $L(0)$, A and $L(1)$
iii . No unmeasured confounding between A and M, given $L(0)$

Importantly, those assumption hold in both Figure 11(a) and (b). Positivity assumptions necessary to identify MRDE and MRIE are detailed in table 4.

Under the identification assumptions, MRDE and MRIE can be expressed using parameters of the observed data distribution (see Table 6) [Vanderweele et al., 2014; VanderWeele, 2017].

In causal structure such as described in Figure 11(a), the interpretation of MRDE and MRIE is analogous to the interpretation of the Pure Natural Direct Effect (PNDE) and the Total Natural Indirect Effect (TNIE), respectively. Note that the definitions of MRDE and MRIE can easily be adapted to get analogues of TNDE and PNIE.

In case of intermediate confounder $L(1)$ of the $M-Y$ relationship affected by the exposure, as in Figure 11(b), interpretation by analogy with path analyses is the following:

- the Marginal Randomised Indirect Effect (MRIE) corresponds to all the directed paths from A to Y going through the mediator M:
  $A \to M \to Y$ and $A \to L(1) \to M \to Y$.
- the Marginal Randomised Direct Effect (MRDE) corresponds to all the directed paths from A to Y which do not go through the mediator M:
  $A \to Y$ and $A \to L(1) \to Y$.

Conditional interventional natural direct and indirect effects. [Zheng and van der Laan, 2012, 2017] defined the *conditional randomised (or interventional) natural direct effect* and *conditional randomised (or interventional) natural indirect effect*.

The *conditional randomised natural direct effect* (CRDE) is the effect on Y that would be observed under the hypothetical intervention of changing the value of A from a^* to a, while the mediator is set to a random draw for each subject from the distribution of M_{a^*} (the counterfactual distribution of M had the exposure been set to $A = a^*$), fully conditional on the past (conditional on $L(0)$ and $L(1)$). Such counterfactual distribution of the mediator is denoted $\Gamma_{a^*|L(0),L(1)}$:

$$\text{MRDE} = \mathbb{E}\left(Y_{a,\Gamma_{a^*|L(0),L(1)}}\right) - \mathbb{E}\left(Y_{a^*,\Gamma_{a^*|L(0),L(1)}}\right)$$

The *conditional randomised natural indirect effect* (CRIE) is the effect on Y that would be observed under the hypothetical intervention of setting the value of the exposure to $A = a$, while shifting the values of M:

- from a random draw for each subject from the counterfactual distribution of the mediator (conditional on $L(0)$ and $L(1)$) had the exposure been set to $A = a^*$ ($M \sim \Gamma_{a^*|L(0),L(1)}$),
- to a random draw from the counterfactual distribution of M had the exposure been set to $A = a$ ($M \sim \Gamma_{a|L(0),L(1)}$).

$$\text{MRIE} = \mathbb{E}\left(Y_{a,\Gamma_{a|L(0),L(1)}}\right) - \mathbb{E}\left(Y_{a,\Gamma_{a^*|L(0),L(1)}}\right)$$

Unlike marginal randomised direct and indirect effects, the sum of the conditional randomised natural direct and indirect effects is equal to the usual Average Total Effect (ATE) defined in equation (8). Moreover, the quantity is well-defined in survival settings [Zheng and van der Laan, 2017].

Identifiability assumptions for Conditional randomised natural (in)direct effects are stronger than for Marginal randomised natural (in)direct effects, with an additional assumption described below (see table 4):

i . No unmeasured confounding between A and Y, given $L(0)$
ii . No unmeasured confounding between M and Y, given $L(0)$, A and $L(1)$
iii . No unmeasured confounding between A and M, given $L(0)$
v . No unmeasured confounding between A and $L(1)$, given $L(0)$

Those assumptions hold in both Figure 11(a) and (b). Positivity assumptions necessary to identify CRDE and CRIE are detailed in table 4.

Under the identification assumptions, CRDE and CRIE can be expressed using parameters of the observed data distribution (see Table 6) [Zheng and van der Laan, 2017].

Like MRDE and MRIE, in causal structure such as described in Figure 11(a), the interpretation of CRDE and CRIE is analogous to the interpretation of the Pure Natural Direct Effect (PNDE) and the Total Natural Indirect Effect (TNIE), respectively. Note that the definitions of CRDE and CRIE can easily be adapted to get analogues of TNDE and PNIE.

In case of intermediate confounder $L(1)$ of the $M-Y$ relationship affected by the exposure, as in Figure 11(b), interpretation by analogy with path analyses is the following:

- the Conditional Randomised Direct Effect (CRIE) corresponds to the directed path from A to Y which goes only through the mediator M:
  $A \to M \to Y$
- the Conditional Randomised Indirect Effect (CRDE) corresponds to all the directed paths from A to Y, except the path going only through M:
  $A \to Y$,
  $A \to L(1) \to Y$,
  and $A \to L(1) \to M \to Y$

Because the Conditional Randomised Direct Effect (CRDE) includes one of the paths which goes through the mediator ($A \to L(1) \to M \to Y$), its interpretation might be less intuitive.

Links between Natural (in)direct Effects, Controlled Direct Effects and Randomised (in)direct Effects, illustrated using simulated data. Using the results from simulated data sets presented in Tables 2 and 3, we can see that:

- Without intermediate confounder $L(1)$ of the $M-Y$ association, affected by the exposure A (left side of the Tables and Figure 11(a)):
  - The PNDE (respectively TNIE) is the same as the CRDE (resp. CRIE). Note that in this example, the Average total effect (ATE) is equal to the Overall effect (OE), so that the PNDE (resp. TNIE) is also the same as the MRDE (resp. MRIE), and

the MRDE can be interpreted as a weighted mean of the controlled direct effects CDE_0 and CDE_1.
  - Moreover, without any $A*M$ interaction affecting the outcome Y (statistical model $\mathcal{M}_1$), the CDE_1 and CDE_0 have the same value as the PNDE.
- In case of intermediate confounder $L(1)$ of the $M-Y$ association, affected by the exposure A (right side of the Tables and Figure 11(b)):
  - Natural (in)direct Effects are not identifiable.
  - The MRDE can be interpreted as a weighted mean of the controlled direct effects CDE_0 and CDE_1. Without any $A*M$ interaction affecting the outcome Y (statistical model $\mathcal{M}_2$), the MRDE is the same as the CDE.
  - The MRDE (resp. MRIE) does not include the same set of paths as the CRDE (resp. CRIE), so that MRDE $\neq$ CRDE (resp. MRIE $\neq$ CRIE).

4.3 Three-way and four-way decomposition of the total effect

In the presence of an interaction $A*M$ between the exposure and the mediator affecting the outcome Y, the effect of changing the exposure from $A = a^*$ to $A = a$ will depend on the value $M = m$ of the mediator. Consequently, the value of the CDE_m will depend on the value fixed for the mediator $M = m$, and the PNDE (respectively TNIE) will be different from the TNDE (resp. PNIE). [VanderWeele, 2013b, 2014] defined several causal quantities to separate interaction effects from the direct and indirect effects, applying a 3-way or a 4-way decomposition.

4.3.1 Three-way decomposition

VanderWeele suggested a decomposition of the Average total effect into [VanderWeele, 2013b]:

1. a *Pure Direct Effect*, equivalent to the PNDE described above. For simplicity, the causal quantities are described for hypothetical interventions on a binary exposure A where $a = 1$ and $a^* = 0$ and binary mediators M.

$$\text{PNDE} = \mathbb{E}(Y_{1,M_0}) - \mathbb{E}(Y_{0,M_0})$$

2. a *Pure Indirect Effect*, equivalent to the PNIE.

$$\text{PNIE} = \mathbb{E}(Y_{0,M_1}) - \mathbb{E}(Y_{0,M_0})$$

3. and a *Mediated Interactive Effect* (MIE):

$$\text{MIE} = \mathbb{E}\left((Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}) \times (M_1 - M_0)\right)$$

The MIE is the average of the product between:

- an additive interaction effect between the exposure and the mediator on the outcome, corresponding to the difference between the effect on $Y_{a,m}$ of a hypothetical joint modification of A and M from $(A = 0, M = 0)$ to $(A = 1, M = 1)$, contrasted with the sum of two individual changes in either A or M, while the other variable is set constant to the reference level $M = 0$ or $A = 0$:

$$[Y_{1,1} - Y_{0,0}] - [(Y_{1,0} - Y_{0,0}) + (Y_{0,1} - Y_{0,0})] = (Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0})$$

- and the effect of the exposure A on the mediator M (denoting M_a the counterfactual value of M had the exposure been set to $A = a$): $(M_1 - M_0)$

The MIE is equal to the difference between the TNDE and the PNDE, as well as the difference between the TNIE and the PNIE. The MIE corresponds to an additive interaction which operates only if the exposure A has an effect on the mediator.

Identification assumptions described for this 3-way decomposition are similar to the identification assumptions of PNDE and TNDE described above (see Table 4). This 3-way decomposition is not identifiable in causal structures with intermediate confounders $L(1)$ of the $M - Y$ relationship affected by A, such as in Figure 11(b).

Under the identification assumptions, the PNDE, PNIE and MIE can be expressed using parameters of the observed data distribution (see Table 7) [VanderWeele, 2013b].

4.3.2 Four-way decomposition

[VanderWeele, 2014] also developped a 4-way decomposition of the Average total effect (ATE) of an exposure A on an outcome Y, into:

1. a "Controlled Direct Effect" (CDE_0) of A on Y, setting the level of the mediator to the reference value $M = 0$:

$$\text{CDE}_0 = \mathbb{E}(Y_{1,0}) - \mathbb{E}(Y_{0,0})$$

   The CDE_0 corresponds to the standard CDE with the value of M fixed to 0, i.e. the effect of the exposure in the absence of the mediator (effect due neither to mediation nor to interaction);
2. a "Reference Interaction Effect" (RIE)

$$\text{RIE} = \mathbb{E}\left[(Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}) \times M_0\right]$$

   The RIE corresponds to the $A * M$ additive interaction effect on the outcome Y which

operates only if the mediator is present in the absence of the exposure (when $M_0 = 1$). This effect is due to the $(A * M)$ interaction only.

3. a "mediated interaction" (MIE), similar to the MIE of the 3-way decomposition:

$$\text{MIE} = \mathbb{E}\left[(Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0}) \times (M_1 - M_0)\right]$$

The MIE corresponds to the effect of A on Y due to both the mediation through the mediator and the interaction with the mediator.

4. and a pure indirect effect equivalent to the PNIE:

$$\text{PNIE} = \mathbb{E}(Y_{0,M_1}) - \mathbb{E}(Y_{0,M_0})$$

The PNIE corresponds to the effect of A on Y due to mediation only.

According to [VanderWeele, 2014], "if the exposure affects the outcome for a particular individual, then at least 1 of 4 things must be the case".

In relation with the 2-way or the 3-way decomposition, we have:

- $\text{PNDE} = \text{CDE}_0 + \text{RIE}$,
- $\text{TNDE} = \text{CDE}_0 + \text{RIE} + \text{MIE}$.

Identification assumptions for the 4-way decomposition are similar to the identification assumptions of the 3-way decomposition (see Table 4). Like the 3-way decomposition, the 4-way decomposition is not identifiable in causal structures described in Figure 11(b).

Under the identification assumptions, the CDE_0, RIE, MIE, and PNIE can be expressed using parameters of the observed data distribution (see Table 7) [VanderWeele, 2014].

Using the results from simulated data sets presented in Tables 2 and 3, we can see that:

- Without time varying confounder $L(1)$ affected by the exposure (left side of the Tables and Figure 11(a)):
  - without $A * M$ interaction, the MIE and RIE are null
    and $\text{PNDE} = \text{CDE}_0 = \text{CDE}_m = \text{MRDE} = \text{CRDE}$;
  - in presence of $A * M$ interaction,
    $\text{PNDE} = \text{CDE}_0 + \text{RIE} = \text{MRDE} = \text{CRDE}$
    and $\text{TNDE} = \text{CDE}_0 + \text{RIE} + \text{MIE}$
    and $\text{TNIE} = \text{PNIE} + \text{MIE} = \text{MRIE} = \text{CRIE}$
- We note that the 3-way or 4-way decompositions are unidentifiable in case of intermediate confounder $L(1)$ of the $M - Y$ association, affected by the exposure A (right side of the Tables and Figure 11(b)).

