Università degli Studi di Cagliari

Dipartimento di Matematica e Informatica
Dottorato di Ricerca in Matematica e Calcolo Scientifico
Ciclo XXVIII

Ph.D. Thesis

Mediation analysis for different types of Causal questions: Effect of Cause and Cause of Effect

S.S.D. SECS-S/01 STATISTICA

Candidate
Rossella Murtas

Supervisor
Prof. Monica Musio

PhD Coordinator
Prof. Giuseppe Rodriguez

Final examination academic year 2014/2015

Abstract

Many statistical analyses aim at a causal explanation of the data. When discussing this topic it is important to specify the exact query we want to talk about. A typical causal question can be categorized in two main classes: questions on the causes of observed effects and questions on the effects of observed causes. In this dissertation we consider both EoC and CoE causal queries from a particular perspective that is Mediation. Mediation Analysis aims to disentangle the pathway between exposure and outcome on a direct effect and an indirect effect arising from the chain exposure-mediator-outcome. In the EoC framework, if the goal is to measure the causal relation between two variables when a third is involved and plays the role of mediator, it is essential to explicitly define several assumptions among variables. However if any of these assumptions is not met, estimates of mediating effects may be affected by bias. This phenomenon, known with the name of *Birth Weight paradox*, has been explained as a consequence of the presence of unmeasured confounding between the mediator and the outcome. In this thesis we discuss these apparent paradoxical results in a real dataset. In addition we suggest useful graphical sensitivity analysis techniques to explain the potential amount of bias capable of producing these paradoxical results. From a CoE perspective, given empirical evidence for the dependence of an outcome variable on an exposure variable, we can typically only provide bounds for the "probability of causation" in the case of an individual who has developed the outcome after being exposed. We show how these bounds can be adapted or improved if further information becomes available. In addition to reviewing existing work on this topic, we provide a new analysis for the case where a mediating variable can be observed. In particular we show how the probability of causation can be bounded in two different cases of partial and complete mediation.

Declaration

I declare that to the best of my knowledge the contents of this thesis are original and my work except where indicated otherwise.

Aknowledgments

Fare un errore diverso ogni giorno non é solo accettabile, é la definizione di progresso. (Robert Brault)

Rossella Murtas gratefully acknowledges INPS for the financial support of her PhD scholarship (Doctor J, Homo Sapiens Sapiens Operational Programme 2012).

Contents

List of Figures

List of Tables

Introduction

Causality is a intuitive concept that we all recognize. For example, is lung cancer caused by smoking? Was contaminated water causing cholera in London in 1854? Can the court infer sex discrimination in a hiring process? However, statisticians have been very careful in formalizing this concept. One reason may be the laborious methods and definitions implemented to study causality. Another explanation may be the complexity to translate real life problems in mathematical notations and formulas. The first step should be to perfectly identify the causal question of interest. This can be categorized in two main classes: questions on the causes of observed effects and questions on the effects of observed causes. This basic distinction, barely familiar in causal inference literature, is fundamental to identify the correct definition of causation. To understand this distinction let us consider the following example. An individual, called Ann, might be subjected to some exposure X, and might develop some outcome Y. For simplicity we will refer to X as a binary decision variable denoting whether or not Ann takes a drug and Y an outcome variable coded as 1 if she dies and 0 if not. We will denote with $X_A = \{0, 1\}$ the value of Ann's exposure and $Y_A = \{0, 1\}$ the value of Ann's outcome.

Questions on the effects of observed causes, named "EoC", are widely known in literature. For example, in medicine, Randomized clinical trials are one of the most rigorous design to assess the effect of a treatment in a population. In the EoC framework we would be interested in asking: "What would happen to Ann if she were to take the drug?" or "What would happen to Ann if she were not to take the drug?". From an individual to a population level, a typical EoC query will be "Is death caused by the drug?". On the other hand, questions on the causes of observed effects "CoE" are quite different and more tricky: they are common in a Court of Law, when we want to asses legal responsibility. For example, let us suppose that Ann has developed the outcome after being exposed, a typical question will be "Knowing that Ann did take the drug and passed away, how likely she would not have died if she had not taken the drug?". In contrast to EoC queries, that are mostly adopted to infer knowledge in the whole population, CoE questions underline a new challenging individual investigation.

In this dissertation we consider both EoC and CoE causal effects invoking

the counterfactual framework. This method examines causality introducing a new type of statistical variables called *Potential Variables*. If X is the exposure and Y the outcome, the potential variable $Y(x)$ will be the hypothetical value of Y that would arise if X was set to x. According to the observed level x, the potential variable can hypothetically incorporate information about what would have happened to the outcome if we would observed a different value of the exposure.

Here we will focus on a particular situation of causal inference that is Mediation Analysis. Mediation aims to assess the extent to which the effect of X on Y is mediated through other pathways and to which this effect is due only by X acting directly on Y. This method aims to disentangle the causality of X on Y on a direct effect and an indirect effect arising from the chain exposure-mediator-outcome. In particular, we will face different problems of Mediation Analysis to both EoC and CoE causal questions.

Mediation Analysis for EoC questions incorporates most of the statistical literature and methods. Usually, it requires the definition of several effects capable of measuring the direct effect of the exposure on the outcome and the indirect effect through the mediator. However, identification of mediation effects requires strong assumptions of no unmeasured confounding in every of these relations: exposure-outcome, exposure-mediator and mediator-outcome. The first two assumptions can easily be verified considering only experimental studies with randomized exposures. The assumption of no unmeasured mediator-outcome confounding can not be easily excluded. Assessing mediation analysis with unmeasured mediator-outcome confounding usually leads to paradoxical results. Hernández-Diaz *et al.* [28] discussed how infants born to smokers have higher risk of both low birth weight (LBW; defined as birth weight <2500g) and infant mortality than infants born to non-smokers, but in the LBW stratum maternal smoking appears not to be harmful for infant mortality relatively to non-smoking. This phenomenon, known with the name of *Birth Weight paradox*, has been explained as a consequence of the presence of unmeasured confounding between the mediator birth weight and the outcome infant mortality. In this thesis we discuss these apparent paradoxical results studying the effect of high parity on wheezing or asthma mediated by low birth weight. In particular, we consider two different cases of a rare and a regular outcome. After partitioning the causal effect into a direct and indirect effects, we examine different techniques to test the sensitivity of these paradoxical results. In addition we suggest useful graphical sensitivity analysis techniques to explain the potential amount of bias capable of producing these paradoxical results. Furthermore, we implemented different Stata handwriting for this sensitivity analysis, that will be collected in a final statistical package.

In contrast, mediation analysis in the CoE framework is a new and interesting challenge in statistical theory. Definition of CoE causal effects is completely

different from EoC questions. It invokes the *Probability of Causation* as given by Dawid (2011) in [16] and also named by Pearl as *Probability of Necessity* [48]. Given that Ann took the drug and passed away, the Probability of Causation in Ann's case is defined as:

$$\mathrm{PC}_A = \mathrm{P}_A(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1)$$

where P_A denotes the probability distribution over attributes of Ann. The probability of causation is a fundamental measure in several fields such as epidemiology and in a court of law. For example, let us suppose that Ann's children filled a criminal lawsuit against a pharmaceutical manufacturer claiming that the drug was the *cause* of her death. Using data on similar individuals, we wish to evaluate, for this case, the probability that the outcome was in fact caused by the exposure. Whenever the probability of causation exceeds 50%, in a civil court, this is considered as preponderance of evidence because causation is "more probable than not". This is also known with the name of "balance of probability" as the general gold standard in most civil cases. However, nowadays, causality without ad hoc mathematical definitions and rules are widely and wrongly used in many courthouse. Given the important implications of the probability of causation, it is clear that we have to focus on studying methods capable of producing a more precise estimate. From a statistical point of view, this definition underlines a bivariate structure between two potential variables associated at the same subject. However, only one of them will be observable while the other will be counterfactual. For this reason, the PC_A is not completely identifiable. We can at least provide useful information on the values to which it must lie. Under appropriate assumptions, these bounds can be tightened if we can make other observations (e.g., on nonexperimental cases), or measure additional variables (e.g., covariates) and even in the case that unobserved variable confounds the exposure-outcome relationship. In addition we propose a novel approach to bound the probability of causation in mediation analysis. In particular, we focus on two different mechanisms of complete and partial mediation. In the first the exposure is supposed to act on the outcome only through the mediator, *i.e.* no direct effect is present. In the latter, both direct and indirect effects are considered. We will see that, considering a complete mediator, we always obtain smaller bounds for PC_A compared with the simple analysis of an exposure acting directly on the outcome. For the case of partial mediation, usual assumptions of no confounding in any relationships will not be enough to obtain smaller bounds. Here we will introduce a further hypothesis on the bivariate distribution of the counterfactual outcome and mediator that, in addition to univariate conditions, will produce new and interesting results.