5 Estimators

Several estimators have been described for the causal quantities of interest in mediation analysis. In this section, we present a summary of these estimators. For didactic purposes, applied examples using R software and packages are presented as online supplementary material. In the online examples, data sets are created from the 4 types of data generating mechanisms described in section 4 (statistical models $\mathcal{M}_1$, $\mathcal{M}_{1*}$, $\mathcal{M}_2$ and $\mathcal{M}_{2*}$).

- Estimators based on traditional regression models;
- G-computation approaches;
- Marginal Structural Models (MSM), usually estimated by Inverse probability of treatment weighting (IPTW);
- Estimators based on the efficient influence function, mainly Targeted Maximum Likelihood Estimation (TMLE).

5.1 Traditional regression models

Using traditional regression models have been described for two-way decomposition (controlled direct effects, natural direct and indirect effects), three-way decomposition and four-way decomposition. The approach is similar to the "product method" or the "difference method" previously described for classical methods in section 2.1.3.

Approaches based on traditional regression models can be applied in the absence of an intermediate confounder $L(1)$ of the $M - Y$ relationship affected by the exposure A (as in Figure 11(a)).

In causal structures with intermediate confounders of the $M - Y$ relationship affected by the exposure (like in Figure 11(a)), NDE and NIE are not identifiable. The use of traditional regression models adjusted for the mediator results in biased estimations of direct (or indirect) effects due to a collider stratification bias [Robins and Greenland, 1992; Cole and Hernàn, 2002]. This bias can be large in case of strong effects of $A(0)$ on M combined to strong effects of $L(1)$ on M [Lepage et al., 2016].

In order to estimate Controlled direct effects, Marginal or Conditional (in)direct effects, which are identifiable in causal structures such as the DAG in Figure 11(b), other estimators are necessary: G-computation, IPTW or TMLE, described in the next sections.

In structures such as the one described in Figure 11(a), when the mediator M and the outcome Y are quantitative variables, VanderWeele *et al.* suggested to apply the following traditional linear regressions of the mediator M and the outcome Y [VanderWeele and Vansteelandt, 2009; Valeri and VanderWeele, 2013]. If necessary, these models can accommodate $(A * M)$ interactions affecting the outcome Y in order to correctly estimate the Controlled Direct Effect

or the Natural Direct and Indirect Effects. Note that estimations of the causal quantities are conditional on the set of confounders $L(0)$ and $L(1)$.

The first stage is to estimate the following models of the mediator and the outcome (as in Baron and Kenny approach):

$$\mathbb{E}(M \mid A, L(0), L(1)) = \beta_0 + \beta_A A + \beta_{L(0)} L(0) + \beta_{L(1)} L(1) \tag{15}$$

$$\mathbb{E}(Y \mid A, M, L(0), L(1)) = \gamma_0 + \gamma_A A + \gamma_M M + +\gamma_{AM}(A * M) + \gamma_{L(0)} L(0) + \gamma_{L(1)} L(1) \tag{16}$$

Assuming the models are correctly specified, the Controlled Direct Effect, Natural Direct and Indirect Effects can be estimated from equations (15) and (16) by:

$$\widehat{\Psi}^{\mathrm{CDE}_m}_{\mathrm{trad}} \mid L(0), L(1) = (\widehat{\gamma}_A + \widehat{\gamma}_{AM} \times m) \times (a - a^*)$$

$$\widehat{\Psi}^{\mathrm{PNDE}}_{\mathrm{trad}} \mid L(0), L(1) = \left[\widehat{\gamma}_A + \widehat{\gamma}_{AM}(\widehat{\beta}_0 + \widehat{\beta}_A a^* + \widehat{\beta}'_{L(0)} L(0) + \widehat{\beta}'_{L(1)} L(1))\right] \times (a - a^*)$$
$$\widehat{\Psi}^{\mathrm{TNIE}}_{\mathrm{trad}} \mid L(0), L(1) = \widehat{\beta}_A \times [\widehat{\gamma}_M + \widehat{\gamma}_{AM} a] \times (a - a^*)$$

$$\widehat{\Psi}^{\mathrm{TNDE}}_{\mathrm{trad}} \mid L(0), L(1) = \left[\widehat{\gamma}_A + \widehat{\gamma}_{AM}(\widehat{\beta}_0 + \widehat{\beta}_A a + \widehat{\beta}'_{L(0)} L(0) + \widehat{\beta}'_{L(1)} L(1))\right] \times (a - a^*)$$
$$\widehat{\Psi}^{\mathrm{PNIE}}_{\mathrm{trad}} \mid L(0), L(1) = \widehat{\beta}_A \times [\widehat{\gamma}_M + \widehat{\gamma}_{AM} a^*] \times (a - a^*)$$

In the absence of $A * M$ interaction, these estimators correspond to the Baron and Kenny ("product of coefficient") approach. For example, with a binary exposure A, where $a = 1$ and $a^* = 0$:

- the Natural Direct Effect (which is equivalent to the Controlled Direct Effect) is estimated by $\widehat{\gamma}_A$;
- the Natural Indirect Effect is estimated by the product $\widehat{\beta}_A \times \widehat{\gamma}_M$.

If the mediator or the outcome are binary variables, Valeri and VanderWeele described alternative approches using logistic or log-linear traditional regression models in order to estimate Controlled Direct Effects, Natural Direct and Indirect Effects. Note that they are expressed on Odds Ratio or Risk Ratio scales for binary outcomes [Valeri and VanderWeele, 2013].

Three-way decomposition. If the mediator M and outcome Y are quantitative variables and the equations (15) and (16) are correctly specified, the 3-way decomposition can be estimated

by [VanderWeele, 2013b]:

$$\begin{aligned}
\widehat{\Psi}^{\text{PNDE}}_{\text{trad}} \mid L(0), L(1) &= \left[\widehat{\gamma}_A + \widehat{\gamma}_{AM}(\widehat{\beta}_0 + \widehat{\beta}_A a^* + \widehat{\beta}'_{L(0)} L(0) + \widehat{\beta}'_{L(1)} L(1))\right] \times (a - a^*) \\
\widehat{\Psi}^{\text{PNIE}}_{\text{trad}} \mid L(0), L(1) &= \widehat{\beta}_A \times [\widehat{\gamma}_M + \widehat{\gamma}_{AM} a^*] \times (a - a^*) \\
\widehat{\Psi}^{\text{MIE}}_{\text{trad}} \mid L(0), L(1) &= \widehat{\beta}_A \times \widehat{\gamma}_{AM} \times (a - a^*) \times (a - a^*)
\end{aligned}$$

Four-way decomposition. Similarly, if the mediator M and outcome Y are quantitative variables and the equations (15) and (16) are correctly specified, the 4-way decomposition can be estimated by [VanderWeele, 2014]:

$$\begin{aligned}
\widehat{\Psi}^{\text{CDE}_0}_{\text{trad}} \mid L(0), L(1) &= \widehat{\gamma}_A \times (a - a^*) \\
\widehat{\Psi}^{\text{PNIE}}_{\text{trad}} \mid L(0), L(1) &= \widehat{\beta}_A \times [\widehat{\gamma}_M + \widehat{\gamma}_{AM} a^*] \times (a - a^*) \\
\widehat{\Psi}^{\text{MIE}}_{\text{trad}} \mid L(0), L(1) &= \widehat{\beta}_A \times \widehat{\gamma}_{AM} \times (a - a^*) \times (a - a^*) \\
\widehat{\Psi}^{\text{RIE}}_{\text{trad}} \mid L(0), L(1) &= \widehat{\gamma}_{AM} \times (\widehat{\beta}_0 + \widehat{\beta}_A a^* + \widehat{\beta}'_{L(0)} L(0) + \widehat{\beta}'_{L(1)} L(1)) \times (a - a^*)
\end{aligned}$$

5.2 G-computation

A simple example of estimation by g-computation can be given for the estimation of the Average Total Effect (ATE). G-computation can be described as a "simple substitution estimator", based on the g-formula described in equation (12) (and Table 5) [Snowden et al., 2011; van der Laan and Rose, 2011].

Firstly, fit a regression of Y on the exposure A and baseline confounders $L(0)$ (using a logistic regression for binary outcomes for example):

$$\bar{Q}(A, L(0)) = \mathbb{E}[Y \mid A, L(0)]$$

Secondly, for each individual, predict the expected values of Y using this model, had the whole population been exposed to $A = 1$ (denoted $\widehat{\bar{Q}}(A = 1, L(0))$, and had the whole population been exposed to $A = 0$ ($\widehat{\bar{Q}}(A = 0, L(0))$. Predicted values are then plugged in the g-formula (for a sample of size n).

$$\widehat{\Psi}^{\text{ATE}}_{\text{gcomp}} = \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{\bar{Q}}(A = 1, L(0)) - \widehat{\bar{Q}}(A = 0, L(0))\right]$$

Confidence intervals can be obtained using bootstrapping.

5.2.1 Parametric G-computation

G-formula estimands of tables 5, 6 and 7 can be obtained using Monte Carlo simulations of the $L(1)_{a'}$, $M_{a'}$, $Y_{a',m}$, $Y_{a,M_{a'}}$, $Y_{a,G_{a'|L(0)}}$ or $Y_{a,\Gamma_{a'|L(0),L(1)}}$ variables under the counterfactual scenarios considered to define the causal quantities of interest [Robins and Hernàn, 2009; Daniel et al., 2013].

As an example for the Controlled direct effect $\Psi^{\text{CDE}_m} = \mathbb{E}(Y_{1,m}) - \mathbb{E}(Y_{0,m})$, the expectation $\mathbb{E}(Y_{a',m})$ (where $a' = \{0,1\}$)) can be estimated from datasets generated by a causal structure where $L(1)$ is affected by the exposure (as in Figure 11(b)):

1. Estimate the probability density function of $L(1)$ conditional on the previous variables $L(0)$ and A. If $L(1) = \{L(1.1), L(1.2), L(1.3), \ldots\}$ is a set of several covariates, it is possible to estimate a model for each of them applying an arbitrary sequential factorisation $f(L(1.1) \mid A, L(0))$, $f(L(1.2) \mid A, L(0), L(1.1))$, $f(L(1.3) \mid A, L(0), L(1.1), L(1.2))$, etc;
2. Simulate individual values of $L(1)_{a'}$ under the counterfactual scenario had $A = a'$ for the whole population, using the $L(1)$ densities estimated at the previous step;
3. Fit a regression of Y conditional on the observed upstream variables $\mathbb{E}(Y \mid L(0), A, L(1), M)$;
4. Using the regression of Y, simulate individual values of $\mathbb{E}(Y \mid l(0)_i, a', l(1)_{a',i}, m)_i$ under the counterfactual scenario setting $A = a'$ and $M = m$ in the whole population, and setting the $L(1)$ values to the simulated values $L(1)_{a'}$ from the second step of the procedure.
5. Estimate $\mathbb{E}(Y_{a',m})$ by the mean of the individual simulated values at the previous step.

Estimations using parametric g-computation have been described to estimate:

- Controlled Direct Effects [Daniel et al., 2011];
- Natural Direct and Indirect Effects [Daniel et al., 2011]. An additional step to estimate the density function of the mediator is necessary, in order to simulate individual values of the mediator $M_{a'}$ under the counterfactual scenario setting $A = a'$.
- Marginal Randomised Direct and Indirect Effects [Lin et al., 2017]. An additional step to estimate the density function of the mediator is necessary, in order to simulate and randomly permute individual values of the mediator $M_{a'}$ under the counterfactual scenario setting $A = a'$.

The approach described for MRDE and MRIE can be adapted to estimate Conditional Randomised Direct and Indirect Effects.

5.2.2 G-computation by Iterative Conditional Expectation

A limit of parametric G-computation is the difficulty to estimate density functions of $L(1)$ variables. Moreover, it is necessary to fit a model for each variable in the set $L(1)$.

An alternative approach is g-computation by iterative conditional expectation (ICE), which

have been described to analyse counterfactual scenarios relevant for Controlled direct effects, Marginal or Randomised natural direct and indirect effects [Bang and Robins, 2005; Petersen et al., 2014; Zheng and van der Laan, 2017]. This approach relies on a smaller number of models to fit (especially if several variables are included in the set $L(1)$).

As an example for Controlled direct effects, the estimand described in Table 5 can be reformulated by iterative conditional expectation,

$$\begin{aligned}\Psi^{\text{CDE}} =&\mathbb{E}(\mathbb{E}_{L(1)}[\mathbb{E}_Y(Y \mid L(1), L(0), A, M = m) \mid L(0), A = a]) \\ &- \mathbb{E}(\mathbb{E}_{L(1)}[\mathbb{E}_Y(Y \mid L(1), L(0), A, M = m) \mid L(0), A = a^*])\end{aligned}$$

and the counterfactual quantity $\mathbb{E}(Y_{a',m})$ can be estimated by the following procedure:

1. Fit a model of the outcome, conditional on all the observed upstream variables

$$\bar{Q}_Y(M) = \mathbb{E}(Y \mid L(0), A, L(1), M)$$

2. For each individual, predict values of Y using the model fitted at the previous step, had the whole population been exposed to $M = m$; Predicted values are denoted $\hat{\bar{Q}}_Y(m)$;
3. Fit a model of the predicted values $\hat{\bar{Q}}_Y(m)$, conditional on the observed variables upstream of $L(1)$

$$\bar{Q}_{L(1)}(A) = \mathbb{E}\left(\hat{\bar{Q}}_Y(m) \mid L(0), A\right)$$

4. Predict individual values of $\bar{Q}_{L(1)}$ using the model fitted at the previous step, had the whole population been exposed to $A = a'$. Predicted values are denoted $\hat{\bar{Q}}_{L(1)}(a')$;
5. Estimate $\mathbb{E}(Y_{a',m})$ by the mean $\sum_i^n \frac{1}{n}\hat{\bar{Q}}_{L(1)}(a')$.

Similar procedures to estimate Marginal or Conditional Direct and Indirect Effects using g-computation by conditional iterative expectation are described in the online supplementary material.

5.2.3 Statistical properties of g-computation estimators

Estimations of direct and indirect effects by g-computation are expected to be unbiased if the models fitted during the procedures ($\bar{Q}$ regressions) are correctly specified [Snowden et al., 2011; van der Laan and Rose, 2011].

Estimates are unaffected by deviation from the positivity assumption, so that the procedure is able to extrapolate beyond the observed data. Considering the positivity assumption is all the more important to avoid conclusions that are only weakly supported by the available data

[Snowden et al., 2011].

Moreover, G-computation is not an asymptotically linear estimator, so that its efficiency properties are not optimal [van der Laan and Rose, 2011].

5.3 Marginal structural models (MSM)

Marginal structural models are models of the expected value of a counterfactual outcome under study. They are used to summarize the causal relationship between the expectation of the counterfactual outcome and the exposures of interest [Robins et al., 2000; Neugebauer and van der Laan, 2007]. In the context of mediation analyses, exposures of interest are the initial exposure A and the mediator M.

5.3.1 MSM for Controlled Direct Effects

The following MSM can be considered to estimate Controlled direct effects [VanderWeele, 2009a; Vansteelandt, 2009]:

$$\mathbb{E}(Y_{a,m}) = \alpha_0 + \alpha_A a + \alpha_M m \tag{17}$$

If we suspect the presence of $A * M$ interaction affecting the outcome, it is possible to add an interaction term:

$$\mathbb{E}(Y_{a,m}) = \alpha_0 + \alpha_A a + \alpha_M m + \alpha_{AM}(a \times m) \tag{18}$$

The controlled direct effects CDE_m can then be expressed using the coefficients of the MSM, for example using equation (18):

$$\begin{aligned}
\Psi^{\text{CDE}_m} =& \mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m}) \\
\Psi^{\text{CDE}_m}_{\text{MSM}} =& (\alpha_0 + \alpha_A a + \alpha_M m + \alpha_{AM}(a \times m)) \\
& - (\alpha_0 + \alpha_A a^* + \alpha_M m + \alpha_{AM}(a^* \times m)) \\
\Psi^{\text{CDE}_m}_{\text{MSM}} =& \alpha_A(a - a^*) + \alpha_{AM}(a - a^*) \times m
\end{aligned}$$

5.3.2 MSM for Natural Direct and Indirect Effects

For Pure Natural Direct Effects and Total Natural Indirect Effects, VanderWeele suggested using two MSM, a model of $Y_{a,m}$ and a model of M_a, conditional on the baseline confounders $L(0)$, where h and g are the link functions chosen by the analyst [VanderWeele, 2009a]:

$$\mathbb{E}(Y_{a,m} \mid l(0)) = h^{-1}(a, m, l(0)) \tag{19}$$

$$\mathbb{E}(M_a \mid l(0)) = g^{-1}(a, l(0)) \tag{20}$$

If h^{-1} is linear in m (meaning that interactions between M and other variables are possible, but not polynomial functions of M or other transformations of M such as $\log(M)$, $\sqrt{M}$, etc), and if A does not affect confounders $L(1)$ of the $M \to Y$ relationship, then

$$\mathbb{E}(Y_{a,M_{a^*}} \mid l(0)) = h^{-1}\left[a, g^{-1}[a^*, l(0)], l(0)\right] \tag{21}$$

Then the PNDE and TNIE as expressed in Table 5 can be reformulated using the MSM functions:

$$\begin{aligned}
\Psi_{\text{MSM}}^{\text{PNDE}} &= \mathbb{E}(Y_{a,M_{a^*}}) - \mathbb{E}(Y_{a^*,M_{a^*}}) \\
&= \sum_{l(0)} \left[h^{-1}(a, g^{-1}[a^*, l(0)], l(0)) - h^{-1}(a^*, g^{-1}[a^*, l(0)], l(0))\right] \times \mathbb{P}(L(0) = l(0)) \qquad (22) \\
\Psi_{\text{MSM}}^{\text{TNIE}} &= \mathbb{E}(Y_{a,M_a}) - \mathbb{E}(Y_{a,M_{a^*}}) \\
&= \sum_{l(0)} \left[h^{-1}(a, g^{-1}[a, l(0)], l(0)) - h^{-1}(a, g^{-1}[a^*, l(0)], l(0))\right] \times \mathbb{P}(L(0) = l(0)) \qquad (23)
\end{aligned}$$

For example, if Y and M are continuous variables, one can use identity functions for g and h:

$$\begin{aligned}
\mathbb{E}(Y_{a,m} \mid l(0)) &= h^{-1}(a, m, l(0)) \\
&= \alpha_0 + \alpha_A a + \alpha_M m + \alpha_{AM}(a \times m) + \alpha'_{L(0)} l(0) \qquad (24) \\
\mathbb{E}(M_a \mid l(0)) &= g^{-1}(a, l(0)) = \beta_0 + \beta_A a + \beta'_{L(0)} l(0) \qquad (25)
\end{aligned}$$

Plugging equations (24) and (25) in equations (22) and (23) gives:

$$\begin{aligned}
\Psi_{\text{MSM}}^{\text{PNDE}} &= \sum_{l(0)} \left[\alpha_A + \alpha_{AM} \times (\beta_0 + \beta_A a^* + \beta'_{L(0)} l(0))\right] \times (a - a^*) \times \mathbb{P}(L(0) = l(0)) \\
\Psi_{\text{MSM}}^{\text{TNIE}} &= \sum_{l(0)} \left[(\alpha_M + \alpha_{AM} a) \times \beta_A \times (a - a^*)\right] \mathbb{P}(L(0) = l(0)) \\
&= (\alpha_M + \alpha_{AM} a) \times \beta_A \times (a - a^*)
\end{aligned}$$

Alternatively, [Lange et al., 2012] introduced a another "unified" approach for estimating natural direct and indirect effects using the following MSM:

$$h\left[\mathbb{E}(Y_{a,M_{a^*}})\right] = \alpha_0 + \alpha_1 a + \alpha_2 a^* + \alpha_3 a \times a^* \tag{26}$$

where h is an appropriate link function. On the difference scales, natural direct and indirect

effects can be expressed using the MSM parameters of equation (26):

$$\begin{aligned}
\Psi^{\text{PNDE}}_{\text{MSM.unified}} &= \mathbb{E}\left(Y_{a,M_{a^*}}\right) - \mathbb{E}\left(Y_{a^*,M_{a^*}}\right) \\
&= h^{-1}\left[\alpha_0 + \alpha_1 a + \alpha_2 a^* + \alpha_3 a \times a^*\right] - h^{-1}\left[\alpha_0 + \alpha_1 a^* + \alpha_2 a^* + \alpha_3 (a^*)^2\right] \\
\Psi^{\text{TNIE}}_{\text{MSM.unified}} &= \mathbb{E}\left(Y_{a,M_a}\right) - \mathbb{E}\left(Y_{a,M_{a^*}}\right) \\
&= h^{-1}\left[\alpha_0 + \alpha_1 a + \alpha_2 a + \alpha_3 a^2\right] - h^{-1}\left[\alpha_0 + \alpha_1 a + \alpha_2 a^* + \alpha_3 a \times a^*\right]
\end{aligned}$$

Applying h as a logit link function, Natural direct and indirect effects can be expressed on odds ratio scales :

$$\begin{aligned}
\Psi^{\text{PNDE}.OR}_{\text{MSM.unified}} &= \frac{\text{odds}[\mathbb{E}(Y_{a,M_{a^*}})]}{\text{odds}[\mathbb{E}(Y_{a^*,M_{a^*}})]} \\
&= \exp\left[\alpha_1(a - a^*) + \alpha_3 a^*(a - a^*)\right] \\
\Psi^{\text{TNIE}.OR}_{\text{MSM.unified}} &= \frac{\text{odds}[\mathbb{E}(Y_{a,M_a})]}{\text{odds}[\mathbb{E}(Y_{a,M_{a^*}})]} \\
&= \exp\left[\alpha_2(a - a^*) + \alpha_3 a(a - a^*)\right]
\end{aligned}$$

5.3.3 MSM for Marginal Randomised Natural Direct and Indirect Effects

As for Natural direct and indirect effects, [VanderWeele, 2017; Vansteelandt and Daniel, 2017] suggested to use two marginal structural models:

$$\mathbb{E}(Y_{a,m} \mid l(0) = h^{-1}(a, m, l(0)) \quad \text{and} \quad \mathbb{P}(M_a = m \mid l(0)) = g^{-1}(a, l(0))$$

and combine those two MSM in order to define the causal quantities necessary for Marginal Randomised direct and indirect effects:

$$\mathbb{E}(Y_{a,G_{a^*|L(0)}}) = \sum_{l(0)} \sum_{m} \mathbb{E}(Y_{a,m} \mid l(0)) \times P(M_{a^*} = m \mid l(0)) \times \mathbb{P}(L(0) = l(0))$$

For example, if Y is continuous and M is binary, one can use the identity function for h (as in equation (24)) and the logit function for g:

$$\mathbb{E}(M_a \mid l(0)) = g^{-1}(\beta_0 + \beta_A a + \beta'_{L(0)} l(0))$$

MRDE and MRIE can then be formulated using the parameters of the MSM:

$$\Psi_{\text{MSM}}^{\text{MRDE}} = \sum_{l(0)} \left[\alpha_A + \alpha_{AM} \times g^{-1}(\beta_0 + \beta_A a^* + \beta'_{L(0)} l(0))\right] (a - a^*) \mathbb{P}(L(0) = l(0))$$

$$\Psi_{\text{MSM}}^{\text{MRIE}} = \sum_{l(0)} (\alpha_M + \alpha_{AM} a) \left[g^{-1}(\beta_0 + \beta_A a + \beta'_{L(0)} l(0)) - g^{-1}(\beta_0 + \beta_A a^* + \beta'_{L(0)} l(0))\right] \mathbb{P}(L(0) = l(0))$$

5.3.4 Estimation of the MSM parameters

Because MSM are models of unobserved counterfactual variables, estimators of the MSM parameters are necessary. Several methods have been described to estimate MSM parameters, based on g-computation, Inverse Probability of Treatment Weighting or double robust methods [Robins et al., 2000; Neugebauer and van der Laan, 2007; VanderWeele, 2009a; Snowden et al., 2011].

Most often, MSM parameters are estimated using Inverse Probability of Treatment Weighting (cf. subsection 5.4.2).