In § 1 we review the major literature of Causality from an interventional § 1.2, a counterfactual § 1.3 and a decision theory § 1.6 perspectives. In § 2 we introduce the EoC and CoE frameworks and the principal differences between these

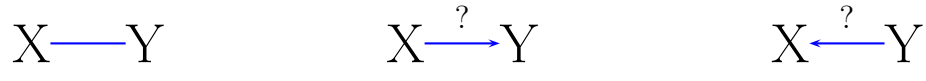
two methods. In § 3 we illustrate several methods to study Mediation Analysis in the EoC framework in § 4 we apply some of these approaches to study the effect of high parity on two different outcomes of wheezing or asthma mediated by birth weight. In particular we will investigate the Birth Weight Paradox using data from an Italian birth cohort called Ninfea. In addition we will introduce several graphical techniques to test the sensitivity of the mediation effects in relation to the paradox. In § 5 we review the literature of CoE causal questions and we introduce new methods to study mediation in the CoE framework. Comparisons between the methods presented in § 5 are described both theoretically and by means of some examples. Finally, § 6 summarizes the findings of our work and describe our further aims while in Appendix A we include a selection of the handwriting implemented in Stata that will be collected in a statistical package.

Chapter 1

Causality

One of the most important topics in statistics is the study of the relationship between an outcome Y and an independent variable X. When we study this relation for continuous variables, we usually express it in terms of correlation. We say that X and Y are perfect positively (negatively) correlated if ρ is equal to one (minus one) and independent if equal to zero. In the main, most of the statistical literature is dedicated to finding the best model capable of describing the data and of easily relating X and Y. These two frameworks collapse when we consider linear regression models, *i.e.* the regression coefficient of X on Y can be simply obtained by multiplying the correlation by one over the standard deviations product.

But what happens when X causes Y? Is correlation a meaningful measure? It does not take into account the nature of the connection [82]. For example, does smoking cause lung cancer? How much of this relationship is due to other explanations or factors? Did a drug cause deaths in a specific population?
Effectively, a strong correlation between X and Y ensures a connection but does not specify the direction. Moreover, the correlation can be due to factors other than X and Y alone. One way to evaluate when such correlation is causal is to consider all possible mechanisms that play a role in this relationship. The diagrams in Figure 1.1 describe three possible mechanisms, all capable of explaining a correlation between X and Y. The question marks in Figure 1.1b and 1.1c encode the uncertainty of inferring causation from correlation alone, without any knowledge of the underlying mechanisms.



(a) (b) (c)

Figure 1.1: Simple Pathways where (a) X and Y are associated, (b) X causes Y and (c) Y causes X.

Understanding these mechanisms is crucial when we talk about the effect of an exposure (X) on a disease (Y). In this setting, the well known Epidemiological measures of Odds Ratio (OR), Relative risk (RR) and Hazard Ratio (HR) can be more useful than mere correlation as they are able to quantify the exposure-disease association.

However, most of the studies in health, social and the behavioral sciences are not associational but causal in nature [53]. For this reason we need to define a proper causal language.

1.1 Causality vs association

Intuitively, one can define causality between two continuous variables X and Y when a change in X produces a change in Y. The usual measure of association, given by

$$\rho = \frac{cov(X, Y)}{dev(X) \cdot dev(Y)}, \tag{1.1}$$

is not sensitive in assessing causality. The correlation is a commutative measure and situations such as 1.1b and 1.1c produce the same result. However, under some assumptions, we will discuss how association can be viewed under the causal inference lens.

Beyond the intuitive definition, several authors provide mathematical and formal definitions of causality. Articulating the condition under which one variable is considered relevant to another, we can create a diagram capable of illustrating the relational associations between exposure and outcome. In Epidemiology we usually depict *Directed Acyclic Graph* as the diagrams shown in Figure 1.1b and 1.1c. The "directed' element is meant to infer causation and "cyclic' to avoid non-biological loops since no single variable can cause itself. Pearl (2014) in [52] defines formulas and rules to convert causal assumptions into conditional independencies implied in a DAG. One of the first attempts to formalise causality is given by the geneticist Sewall Wright (1921) in [82]. With his *path analysis*, Wright was able to quantify causal effects linking a regression coefficient with every path in the diagram. The *Structural equation model* approach (SEM) generalizes path analysis defining a statistical model for every endogenous variable in a DAG. However, the majority of the literature and methods of SEM are restricted to continuous outcomes while *nonparametric* models can be used in order to avoid having to specify a precise functional form. A completely different approach based on *Decision Theory* will be discuss in § 1.6. The last and somewhat more tricky definition is a result of the *counterfactual* semantic approach attributed to Lewis (1973) in [36] and mathematically formulated by Rubin (1974) in [65]. When we describe causation as changes in Y caused by changes in X, we should measure, for every instance,

how observing or not the exposure will lead to observing or not the outcome. The term counterfactual means that these outcomes represent situations that may not actually occur (they may be counter to the fact situation). This problem is defined by Holland (1986) in [30] as "the fundamental problem of causal inference": No man ever steps in the same river twice (Heraclitus 535-475 b.C.).
In a simple framework all these definitions should produce the same results. In § 1.2 we will discuss the DAG approach in greater details while in § 1.3 we will discuss causality in terms of counterfactual variables. In § 1.6 we will introduce the Decision theory approach to causal inference. Path Analysis will be described in § 3.1.1 as a particular approach to *Mediation analysis*.

1.2 Directed Acyclic Graph

As articulating in the previous section, when we talk about causation, introducing a graph like Figure 1.1 is mandatory. A graph is a network composed of a set of *links* each of them connecting two *nodes*. Nowadays graph theory is becoming increasingly significant in a variety of fields such as engineering (logistics, networks), computer science (algorithms, decision theory) and medicine (genomics, pathways of an infection, correlation between drugs). In statistics, nodes usually represent variables and links represent our knowledge of an ideal relationship between them. Every link in a graph can be direct (for which the link will be depicted by an arrow), undirect (no arrowhead) or bidirectional (double ended arrowheads). A graph is called *directed* if all links are directed. A graph that does not contain cycles is called *acyclic*.

In medicine the research interest usually lies with a particular *exposure* and a particular *outcome*. For example:

- Knowing that Ann took a drug (exposure) and passed away (outcome), how likely is that she would not have died if she had not taken the drug?
- Does a particular diet (exposure) influence the relapse of breast cancer (outcome)?
- Does the measles-mumps-rubella (MMR) vaccine (exposure) cause autism (outcome)?

Such types of relationships are always time dependent, *i.e.* the drug is taken before death occurred *etc.*. For this reason we will consider only *directed acyclic graph* also known as DAG.

Let us consider the following example [51]. Figure 1.2 describes a simple DAG where X_1 represents the season, X_2 whether rain falls, X_3 whether the

sprinkler is on, X_4 whether the pavement gets wet and X_5 whether the pavement would be slippery. All these variables are binary except the season X_1 which can have 4 possible values. In this DAG, the variable X_5 has not child, has one parents (X_4) and three ancestors (X_1, X_2 and X_3).

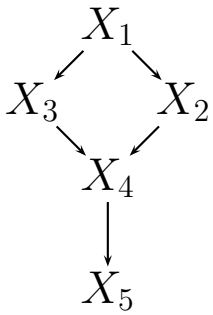


Figure 1.2: Example of a Directed Acyclic Graph representing dependencies between five variables

A DAG such as the one shown in Figure 1.2 is called a *Bayesian Network* by Pearl (1985) in [47]. Mostly because of the subjective nature of the information encoded within it and for the connection with Bayes's Theorem. The role of Bayesian Networks in statistical modelling is straightforward. They provide a logical diagram and facilitate an optimal representation of joint distributions. To clarify the last point, suppose we have n binary variables $X_1, \ldots, X_n$ with joint distribution $P(X_1 = x_1, \ldots, X_n = x_n)$ that we will denote $P(x_1, \ldots, x_n)$. If we need to describe $P(x_1, \ldots, x_n)$ for every $x_i = \{0, 1\}$, we will produce 2^n numbers. We can always decompose P as the product

$$P(x_1, \ldots, x_n) = P(x_n|x_1, \ldots, x_{n-1}) \cdots P(x_2|x_1) \cdot P(x_1) \tag{1.2}$$

for every order of $X_1, \ldots, X_n$. For example in the DAG in Figure 1.2, considering the natural order of the variables, we will have