5.4 Inverse probability of treatment weighting (IPTW)

Intuitively, estimators based on Inverse probability of treatment weighting (IPTW) operates by assigning a weight to each individual so that baseline and intermediate confounders are balanced relative to the exposure A and the mediator M in the new pseudo-population, so that there is no confounding between A (or M) and $Y_{a,m}$ [Hernàn and Robins, 2006].

5.4.1 Estimating Average Total Effect and Controlled Direct Effects by IPTW

For Average Total Effects ($\text{ATE} = \mathbb{E}(Y_a) - \mathbb{E}(Y_a^*)$**),** it is possible to estimate $\mathbb{E}(Y_{a'})$ applying the following weight to the outcome of each individual (where $g(A_i \mid L(0)_i)$ is the probability of receiving his observed exposure A_i, given $L(0)_i$):

$$w_{\text{ATE}} = \frac{I(A_i = a')}{g(A_i \mid L(0)_i)}$$

so that the ATE can be estimated by the following Horvitz and Thompson estimator [Horvitz and Thompson, 1952; Hernàn and Robins, 2006]:

$$\widehat{\Psi}_{\text{IPTW}}^{ATE} = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a)}{\widehat{g}(A_i = a \mid L(0)_i)} Y_i - \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a^*)}{\widehat{g}(A_i = a^* \mid L(0)_i)} Y_i$$

In case of positivity violation (if $g(A_i \mid L(0)_i) = 0$ in some strata $L(0)$), weights cannot be computed. Near positivity violation (if $g(A_i \mid L(0)_i) \approx 0$ in some strata $L(0)$) will result in

extreme weights, increasing the variance of the IPTW estimator.

In order to reduce variability resulting from near positivity violation, common approaches are:

- to truncate the weights (for example at the 1st and 99th percentiles, or applying a data-adaptive selection of the truncation level),
- to trim the weights (drop units with propensity scores outside a given interval),

but this will also result in a biased IPTW estimator [Cole and Hernàn, 2008; Bembom and van der Laan, 2008; Crump et al., 2009; Stürmer et al., 2010; Xiao et al., 2013; Lepage et al., 2016; Ju et al., 2019].

Alternatively, a "stabilized" IPTW estimator can be applied using a modified Horvitz-Thomson estimator [Hernàn and Robins, 2006]:

$$\widehat{\Psi}^{ATE}_{\text{sIPTW}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)g^*(A_i=a)}{\widehat{g}(A_i=a|L(0)_i)}Y_i}{\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)g^*(A_i=a)}{\widehat{g}(A_i=a|L(0)_i)}} - \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a^*)g^*(A_i=a^*)}{\widehat{g}(A_i=a^*|L(0)_i)}Y_i}{\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a^*)g^*(A_i=a^*)}{\widehat{g}(A_i=a^*|L(0)_i)}}$$

where $g^*(A_i = a)$ is a non-null function of A (for example, $g^*(A = a) = P(A = a)$).

Using stabilised IPTW estimator enables to get a bounded estimator and a weaker positivity condition (the denominator can be zero if the numerator is zero) [Cole and Hernàn, 2008].

For controlled direct effects ($\text{CDE}_m = \mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m})$**),** the Horvitz-Thomson IPTW estimator is:

$$\begin{aligned}\widehat{\Psi}^{\text{CDE}_m}_{\text{IPTW}} =& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i = a, M_i = m)}{\widehat{g}(A_i = a \mid L(0)_i) \times \widehat{g}(M_i = m \mid L(1)_i, A_i, L(0)_i)}Y_i\right) \\ & - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i = a^*, M_i = m)}{\widehat{g}(A_i = a^* \mid L(0)_i) \times \widehat{g}(M_i = m \mid L(1)_i, A_i, L(0)_i)}Y_i\right)\end{aligned}$$

and the alternative "stabilized" IPTW estimator is:

$$\begin{aligned}\widehat{\Psi}^{\text{CDE}_m}_{\text{sIPTW}} =& \frac{\sum_{i=1}^{n}\left(\frac{I(A_i=a,M_i=m)}{\widehat{g}(A_i=a|L(0)_i)\times\widehat{g}(M_i=m|L(1)_i,A_i,L(0)_i)}Y_i\right)}{\sum_{i=1}^{n}\left(\frac{I(A_i=a,M_i=m)}{\widehat{g}(A_i=a|L(0)_i)\times\widehat{g}(M_i=m|L(1)_i,A_i,L(0)_i)}\right)} \\ & - \frac{\sum_{i=1}^{n}\left(\frac{I(A_i=a^*,M_i=m)}{\widehat{g}(A_i=a^*|L(0)_i)\times\widehat{g}(M_i=m|L(1)_i,A_i,L(0)_i)}Y_i\right)}{\sum_{i=1}^{n}\left(\frac{I(A_i=a^*,M_i=m)}{\widehat{g}(A_i=a^*|L(0)_i)\times\widehat{g}(M_i=m|L(1)_i,A_i,L(0)_i)}\right)}\end{aligned}$$

For conditional randomised direct and indirect effects ($\text{CRDE} = \mathbb{E}(Y_{a,\Gamma_{a^*|L(0),L(1)}}) - \mathbb{E}(Y_{a^*,\Gamma_{a^*|L(0),L(1)}})$**, and** $\text{CRIE} = \mathbb{E}(Y_{a,\Gamma_{a|L(0),L(1)}}) - \mathbb{E}(Y_{a,\Gamma_{a^*|L(0),L(1)}})$**),** [Zheng and van der Laan, 2017] described the

following IPTW estimator:

$$\widehat{\mathbb{E}}(Y_{a,\Gamma_{a'|L(0),L(1)}}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i = a) \times Y_i}{\widehat{g}(A_i = a_i \mid L(0)_i)} \times \frac{\widehat{g}(M_i = m_i \mid a', L(1)_i, L(0)_i)}{\widehat{g}(M_i = m_i \mid a, L(1)_i, L(0)_i)}\right)$$

5.4.2 Estimation of MSM parameters

IPTW methods are frequently applied to estimate parameters of Marginal Structural Models described in subsection 5.3).

For controlled direct effects, parameters of the MSM (17) and (18) can be estimated by fitting a weighted generalised linear model of the observed outcomes Y_i on the exposure A and the mediator M (according to the models (17) or (18)), where the individual weights are:

$$w_i = w_{A,i} \times w_{M,i}$$

where

$$w_{A,i} = \frac{1}{\widehat{g}(A_i \mid L(0)_i)}, \quad \text{and} \quad w_{M,i} = \frac{1}{\widehat{g}(M_i \mid L(0)_i, A_i, L(1)_i)}$$

Alternatively, a recommended approach is to use "stabilised" weights, where individual weights are:

$$sw_i = sw_{A,i} \times sw_{M,i}.$$

where

$$sw_{A,i} = \frac{\widehat{g}(A_i)}{\widehat{g}(A_i \mid L(0)_i)}, \quad \text{and} \quad sw_{M,i} = \frac{\widehat{g}(M_i \mid A_i)}{\widehat{g}(M_i \mid L(0)_i, A_i, L(1)_i)}.$$

For natural direct and indirect effects, parameters of the MSM (19) can be estimated by fitting a weighted generalised linear model of the observed outcomes Y_i on the exposure A, the mediator M and the baseline confounders $L(0)$, according to the model (19). The individual weights are the same than the weights w_i or the stabilised weights sw_i used for controlled direct effects. Numerators of stabilised weights can also be defined conditional on $L(0)$, resulting in $sw_{A,i} = 1$ regarding the stabilised weight associated with the exposure A.

Parameters of the MSM (20) can be estimated by fitting a weighted generalised linear model of the observed mediator M_i on the exposure A and the baseline confounders $L(0)$, according to the model (20). The individual weights are $w_{A,i}$ (or $sw_{A,i}$ for stabilised weights) as defined for the estimation of controlled direct effects.

For the "unified" approach described by Lange *et al.*, parameters of the MSM (26) can be estimated by the following procedure for binary exposures [Lange et al., 2012]:

i. Create a new data set by repeating twice the original data set;
ii. Add a variable A^* which should be equal to $A^* = A$ in the first replication of the data set, and $A^* = 1 - A$ in the second data set;
iii. Regress the outcome Y on A and A^* as in the formula of the MSM (26) (with the appropriate link function), applying a weighted regression where weights and stabilised weights are respectively:

$$w_{unified,i} = \frac{1}{\widehat{g}(A_i \mid L(0))} \times \frac{\widehat{g}(M_i \mid A = A_i^*, L(0)_i, L(1)_i)}{\widehat{g}(M_i \mid A = A_i, L(0)_i, L(1)_i)}$$
$$\text{and } sw_{unified,i} = \frac{\widehat{g}(A_i)}{\widehat{g}(A_i \mid L(0))} \times \frac{\widehat{g}(M_i \mid A = A_i^*, L(0)_i, L(1)_i)}{\widehat{g}(M_i \mid A = A_i, L(0)_i, L(1)_i)}$$

For marginal randomised natural direct and indirect effects, parameters of the two MSM can be estimated applying the same approach as the approach described in the previous paragraph for natural direct and indirect effects.

5.4.3 Statistical properties of IPTW estimators

IPTW estimators are expected to be unbiased if the models fitted to construct the weights ($g(A \mid L(0))$, $g(M \mid L(0), A, L(1))$) are consistent [Snowden et al., 2011; van der Laan and Rose, 2011].

As indicated earlier, IPTW estimators can be strongly affected by positivity violation, which is expected with data sparsity (with large sets of $L(0)$ and $L(1)$ confounders including continuous variables, or if the exposures A and M are high-dimensional variables). Positivity violation will result in IPTW estimators with increased variance. Using stabilised weights can partially mitigate this variability [Robins et al., 2000; van der Laan and Rose, 2011; Lepage et al., 2016].

Other approaches to reduce variability of IPTW estimators are:

- to truncate the weights. However, weight truncation will also result in increased bias (estimators of $g(A \mid L(0))$ and $g(M \mid L(0), A, L(1))$ are no longer consistent after weight truncation) [Cole and Hernàn, 2008; Lepage et al., 2016]. Several authors suggested improved procedures to choose the truncation levels, using data-adaptive selection of optimal truncation levels [Bembom and van der Laan, 2008; Xiao et al., 2013; Ju et al., 2019].
- to trim the weights, dropping units with propensity scores outside a given interval. Several estimators have been suggested to optimise the trimming strategy and the standard error of the estimations [Stürmer et al., 2010; Yang and Ding, 2018; Garès et al., 2021].

5.5 Doubly robust efficient methods

Double robust methods can be used to mitigate the influence of misspecification of models applied in g-computation or IPTW estimators. For example, Targeted Maximum Likelihood Estimation (TMLE) or Augmented Inverse Probability of Treatment Weighted (A-IPTW) have been described as doubly-robust estimators. They rely on both models of the outcomes ($\bar{Q}$ used in iterative g-computation) and propensity score models (g used in IPTW).

If either $\bar{Q}$ or g models are consistently estimated, double robust methods will be consistent. Moreover, they are efficient if both $\bar{Q}$ and g models are consistently estimated: they can achieve the Cramer-Rao lower bound for the variance of unbiased estimators (the smallest asymptotic variance) [van der Laan and Rose, 2011; Porter et al., 2011; Luque-Fernandez et al., 2018].

Both approach rely on the estimation of the efficient influence curve. Compared to TMLE, A-IPTW is described as less robust to positivity violation, and it might produce estimates outside of the statistical model space [van der Laan and Rose, 2011]. Data-adaptive (machine learning) algorithm can be applied in order to obtain consistant estimates of $\bar{Q}$ and g functions and optimize the statistical properties of the estimators.

- TMLE and A-IPTW procedures have been described, with statistical packages, for the estimation of Average treatment effects (ATE) [van der Laan and Rubin, 2006; van der Laan and Rose, 2011; Gruber and van der Laan, 2012; Zhong et al., 2021].
- A TMLE procedure for repeated exposures has been developped and can be applied to estimate controlled direct effects (CDE) [van der Laan and Gruber, 2012; Petersen et al., 2014; Lendle et al., 2017].
- TMLE and an alternative double robust "one-step" estimator have been described for marginal randomised direct and indirect effects [Rudolph et al., 2017, 2018, 2019; Díaz et al., 2021a]. The corresponding "medoutcon" R package is under development [Hejazi et al., 2018].
- A TMLE procedure for conditional randomised direct and indirect effects has also been described [Zheng and van der Laan, 2017].