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_5|x_1, x_2, x_3, x_4) \cdot P(x_4|x_1, x_2, x_3) \cdot P(x_3|x_1, x_2) \cdot P(x_2|x_1) \cdot P(x_1).$$

Whenever a variable X_j is independent to some predecessors, (1.2) will lead to a more economical joint probability function. These independent relationship can easily be read in a directed acyclic graph (the mathematical rules are given below). For example from the graph in Figure 1.2 we can see that whenever the pavement is wet, it will be slippery independently by the season, the rain and the irrigation system. We will use Dawid's notation $\perp\!\!\!\perp$ to denote conditional independencies [14]. In the seasonal example, the independence of X_5 from all other variables once we know X_4 will be represented as $X_5 \perp\!\!\!\perp X_1|X_4$, $X_5 \perp\!\!\!\perp X_2|X_4$ and $X_5 \perp\!\!\!\perp X_3|X_4$.

The joint probability distribution will then become

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_5|x_4) \cdot P(x_4|x_2, x_3) \cdot P(x_3|x_1) \cdot P(x_2|x_1) \cdot P(x_1). \quad (1.3)$$

As we can see from (1.3) every variable in a Dag depends only on its parents, known as *Markovian parents*. Mathematically, if we have n discrete variables $X_1, \ldots, X_n$, with joint distribution $P(x_1, \ldots, x_n)$, we can always decompose P as the product

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|pa_i)$$

where pa_i denotes the parents of X_i. An elegant and feasible definition to test graphical independence is the *d-separation* criterion [51].

Definition 1.2.1 (d-Separation) *A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if*

1. *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or*

2. *p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z*

A set Z is said to d-separate X form Y if and only if Z blocks every path from a node in X to a node in Y and then $X \perp\!\!\!\perp Y|Z$.

In Figure 1.2 rain and slippery are connected by a chain ($X_2 \to X_4 \to X_5$), if we know it is going to rain the pavement will be slippery, hence X_2 and X_5 are associated. But if we know that the pavement is wet we no longer need to know whether it is raining. Rain and slippery will become independent because wetness blocks the pathway ($X_2 \perp\!\!\!\perp X_5|X_4$). On the other hand sprinkler and rain are dependent because they are both children of the season. But if we know the season they will not be correlated anymore ($X_2 \perp\!\!\!\perp X_3|X_5$).
As a consequence of Definition 1.2.1, we call *open* a path that contains a confounder or a mediator. It becomes blocked after we condition on them. A path is said blocked if it contains a collider and it said to be opened if we condition on it. Variables located in open paths are correlated.

In subsection 1.2 we will discuss the problem arising from conditioning on a collider.

Another fundamental concept in graph theory is *intervention*. In the example of Figure 1.2 we could consider what would have happened to the pavement if we **made sure** that the sprinkler was off. This causal question involves a certain

intervention on a variable in the model. It captures the change in the system implied by setting the sprinkler off. In fact, we are testing the sensibility of the system after intervening on X_3. We further assume that the change is local, affecting only its descendants. The graph can be adapted simply by applying this intervention to the system, *i.e.* deleting all arrows pointing towards X_3.

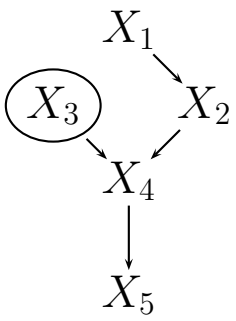


Figure 1.3: General Directed Acyclic Graph representing dependencies between five variables after intervention on X_3

The act of intervening on X_3 will be described by the notation $do(X_3 = off)$ that means setting the sprinkler off. The resulting joint probability distribution will be

$$P_{do(X_3=off)}(x_1, x_2, x_4, x_5) = P(x_5|x_4) \cdot P(x_4|x_2, X_3 = off) \cdot P(x_2|x_1) \cdot P(x_1). \quad (1.4)$$

As noted by Pearl (2009) in [51] the action $do(X_3 = off)$ and the observation $X_3 = off$ are different. In the first we are conditioning on $X_3 = off$ in a graph obtained mutilating the arrow from X_1 to X_3 while the second means simply observing $X_3 = off$. This is the main difference between prediction type causal queries and intervention type, *i.e.* observing and intervening.