6 Discussion

6.1 In summary

For the last twenty years, classical methods in mediation analyses (difference in coefficients, product of coefficients, path analyses and structural equation modelling) have been supplemented by concepts and methods from the causal inference literature:

- Non parametric causal models and graphical approaches (DAGs), to describe hypotheses on the causal structure of the data generating system;
- Counterfactual expressions and notations of causal quantities of interest, allowing more precise definitions of direct and indirect effects dealing with interactions and intermediate confounding affected by the initial exposure;
- Specification of assumptions needed to identify and estimate the causal quantities of interest (consistency, sequential randomisation assumption, positivity);
- Several estimators (g-computation, IPTW, double robust estimators) which can be implemented, with statistical properties varying from one family of estimators to another.

6.2 Additionnal topics not covered in this report

6.2.1 More general longitudinal structures and time-to-event outcomes

In survival contexts and causal models without mediator-outcome confounders affected by the exposure (as in Figure 11(a)), Lange and Hansen suggested using the Aalen additive hazard model to estimate Natural direct and indirect effects [Lange and Hansen, 2011]:

$$\gamma(t, A, M, L(0), L(1)) = \lambda_0(t) + \lambda_A(t)A + \lambda_M(t)M + \lambda_{L(0)}(t)L(0) + \lambda_{L(1)}(t)L(1)$$

where $\gamma(t)$ is the hazard function (event rate at time t, given that a person survived until time t or later) and $\lambda(t)$ are potentially time-dependant coefficient functions.

Assuming a continuous mediator that can be modeled using a linear regression,

$$M = \beta_0 + \beta_A A + \beta_{L(0)}L(0) + \beta_{L(1)}L(1) + \varepsilon$$

and under the usual "no unmeasured confounding" assumptions (i), (ii), (iii) and (iv) necessary for the identification of Natural direct and indirect effects (cf. Table 4), the Pure Natural Direct Effect and Total Natural Indirect Effect can be estimated respectively by $\lambda_A(a - a^*)$ and $\lambda_M \times \beta_A(a - a^*)$ if the exposure and the mediator have no time-dependent effects in the Aalen model ($\lambda_A(t)$ and $\lambda_M(t)$ are constant) [Lange and Hansen, 2011].

In the same context, VanderWeele described alternative approaches using [Vanderweele, 2011]:

- accelerated failure time models and proportional hazard models, which can be applied to estimate PNDE and TNIE on the mean survival time scale,
- proportional hazard models, which can be applied to estimate natural direct and indirect effects on the log hazard ratio difference scale, provided the outcome is rare.

Recently, Tai et al. redefined Natural direct and indirect effects in order to take into account death-truncation of the mediator in survival analyses [Tai et al., 2021].

More general longitudinal structures implying repeated exposures $A(t)$ and mediators $M(t)$, with time-varying covariates $L(t)$ and $R(t)$, such as shown in Figure 12, have been described. In those complex causal models, it is possible to estimate:

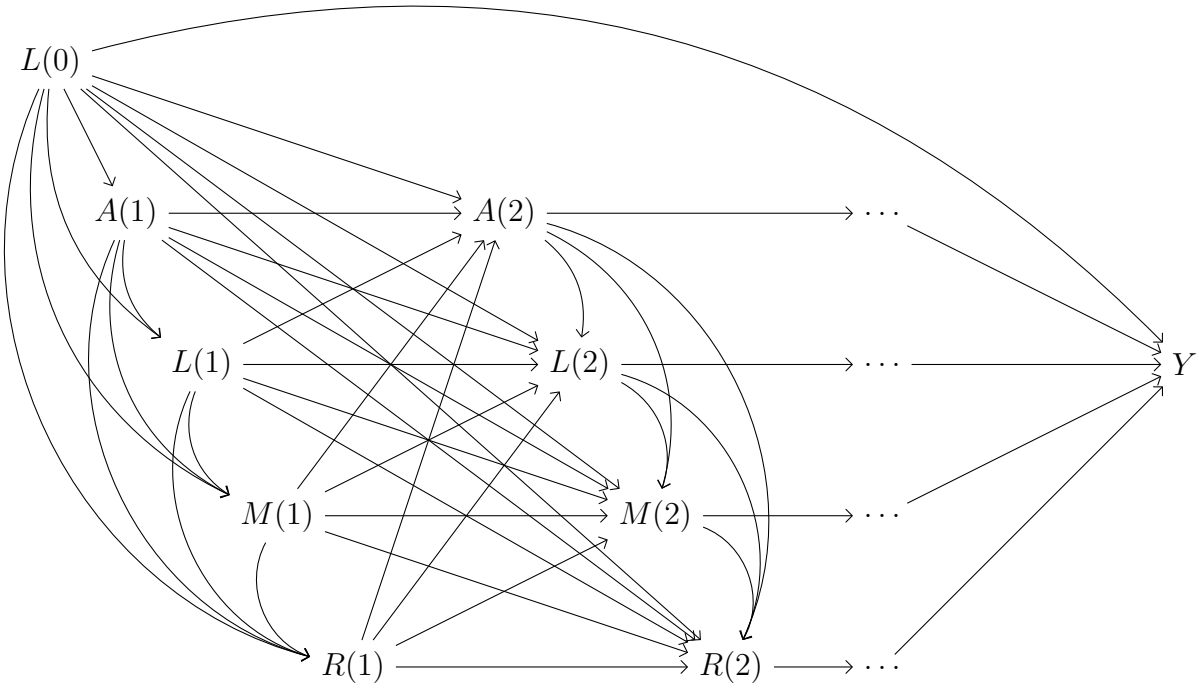
- Controlled direct effects, which can be considered as effects of repeated exposures with time-varying covariates [Robins and Hernàn, 2009].
- Marginal randomised direct and indirect effects as detailed by VanderWeele *et al.* [Vanderweele et al., 2014; Lin et al., 2017; VanderWeele, 2017]. However, as indicated earlier, MRDE and MRIE are not well defined in survival settings if participants can die before mediator occurrences [Zheng and van der Laan, 2017].
- Conditional randomised direct and indirect effects as detailed by Zheng and van der Laan [Zheng and van der Laan, 2017]. Considering that $Y(t)$ is a time dependant outcome included in the $L(t)$ set, CRDE and CRIE are well defined when conditioning on the participant's time-varying history. It's also possible to model informative censoring mechanisms, considering an indicator of remaining uncensored at time t in the $A(t)$ set of exposure variables. An indicator of being monitored at time t can also be added in the $A(t)$ set, in order to take into account missing values at time t for participants who missed a visit at time t but who are still alive and uncensored.

6.2.2 Multiple mediators

In causal structures without mediator-outcome confounders affected by the exposure (as in Figure 11(a)), [VanderWeele and Vansteelandt, 2014] described a situation with multiple mediators where $M = (M^{(1)}, M^{(2)}, \ldots, M^{(k)})$ is a vector including all the mediators of interest. In this context, they suggested using traditional regression approaches in order to estimate Natural direct and indirect effects or controlled direct effets:

i. Assess all the mediators of interest as a single mediator M, defined as the entire vector of mediators.
ii. or assess mediators sequentially: first $M^{(1)}$, then $(M^{(1)}, M^{(2)})$ jointly, then $(M^{(1)}, M^{(2)}, M^{(3)})$ jointly, etc.

Figure 12: More general longitudinal causal models, with repeated exposures and mediators



The first approach does not require knowing the ordering of the mediators. If ordering of the mediator is known, the second approach can correctly give more details on the portion of the total effect mediated through $M^{(1)}$, the portion of the total effect mediated through both $(M^{(1)}, M^{(2)})$ and deduce the additional contribution of $M^{(2)}$ beyond $M^{(1)}$. This additional contribution would correspond to the additional effect mediated only through $M^{(2)}$, which is added to the possible paths going through both $M^{(1)}$ and $M^{(2)}$ (if $M^{(1)}$ affects $M^{(2)}$). The sequential analysis can then be continued to assess the effect mediated through $(M^{(1)}, M^{(2)}, M^{(3)})$, etc.

[Steen et al., 2017] extented this sequential method in a similar context (no mediator-outcome confounder affected by the exposure) and gave an example implying 2 mediators $M = (M^{(1)}, M^{(2)})$. They applied the "unified" MSM approach in order to obtain a 3-way decomposition with [Lange et al., 2012]:

- a Natural direct effect (not mediated by $M^{(1)}$, nor $M^{(2)}$), corresponding to the path $A \to Y$;
- a Natural indirect effect with respect to $M^{(1)}$, implying two paths: $A \to M^{(1)} \to Y$ and $A \to M^{(1)} \to M^{(2)} \to Y$;
- a partial indirect effect with respect to $M^{(2)}$, implying the remaining path $A \to M^{(2)} \to Y$.

They described 6 possible decompositions, according to the way interaction terms are considered in definitions of direct and indirect effects [Steen et al., 2017].

More generally, Marginal and Conditional Randomised Direct and Indirect effects can be applied when dealing with a set of intermediate variables (with known ordering) in which some are considered as mediators of interest and the others as time-varying confounders [Vanderweele et al., 2014; VanderWeele, 2017; Zheng and van der Laan, 2017].

In a causal structure implying 2 mediators $M^{(1)}$ and $M^{(2)}$, [Vansteelandt and Daniel, 2017] applied the principles of interventional effects (drawing mediators in counterfactual distributions under $A = a$ or $A = a^*$) to decompose an overall effect into:

- a direct effect of $A \to Y$;
- an indirect effect via the first mediator;
- an indirect effect via the second mediator, including the path via the first and second mediator if the first mediator affects the second;
- an indirect effects corresponding to interaction effects between mediators on the outcome, or exposure-mediator interactions.

Interestingly, this approach can be applied if the structural dependence between the mediators is unknown (regarding the direction of the causal effects from one mediator to the other, or the presence of an unmeasured common cause) [Vansteelandt and Daniel, 2017].

6.2.3 Addressing potential biases: Measurement errors, unmeasured confounding, selection bias

Measurement errors. Measurement errors and misclassification regarding the exposure, mediator, outcome or confounders can lead to bias in the estimation of the causal quantities of interest. For estimators based on traditional regression methods, and assuming classical measurement error models, simple approaches such as regression calibration or SIMEX (simulation extrapolation) or more complex methods (likelihood methods, bayesian methods, moment reconstruction and moment-adjusted imputation, multiple imputation) can be applied in order to estimate average total effects [Keogh et al., 2020; Shaw et al., 2020].

Regarding causal quantities of interest in mediation analyses, various methods have been described to take into account measurement errors. Most of those methods and results were discussed in the causal framework of Figure 11(a) without mediator-outcome confounding affected by the exposure.

- For measurement errors regarding binary or continuous mediators, effect of nondifferential measurement errors depends on the presence of exposure-mediator $A*M$ interactions and the signs of the effects of the exposure on the mediator ($A \to M$) and the effect of the mediator on the outcome ($M \to Y$) [Ogburn and VanderWeele, 2012; VanderWeele et al., 2012; Blakely et al., 2013; Valeri and Vanderweele, 2014; Valeri et al., 2014]. Methods to correct bias or apply sensitivity analyses include regression calibration, EM algorithms, SIMEX, method of moments [le Cessie et al., 2012; Valeri and Vanderweele, 2014; Valeri

et al., 2014].

- For misclassification regarding binary exposures, SIMEX methods or EM algorithms have been described to correct bias [Valeri et al., 2017; Jiang and VanderWeele, 2019].
- For measurement errors regarding the outcome, [Jiang and VanderWeele, 2015] discussed the expected bias under classical nondifferential measurement error models for continuous or binary outcomes. Regression calibration or EM correction methods are described.

Unmeasured confounding. Estimation of causal quantities of interest in mediation analyses rely strongly on the sequential randomisation assumptions or other forms of "no unmeasured confounding" (Table 4). These assumptions can be assessed by sensitivity analysis (or bias analysis). For estimation of total effects, sensitivity analyses aim to assess "the combination of bias parameters that could wholly explain the observed association if no effect of the exposure on the outcome truly existed" [Lash et al., 2009; Imbens and Rubin, 2015].

More recently, [Ding and VanderWeele, 2016] derived a bounding factor without imposing assumptions on the unmeasured confounders. They introduced the "E-value", corresponding to the "minimum strength of association that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment–outcome association, conditional on the measured covariates" [Ding and VanderWeele, 2016; VanderWeele and Ding, 2017]. An R package is available for the calculation of E-values [Mathur et al., 2018]. It should be noted that carrying out sensitivity analyses with E-value is more appropriate with causal effects expressed on the risk ratio (or odds-ratio) scale [Ding and VanderWeele, 2016]. Because many causal quantities of interest in mediation analyses are expressed on difference scales, it seems useful to express the results also on risk ratio scales. In the context of mediation analyses without mediator-outcome confounders affected by the exposure (as in Figure 11(a)), Smith and VanderWeele showed that E-values could be applied to most situations in order to assess the unmeasured mediator-outcome confounding assumption [Smith and VanderWeele, 2019].

A more general approach to sensitivity analyses was presented by [Díaz and van der Laan, 2013], in order to assess violation of randomisation assumptions or bias due to measurement errors. The approach does not rely on additional models and can be implemented with asymptotically linear estimators such as TMLE.

Selection bias. [Valeri and Coull, 2016] discussed selection bias arising from missing data and its consequences on the estimation of direct and indirect effects. They suggest using nonparametric sensitivity analyses. More generally, the usual approaches described for dealing with missing data can be applied for mediation analysis [Carpenter and Smuk, 2021].

6.3 Future prospects

The causal inference approaches are framed in confirmatory analyses where the structural hypotheses are assumed to be correct for the causal model. To our knowledge, we lack some tools and guidelines to conduct mediation analyses with more exploratory objectives. For example, beginning with a general scientific objective of exploring the intermediate mechanisms between an initial exposure and the outcome, using a large set of variables:

- Starting from an exploratory approach, how can we state relevant structural causal models (DAGs) combining theoretical knowledge and observed data?
- Dealing with a large set of intermediate variables, how can we explore and quantify the larger indirect effects, identify the most interesting mediators in order to understand mechanisms or suggest possible interventions?

References

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–73, 2005.

R. M. Baron and D. A. Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–82, 1986.

Oliver Bembom and Mark van der Laan. Data-adaptive selection of the adjustment set in variable importance estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, page Paper 231, 2008.

Tony Blakely, Sarah McKenzie, and Kristie Carter. Misclassification of the mediator matters when estimating indirect effects. *Journal of epidemiology and community health*, 67(5):458–66, 2013.

K.A. Bollen and R.A. Stine. Bootstrapping goodness-of-fit measures in structural equation models. In K.A. Bollen and J.S. Long, editors, *Testing structural equation models*, page 111–135. Sage, Newbury Park, CA, 1993.

Kenneth A. Bollen. *Structural Equation Models with Observed Variables*. John Wiley & Sons, Ltd, 1989. ISBN 9781118619179. doi: https://doi.org/10.1002/9781118619179.ch4.

Kenneth A. Bollen and Pamela Paxton. Interactions of latent variables in structural equation models. *Strucural Equation Modeling*, 5(3):267–293, 1998.

M.W. Browne. Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37:62–83, 1984.

M.W. Browne and A. Shapiro. Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, 41:193–208, 1988.

J. R. Carpenter and M. Smuk. Missing data: A statistical framework for practice. *Biometrical journal*, 63(5):915–947, 2021.

Stephen R Cole and Miguel A Hernàn. Fallibility in estimating direct effects. *International Journal of Epidemiology*, 31(1):163–5, 2002.

Stephen R Cole and Miguel A Hernàn. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–64, 2008.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

Rhian M. Daniel, Bianca L. De Stavola, and Simon N. Cousens. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, 11(4):479–517, 2011.

Rhian M Daniel, Michael G Kenward, Simon N Cousens, and Bianca L De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, 2012.

R.M. Daniel, S.N. Cousens, B.L. De Stavola, M.G. Kenward, and J.A.C. Sterne. Methods for dealing with time-dependent confounding. *Statistics in medicine*, 32(9):1584–618, 2013.

V Didelez, A Dawid, and S Geneletti. Direct and indirect effects of sequential treatments. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, 2006.

Peng Ding and Tyler J. VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27:368–377, 2016. ISSN 1531-5487.

O Dumas, V Siroux, N Le Moual, and R Varraso. Approches d'analyse causale en épidémiologie. *Revue d'Épidémiologie et de Santé Publique*, 62(1):53–63, 2014.

Ivan Díaz, Nima S Hejazi, Kara E Rudolph, and Mark J van der Laan. Non-parametric efficient causal mediation with intermediate confounders. *Biometrika*, 108(3):627–641, 2021a.

Iván Díaz and Mark J van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics*, 9 (2):149–60, 2013.

Iván Díaz, Nicholas Williams, Katherine L. Hoffman, and Edward J. Schenck. Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, pages 1–16, 2021b.

Valérie Garès, Guillaume Chauvet, and David Hajage. Variance estimators for weighted and stratified linear dose–response function estimators using generalized propensity score. *Biometrical Journal*, 2021.

M Maria Glymour. Using causal diagrams to understand common problems in social epidemiology. In J Michael Oakes and Jay S Kaufman, editors, *Methods in social epidemiology*, pages 393–428. Jossey-Bass, San Francisco, CA, 2006.

M Maria Glymour, Jennifer Weuve, Lisa F Berkman, Ichiro Kawachi, and James M Robins. When is baseline adjustment useful in analyses of change? an example with education and cognitive change. *American Journal of Epidemiology*, 162(3):267–78, 2005.

Maria M. Glymour. Using causal diagrams to understand common problems in social epidemiology. In Oakes J. Michael and Kaufman Jay S., editors, *Methods in social epidemiology, second edition*, pages 458–492. Jossey-Bass and Pfeiffer Imprint, Wiley, San Francisco, CA, 2017.

S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

Susan Gruber and Mark J. van der Laan. tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(13):1–35, 2012.

Andrew F Hayes. An index and test of linear moderated mediation. *Multivariate Behavioral Research*, 50(1):1–22, 2015.

Nima S Hejazi, Ivan Díaz, and Kara E Rudolph. *medoutcon: Efficient causal mediation analysis with intermediate confounders*, 2018. URL https://github.com/nhejazi/medoutcon. R package version 0.1.5.

Miguel A Hernàn and Stephen R Cole. Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, 170(8):959–62; discussion 963–4, 2009.

Miguel A. Hernàn and James M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586, 2006.

Miguel A Hernàn, Sonia Hernàndez-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

Miguel A Hernán, John Hsu, and Brian Healy. A second chance to get causal inference right: a classification of data science tasks. *Chance*, 32(1):42–49, 2019.

Miguel A. Hernán. Does water kill? a call for less casual causal inferences. *Annals of Epidemiology*, 26(10):674–680, 2016. ISSN 1047-2797.

Austin Bradford Hill. The environment and disease: association or causation? *Sage Publications*, 1965.

Paul W. Holland. Statistics and causal inference. *Journal of the american statistical association*, 81(396):945–960, 1986.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Dawn Iacobucci. *Mediation analysis.* Quantitative applications in the social sciences. Sage, Los Angeles, 2008.

Guido W. Imbens and Donald B. Rubin. Sensitivity analysis and bounds. In Guido W. Imbens and Donald B. Rubin, editors, *Causal inference for statistics, social, and biomedical sciences. An introduction*, pages 496–509. Cambridge University Press, New York, 2015.

Zhichao Jiang and Tyler J VanderWeele. Causal mediation analysis in the presence of a mismeasured outcome. *Epidemiology*, 26(1):e8–e9, 2015.

Zhichao Jiang and Tyler J VanderWeele. Causal mediation analysis in the presence of a misclassified binary exposure. *Epidemiologic Methods*, 8(1):20160006, 2019.

Cheng Ju, Joshua Schwab, and Mark J van der Laan. On adaptive propensity score truncation in causal inference. *Statistical methods in medical research*, 28(6):1741–1760, 2019.

Karl G. Jöreskog and Fan Yang. Nonlinear structural equation models: the kenny-judd model with interaction effects. In G.A. Marcoulides and R.E. Schumacker, editors, *Advanced Structural Equation Modeling: Issues and Techniques*, pages 57–88. Psychology Press, New York, 1st edition, 1996.

J. S. Kaufman, R. F. Maclehose, and S. Kaufman. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives and Innovations*, 1(1):4, 2004.

Michelle Kelly-Irving, Benoit Lepage, Dominique Dedieu, Mel Bartley, David Blane, Pascale Grosclaude, Thierry Lang, and Cyrille Delpierre. Adverse childhood experiences and premature all-cause mortality. *European journal of epidemiology*, 28(9):721–34, 2013.

David A. Kenny and Charles M. Judd. Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96(1):201–210, 1984.

Ruth H Keogh, Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Helmut Küchenhoff, Janet A Tooze, Michael P Wallace, Victor Kipnis, and Laurence S Freedman. Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1-basic theory and simple methods of adjustment. *Statistics in medicine*, 39(16):2197–2231, 2020.

Emil Kupek. Log-linear transformation of binary variables: a suitable input for sem. *Structural Equation Modeling: a multidisciplinary journal*, 12(1):28–40, 2005.

Emil Kupek. Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders. *BMC medical research methodology*, 6(1):13, 2006.

Theis Lange and Jorgen V Hansen. Direct and Indirect Effects in a Survival Context. *Epidemiology*, 2(4):575–581, 2011.

Theis Lange, Stijn Vansteelandt, and Maarten Bekaert. A Simple Unified Approach for Estimating Natural Direct and Indirect Effects. *American Journal of Epidemiology*, 176(3): 190–195, 07 2012.

Timothy L. Lash, Matthew P. Fox, and Fink Aliza K. Unmeasured and unknown confounders. In Timothy L. Lash, Matthew P. Fox, and Fink Aliza K., editors, *Applying quantitative bias analysis to epidemiologic data*, Statistics for biology and health, pages 59–78. Springer, New York, 2009.

Debbie A. Lawlor, Kate Tilling, and George Davey Smith. Triangulation in aetiological epidemiology. *International journal of epidemiology*, 45:1866–1886, 2016.

Saskia le Cessie, Jan Debeij, Frits R Rosendaal, Suzanne C Cannegieter, and Jan P Vandenbroucke. Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*, 23(4):551–60, 2012.

Samuel D. Lendle, Joshua Schwab, Maya L Petersen, and Mark J. van der Laan. ltmle: An R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, 81(1):1–21, 2017.

B. Lepage, D. Dedieu, N. Savy, and T. Lang. Estimating controlled direct effects in the presence of intermediate confounding of the mediator-outcome relationship: Comparison of five different methods. *Statistical Methods in Medical Research*, 25(2):553–570, 2016.

Benoît Lepage, Sébastien Lamy, Dominique Dedieu, Nicolas Savy, and Thierry Lang. Estimating the causal effect of an exposure on change from baseline using directed acyclic graphs and path analysis. *Epidemiology*, 26(1):122–129, 2015.

Sheng-Hsuan Lin, Jessica Young, Roger Logan, Eric J. Tchetgen Tchetgen, and Tyler J. VanderWeele. Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators and confounders. *Epidemiology*, 28(2):266–274, 2017.

John C. Loehlin. *Latent variable models : an introduction to factor, path, and structural equation analysis.* L. Erlbaum Associates, Mahwah, N.J., 4th edition, 2004.

Miguel Angel Luque-Fernandez, Michael Schomaker, Bernard Rachet, and Mireille E. Schnitzer. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16):2530–2546, 2018.

David P MacKinnon, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West, and Virgil Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83–104, 2002.

David Peter MacKinnon. *Introduction to statistical mediation analysis*. Lawrence Erlbaum Associates, New York, 2008.

Maya B. Mathur, Peng Ding, Corrine A. Riddell, and Tyler J. VanderWeele. Web site and r package for computing e-values. *Epidemiology*, 29(5):e45–e46, 2018.

Bengt Muthén. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132, 1984.

Bengt Muthén. Goodness of fit with categorical and other non-normal variables. In K.A. Bollen and J.S. Long, editors, *Testing Structural Equation Models*, pages 205–243. Sage, Newbury Park, CA, 1993.

Romain Neugebauer and Mark van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137:419–434, 2007.

Trang Quynh Nguyen, Ian Schmid, and Elizabeth A Stuart. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological methods*, 2020.

Elizabeth L Ogburn and Tyler J VanderWeele. Analytic results on the bias due to nondifferential misclassification of a binary mediator. *American journal of epidemiology*, 176(6):555–61, 2012.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2): Article7, 2010a.

J. Pearl. On the consistency rule in causal inference. axiom, definition, assumption, or theorem? *Epidemiology*, 21(6):872–875, 2010b.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Judea. Pearl. Direct and indirect effects. In D. Koller and J. Breese, editors, *Proceedings of the Seventeenth Conference on Uncertainy in Articial Intelligence*, pages 411–420. Morgan Kaufmann, San Francisco, CA, 2001.

Judea. Pearl. *Causality: models, reasoning, and inference.* Cambridge University Press, Cambridge, U.K.; New York, second edition, 2009a.

Judea Pearl. Causality and structural models in social science and economics. In Judea Pearl, editor, *Causality: models, reasoning, and inference*, pages 155–194. Cambridge University Press, Cambridge, U.K.; New York, second edition, 2009b.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Maya Petersen, Joshua Schwab, Susan Gruber, Nello Blaser, Michael Schomaker, and Mark van der Laan. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference*, 2(2):147–185, 2014.

Maya L Petersen and Mark J van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*, 25(3):418, 2014.

Maya L Petersen, Sandra E Sinisi, and Mark J van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

Kristin E Porter, Susan Gruber, Mark J van der Laan, and Jasjeet S Sekhon. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 7(1):31, 2011.

Kristopher J Preacher, Derek D Rucker, and Andrew F Hayes. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate behavioral research*, 42(1):185–227, 2007.

David H. Rehkopf, M. Maria Glymour, and Theresa L. Osypuk. The consistency assumption for causal inference in social epidemiology: When a rose is not a rose. *Current Epidemiology Reports*, 3(1):63–71, 2016.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–55, 1992.

J. M. Robins and M. A. Hernàn. Estimation of the causal effects of time-varying exposures. In Garrett M. Fitzmaurice, editor, *Longitudinal Data Analysis*, Chapman & Hall/CRC handbooks of modern statistical methods, pages 553–599. CRC Press, Boca Raton, 2009.

J M Robins, M A Hernán, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–60, 2000.

Kenneth J Rothman, S Greenland, Charles Poole, and Lash Timothy L. Causation and causal inference. In Kenneth J Rothman, S Greenland, and Lash Timothy L, editors, *Modern Epidemiology.*, pages 5–31. Lippincott Williams & Wilkins, Philadelphia, USA, 3rd edition, 2008.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 56(5):688–701, 1974.

Donald B. Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47(4):1213–1234, 1991.

Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Kara E Rudolph, Oleg Sofrygin, Wenjing Zheng, and Mark J van der Laan. Robust and flexible estimation of stochastic mediation effects: A proposed method and example in a randomized trial setting. *Epidemiologic Methods*, 7(1):20170007, 2017.

Kara E Rudolph, Oleg Sofrygin, Nicole M Schmidt, Rebecca Crowder, M Maria Glymour, Jennifer Ahern, and Theresa L Osypuk. Mediation of neighborhood effects on adolescent substance use by the school and peer environments. *Epidemiology*, 29(4):590–598, 2018.

Kara E Rudolph, Dana E Goin, Diana Paksarian, Rebecca Crowder, Kathleen R Merikangas, and Elizabeth A Stuart. Causal mediation analysis with observational data: Considerations and illustration examining mechanisms linking neighborhood poverty to adolescent substance use. *American journal of epidemiology*, 188(3):598–608, 2019.

Sharon Schwartz, Seth J. Prins, Ulka B. Campbell, and Nicolle M. Gatto. Is the "well-defined intervention assumption" politically conservative? *Social Science and Medicine*, 166:254–257, 2016.

Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Ruth H Keogh, Victor Kipnis, Janet A Tooze, Michael P Wallace, Helmut Küchenhoff, and Laurence S Freedman. Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2-more complex methods of adjustment and advanced topics. *Statistics in medicine*, 39(16):2232–2263, 2020.

Louisa H. Smith and Tyler J. VanderWeele. Mediational e-values: Approximate sensitivity analysis for unmeasured mediator-outcome confounding. *Epidemiology (Cambridge, Mass.)*, 30:835–837, 2019.

Jonathan M Snowden, Sherri Rose, and Kathleen M Mortimer. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*, 173(7):731–8, 2011.

Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Flexible Mediation Analysis With Multiple Mediators. *American Journal of Epidemiology*, 186(2):184–193, 2017.

Til Stürmer, Kenneth J. Rothman, Jerry Avorn, and Robert J. Glynn. Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution—A Simulation Study. *American Journal of Epidemiology*, 172(7):843–854, 2010.

An-Shun Tai, Chun-An Tsai, and Sheng-Hsuan Lin. Survival mediation analysis with the death-truncated mediator: The completeness of the survival mediation parameter. *Statistics in medicine*, 40(17):3953–3974, 2021.

P Tarka. An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, 52(1):313–354, 2018.

Peter W. G. Tennant, Eleanor J. Murray, Kellyn F. Arnold, Laurie Berrie, Matthew P. Fox, Sarah C. Gadd, Wendy J. Harrison, Claire Keeble, Lynsie R. Ranker, Johannes Textor, Georgia D. Tomova, Mark S. Gilthorpe, and George T. H. Ellison. Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, 2020.

J. Textor, B. van der Zander, M.S. Gilthorpe, M. Liskiewicz, and G.Th. Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016.

Linda Valeri. *Statistical methods for causal mediation analysis*. Ph.d. thesis, Havard university, 2013.

Linda Valeri and Brent A Coull. Estimating causal contrasts involving intermediate variables in the presence of selection bias. *Statistics in medicine*, 35(126):4779–4793, 2016.

Linda Valeri and Tyler J VanderWeele. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros. *Psychological methods*, 18(2):137, 2013.

Linda Valeri and Tyler J Vanderweele. The estimation of direct and indirect causal effects in the presence of misclassified binary mediator. *Biostatistics*, 15(3):498–512, 2014.

Linda Valeri, Xihong Lin, and Tyler J VanderWeele. Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Statistics in medicine*, 33(28):4875–90, 2014.

Linda Valeri, Sarah L Reese, Shanshan Zhao, Christian M Page, Wenche Nystad, Brent A Coull, and Stephanie J London. Misclassified exposure in epigenetic mediation analyses. does dna methylation mediate effects of smoking on birthweight? *Epigenomics*, 9(3):253–265, 2017.

Mark van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Mark J van der Laan and Susan Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1): Article 8, 2012.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Series in Statistics. Springer, New York, NY, 1st edition, 2011.

T. J. VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26, 2009a.

T. J. Vanderweele. Causal mediation analysis with survival data. *Epidemiology*, 22(4):582–5, 2011.

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

T J Vanderweele, S Vansteelandt, and J M Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306, 2014.

Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115, 2014.

Tyler J VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009b.

Tyler J VanderWeele. Invited commentary: structural equation models and epidemiologic analysis. *American Journal of Epidemiology*, 176(7):608–612, 2012.

Tyler J VanderWeele. Policy-relevant proportions for direct effects. *Epidemiology (Cambridge, Mass.)*, 24(1):175, 2013a.

Tyler J VanderWeele. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology (Cambridge, Mass.)*, 24(2):224, 2013b.

Tyler J VanderWeele. A unification of mediation and interaction: a four-way decomposition. *Epidemiology (Cambridge, Mass.)*, 25(5):749, 2014.

Tyler J VanderWeele. Mediation analysis: a practitioner's guide. *Annual review of public health*, 37:17–32, 2016.

Tyler J VanderWeele. Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):917–938, 2017.

Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the e-value. *Annals of internal medicine*, 167:268–274, 2017.

Tyler J VanderWeele and James M Robins. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568, 2007.

Tyler J VanderWeele, Linda Valeri, and Elizabeth L Ogburn. The role of measurement error and misclassification in mediation analysis: mediation and measurement error. *Epidemiology*, 23(4):561–4, 2012.

S. Vansteelandt. Estimating direct effects in cohort and case-control studies. *Epidemiology*, 20 (6):851–60, 2009.

S. Vansteelandt and R. M. Daniel. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258–265, 2017.

Aolin Wang and Onyebuchi A Arah. G-computation demonstration in causal mediation analysis. *European journal of epidemiology*, 30(10):1119–1127, 2015.

Daniel Westreich and Stephen R Cole. Invited commentary: positivity in practice. *American journal of epidemiology*, 171(6):656–63, 2010.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.

Sewall Wright. Path coefficients and path regressions: alternative or complementary concepts? *Biometrics*, 16(2):189–202, 1960.

Yongling Xiao, Erica E.M. Moodie, and Michal Abrahamowicz. Comparison of approaches to weight truncation for marginal structural cox models. *Epidemiologic Methods*, 2(1):1–20, 2013.

S Yang and P Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 03 2018.

Wenjing Zheng and Mark van der Laan. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *Journal of causal inference*, 5(2), 2017.

Wenjing Zheng and Mark J van der Laan. Causal mediation in a survival setting with time-dependent mediators. *Collection Of Biostatistics Research Archive*, 2012.

Yongqi Zhong, Edward H. Kennedy, Lisa M. Bodnar, and Ashley I. Naimi. Aipw: An r package for augmented inverse probability–weighted estimation of average causal effects. *American Journal of Epidemiology*, 2021.

7 Appendix

7.1 Average total effect (ATE)

Using counterfactual notations, the *average total effect* (ATE) is defined as

$$\text{ATE} = \mathbb{E}(Y_{A=a}) - \mathbb{E}(Y_{A=a^*})$$

Under the identification assumption described in paragraph 4.1.1, the ATE can be expressed using parameters of the observed data distribution. For example, under the hypothetical intervention setting $A = a$,

$$\mathbb{E}(Y_{A=a}) = \sum_{l(0)} \mathbb{E}\left[Y_{A=a} \mid l(0)\right] \times P\left(L(0) = l(0)\right)$$

under the randomisation assumption (9), we have:

$$\mathbb{E}(Y_{A=a}) = \sum_{l(0)} \mathbb{E}\left[Y_{A=a} \mid A = a, l(0)\right] \times P\left(L(0) = l(0)\right)$$

under the consistency assumption (11), we have:

$$\mathbb{E}(Y_{A=a}) = \sum_{l(0)} \mathbb{E}\left[Y \mid A = a, l(0)\right] \times P\left(L(0) = l(0)\right)$$

$\mathbb{E}\left[Y \mid A = a, l(0)\right] \times P\left(L(0) = l(0)\right)$ is called the *g-formula* and is based on a truncated factorisation corresponding to a DAG of Figure 7(a) in which the hypothetical intervention on A would locally modify the graph, removing the arrow from $L(0) \rightarrow A(0)$.

The ATE estimand is then given by:

$$\Psi^{\text{ATE}} = \sum_{l(0)} \left[\mathbb{E}(Y \mid A = a, l(0)) - \mathbb{E}(Y \mid A = a^*, l(0))\right] \times P(L(0) = l(0)) \tag{27}$$

7.2 Identification assumptions

Identification assumptions for the causal quantities of interest in mediation analyses are presented in the Table 4.

7.3 Estimands

Under the identification assumptions, the estimands of the causal quantities of interest in mediation analyses are presented in Tables 5, 6 and 7.

Table 4: Identification assumptions

Causal quantities	**Randomisation and structural assumptions**	**Positivity assumptions**
ATE	i) No unmeasured confounding between A and Y, given $L(0)$ $(A \perp\!\!\!\perp Y_a \mid L(0))$	1) if $P(l0) \neq 0$, then $\forall a' \in \{a, a^*\}, P(A = a' \mid l(0)) > 0$
2-Way decomposition		
CDE and EE	i) No unmeasured confounding between A and Y, given $L(0)$ $(A \perp\!\!\!\perp Y_{a,m} \mid L(0))$ ii) No unmeasured confounding between M and Y, given $L(0)$ $(M \perp\!\!\!\perp Y_{a,m} \mid A, L(0), L(1))$	1) and 2) $\forall a' \in \{a, a^*\}$, if $P(a', l(1), l(0)) > 0$, then $P(m \mid l(1), a'l(0)) > 0$,
PNDE and TNIE (or TNDE and PNIE)	i) and ii) and iii) No unmeasured confounding between A and M, given $L(0)$ $(A \perp\!\!\!\perp M_a \mid L(0))$ iv) No confounder $L(1)$ of the $M - Y$ relationship is affected by A $(M_{a^*} \perp\!\!\!\perp Y_{am} \mid L(0), L(1))$	1) and 2) If $P(l(1), a, l(0)) > 0$ and $P(m \mid l(1), a^*, l(0)) > 0$, then, $P(m \mid l(1), a, l(0)) > 0$ 3) if $P(l(1) \mid a, l(0)) > 0$, then $P(l(1) \mid a^*, l(0)) > 0$
MRDE and MRIE	i) and ii) and iii)	1) and 2) If $P(l(1), a, l(0)) > 0$ and $P(m \mid l(1), a^*, l(0)) > 0$, then, $P(m \mid l(1), a, l(0)) > 0$ 3) if $P(l(1) \mid a, l(0)) > 0$, then $P(l(1) \mid a^*, l(0)) > 0$
CRDE and CRIE	i) and ii) and iii) and v) No unmeasured confounding between A and $L(1)$, given $L(0)$ $(A \perp\!\!\!\perp L(1)_a \mid L(0))$	1) and 2) If $P(l(1), a, l(0)) > 0$ and $P(m \mid l(1), a^*, l(0)) > 0$, then, $P(m \mid l(1), a, l(0)) > 0$ 3) if $P(l(1) \mid a, l(0)) > 0$, then $P(l(1) \mid a^*, l(0)) > 0$
3-Way and 4-way decomposition		
PNDE, MIE, and PNIE or CDE_0, MIE, RIE, and PNIE	i) and ii) and iii) and iv)	1) and 2) If $P(l(1), a, l(0)) > 0$ and $P(m \mid l(1), a^*, l(0)) > 0$, then, $P(m \mid l(1), a, l(0)) > 0$ 3) If $P(l(1) \mid a, l(0)) > 0$, then $P(l(1) \mid a^*, l(0)) > 0$ 4) $\forall a' \in \{a, a^*\}$, if $P(a', l(0), l(1)) > 0$, then $\forall m' \in \{m, m^*\}, P(m' \mid l(0), a', l(1)) > 0$

Table 5: Estimands corresponding to the causal quantities of interest

Causal quantities	**Estimand**
Average Total Effect (ATE)	$\Psi^{\text{ATE}} = \sum_{l(0)} \left[\mathbb{E}(Y \mid A = a, l(0)) - \mathbb{E}(Y \mid A = a^*, l(0))\right] \times P(L(0) = l(0))$
2-Way decomposition (1)	
Controlled Direct Effect (CDE) in Fig. 11(a)	$\Psi^{\text{CDE}_m} = \sum_{l(0),l(1)} \left[\mathbb{E}(Y \mid m, l(1), a, l(0)) - \mathbb{E}(Y \mid m, l(1), a^*, l(0))\right] \times P(L(0) = l(0), L(1) = l(1))$
Controlled Direct Effect (CDE) in Fig. 11(b)	$\Psi^{\text{CDE}_m} = \sum_{l(0),l(1)} [\mathbb{E}(Y \mid m, l(1), a, l(0)) \times P(L(1) = l(1) \mid a, l(0)) - \mathbb{E}(Y \mid m, l(1), a^*, l(0)) \times P(L(1) = l(1) \mid a^*, l(0))] \times P(L(0) = l(0)$
2-Way decomposition (2)	
Pure Natural Direct Effect (PNDE)	$\Psi^{\text{PNDE}} = \sum_{l(0),l(1),m} \left[\mathbb{E}[Y \mid m, l(1), a, l(0)] - \mathbb{E}[Y \mid m, l(1), a^*, l(0)]\right] \times P(M = m \mid l(1), a^*, l(0)) \times P(L(0) = l(0), L(1) = l(1))$
Total Natural Indirect Effect	$\Psi^{\text{TNIE}} = \sum_{l(0),l(1),m} \mathbb{E}[Y \mid m, l(1), a, l(0)] \times \left[P(M = m \mid l(1), a, l(0)) - P(M = m \mid l(1), a^*, l(0))\right] \times P(L(0) = l(0), L(1) = l(1))$
2-Way decomposition (3)	
Total Natural Direct Effect	$\Psi^{\text{TNDE}} = \sum_{l(0),l(1),m} \left[\mathbb{E}[Y \mid m, l(1), a, l(0)] - \mathbb{E}[Y \mid m, l(1), a^*, l(0)]\right] \times P(M = m \mid l(1), a, l(0)) \times P(L(0) = l(0), L(1) = l(1))$
Pure Natural Indirect Effect (PNIE)	$\Psi^{\text{PNIE}} = \sum_{l(0),l(1),m} \mathbb{E}[Y \mid m, l(1), a^*, l(0)] \times \left[P(M = m \mid l(1), a, l(0)) - P(M = m \mid l(1), a^*, l(0))\right] \times P(L(0) = l(0), L(1) = l(1))$

(table adapted from [VanderWeele, 2014; Wang and Arah, 2015])

Table 6: Estimands corresponding to the causal quantities of interest (continued)

Causal quantities	**Estimand**
2-Way decomposition (4)†	
Marginal Randomised Direct Effect	$\Psi^{\text{MRDE}} = \mathbb{E}(Y_{a,G_{a^*\mid L(0)}}) - \mathbb{E}(Y_{a^*,G_{a^*\mid L(0)}})$
Marginal Randomised Indirect Effect	$\Psi^{\text{MRIE}} = \mathbb{E}(Y_{a,G_{a\mid L(0)}}) - \mathbb{E}(Y_{a,G_{a^*\mid L(0)}})$
Overall Effect	$\Psi^{\text{OE}} = \mathbb{E}(Y_{a,G_{a\mid L(0)}}) - \mathbb{E}(Y_{a^*,G_{a^*\mid L(0)}})$
	where, in Fig. 11(a), $\mathbb{E}(Y_{a,G_{a'\mid L(0)}}) = \sum_{l(0),l(1),m} \mathbb{E}(Y \mid m, l(1), a, l(0))$ $\times[\sum_{l(1)'} P(M = m \mid l(1)', a', l(0))P(L(1) = l(1)')]$ $\times P(L(0) = l(0), L(1) = l(1))$
	where, in Fig. 11(b), $\mathbb{E}(Y_{a,G_{a'\mid L(0)}}) = \sum_{l(0),l(1),m} \mathbb{E}(Y \mid m, l(1), a, l(0))$ $\times[\sum_{l(1)'} P(M = m \mid l(1)', a', l(0))P(L(1) = l(1)' \mid a', l(0))]$ $\times P(L(1) = l(1) \mid a, l(0)) \times P(L(0) = l(0))$
2-Way decomposition (5)	
Conditional Randomised Direct Effect	$\Psi^{\text{MRDE}} = \mathbb{E}(Y_{a,\Gamma_{a^*\mid L(0),L(1)}}) - \mathbb{E}(Y_{a^*,\Gamma_{a^*\mid L(0),L(1)}})$
Conditional Randomised Indirect Effect	$\Psi^{\text{MRIE}} = \mathbb{E}(Y_{a,\Gamma_{a\mid L(0),L(1)}}) - \mathbb{E}(Y_{a,\Gamma_{a^*\mid L(0),L(1)}})$
	where, in Fig. 11(a), $\mathbb{E}(Y_{a,\Gamma_{a'\mid L(0),L(1)}}) = \sum_{l(0),l(1),m} \mathbb{E}(Y \mid m, l(1), a, l(0)) \times P(M = m \mid l(1), a', l(0))$ $\times P(L(0) = l(0), L(1) = l(1))$
	where, in Fig. 11(b), $\mathbb{E}(Y_{a,\Gamma_{a'\mid L(0),L(1)}}) = \sum_{l(0),l(1),m} \mathbb{E}(Y \mid m, l(1), a, l(0)) \times P(M = m \mid l(1), a', l(0))$ $\times P(L(1) = l(1) \mid a, l(0)) \times P(L(0) = l(0))$

† *the sum is equal to the Overall Effect, OE* $=$ MRDE $+$ MRIE

(table adapted from [VanderWeele, 2014; Wang and Arah, 2015])

Table 7: Estimands corresponding to the causal quantities of interest (continued)

Causal quantities	**Estimand**
3-Way decomposition, where $a = 1$, $a^* = 0$ and M is binary	
Pure Natural Direct Effect (PNDE)	$\Psi^{\text{PNDE}} = \sum_{l(0),l(1),m} [\mathbb{E}[Y \mid m, l(1), A = 1, l(0)] - \mathbb{E}[Y \mid m, l(1), A = 0, l(0)]]$ $\times P(M = m \mid l(1), A = 0, l(0)) \times P(L(0) = l(0), L(1) = l(1))$
Mediated Interaction Effect (MIE)	$\Psi^{\text{MIE}} = \sum_{l(0),l(1),m} [\mathbb{E}(Y \mid A = 1, M = 1, l(0), l(1)) - \mathbb{E}(Y \mid A = 1, M = 0, l(0), (l1))$ $-\mathbb{E}(Y \mid A = 0, M = 1, l(0), l(1)) + \mathbb{E}(Y \mid A = 0, M = 0, l(0), l(1))]$ $\times [\mathbb{E}(M \mid A = 1, l(0), l(1)) - \mathbb{E}(M \mid A = 0, l(0), l(1)] \times P(L(0) = l(0), L(1) = l(1))$
Pure Natural Indirect Effect (PNIE)	$\Psi^{\text{PNIE}} = \sum_{l(0),l(1)} [\mathbb{E}(Y \mid M = 1, l(1), A = 0, l(0)) - \mathbb{E}(Y \mid M = 0, l(1), A = 0, l(0))]$ $\times [P(M = 1 \mid l(1), A = 1, l(0)) - P(M = 1 \mid l(1), A = 0, l(0))]$ $\times P(L(0) = l(0), L(1) = l(1))$
4-Way decomposition, where $a = 1$, $a^* = 0$ and M is binary	
Controlled Direct Effect (CDE)	$\Psi^{\text{CDE}_0} = \sum_{l(0),l(1)} [\mathbb{E}(Y \mid M = 0, l(1), A = 1, l(0)) - \mathbb{E}(Y \mid M = 0, l(1), A = 0, l(0))]$ $\times P(L(0) = l(0), L(1) = l(1))$
Mediated Interaction Effect (MIE)	$\Psi^{\text{MIE}} = \sum_{l(0),l(1)} [\mathbb{E}(Y \mid A = 1, M = 1, l(0), l(1)) - \mathbb{E}(Y \mid A = 1, M = 0, l(0), (l1))$ $-\mathbb{E}(Y \mid A = 0, M = 1, l(0), l(1)) + \mathbb{E}(Y \mid A = 0, M = 0, l(0), l(1))]$ $\times [\mathbb{E}(M \mid A = 1, l(0), l(1)) - \mathbb{E}(M \mid A = 0, l(0), l(1)] \times P(L(0) = l(0), L(1) = l(1))$
Reference Interaction Effect (RIE)	$\Psi^{\text{RIE}} = \sum_{l(0),l(1)} [\mathbb{E}(Y \mid A = 1, M = 1, l(0), l(1)) - \mathbb{E}(Y \mid A = 1, M = 0, l(0), (l1))$ $-\mathbb{E}(Y \mid A = 0, M = 1, l(0), l(1)) + \mathbb{E}(Y \mid A = 0, M = 0, l(0), l(1))]$ $\times P(M = 1 \mid l(1), A = 0, l(0)) \times P(L(0) = l(0), L(1) = l(1))$
Pure Natural Indirect Effect (PNIE)	$\Psi^{\text{PNIE}} = \sum_{l(0),l(1)} [\mathbb{E}(Y \mid M = 1, l(1), A = 0, l(0)) - \mathbb{E}(Y \mid M = 0, l(1), A = 0, l(0))]$ $\times [P(M = 1 \mid l(1), A = 1, l(0)) - P(M = 1 \mid l(1), A = 0, l(0))]$ $\times P(L(0) = l(0), L(1) = l(1))$

(table adapted from [VanderWeele, 2014; Wang and Arah, 2015])