Underling assumptions in a DAG

A DAG such as the one shown in Figure 1.2 underlines some relational assumptions between variables. From a DAG we are able not only to deduce conditional assumptions but also to derive knowledge on the absence of causality. In fact, an arrow from a variable to another means that the first *may* cause the second. The absence of an arrow reflects our knowledge of no known association.

Collider Bias

There are different methods to test whether a variable is a risk factor. The most naive is to include this variable in a model for the outcome. In this section we will discuss why, if we are assessing causality, this method will be biased. Definition 1.2.1 can in fact be used as a tool to decide whether or not conditioning on a variable.

One of the simplest example arises in the presence of a collider such as Figure 1.4a.

(a) (b)

Figure 1.4: DAG illustrating the collider bias problem, which arises after conditioning on a collider

As previously discussed, the absence of an arrow between X and Y represents our knowledge of no association (1.4a). From Definition 1.2.1, the path $X \rightarrow Z \leftarrow Y$ is blocked since it contains an inverted fork. Adjusting for Z (action represented by a box around Z) will open this pathway creating a spurious association between X and Y. The dashed line in Figure 1.4b represents this false association. In the example by Pearl, if we know that the pavement is wet (or slippery) we have only two possible explanations; the sprinkler is on or it is raining. Refuting one of them increases the probability of the other (see Figure 1.5). This is usually called *collider bias* and its consequences will be discussed in detail in the following chapters.

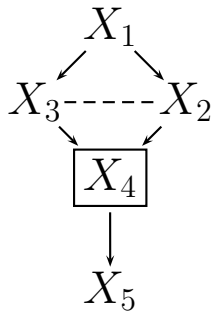


Figure 1.5: DAG representing dependencies between five variables after conditioning on X_4

It is important to note that conditioning on X_5 will produce the same effect as conditioning on X_4 because the latter is its descendant. This distinction is fundamental; assessing causality without constructing ad hoc graphs will invalidate the analysis.

1.2.1 Causal effects: computation and identifiability

Generalizing (1.3) to a set of n discrete variables $X_1, \ldots, X_n$, the intervention on X_i will decompose the joint probability distribution as the product

$$P(x_1, \ldots, x_n | do(X_i = x_i)) = \prod_{j \neq i} P(x_j | pa_j). \tag{1.5}$$

Multiplying and dividing (1.5) by $P(x_i|pa_i)$ we simply get a conditional probability

$$\begin{aligned} P(x_1, \ldots, x_n | do(X_i = x_i)) &= \frac{P(x_1, \ldots, x_n)}{P(x_i|pa_i)} \\ &= P(x_1, \ldots, x_n | X_i = x_i, pa_i) P(pa_i). \end{aligned} \tag{1.6}$$

The joint post-intervention distribution will be the product of the (conditional) pre-intervention distribution and the distribution of the parents of X_i not affected by intervention. Summing equation (1.6) over $X_1, \ldots, X_{n-1}$ and assuming $Y = X_n$ will lead back to the causal effect of X on Y defined intuitively in § 1.1

$$P(Y = y | do(X = x)) = \sum_{pa_x} P(Y = y | X = x, pa_x) \cdot P(pa_x). \tag{1.7}$$

Equation (1.7) produces an immediate algorithm to calculate the causal effect of X on Y. Given a causal diagram in which **all parents of X are observable** we can estimate the causal effect of X on Y from nonexperimental observations. Problems occur when not all parents of X are observable. In the following sections we will introduce two graphical tests to determine a sufficient set of variables capable of estimating $P(Y = y|do(X = x))$.

1.2.2 Back-door Criterion

In the previous section we proved that by measuring all parents of one exposure we are able to compute causal effect form data. However, as mentioned in subsection 1.2, we have to be careful about what adjust for. In this section we will examine a well known method that can provide a sufficient set of variables capable of identifying the causal effect of X on Y.

Definition 1.2.2 (Back-door Criterion [51]) *Given a DAG G, a set of variable Z satisfies the back-door criterion relative to (X, Y) in G if Z satisfies the following conditions*

- *no node in Z is a descendant of X;*
- *Z d-separates every path between X and Y that starts with an arrow pointing into X.*

The term "back-door' hails from the second condition, *i.e.* only pathways starting with an arrow pointing into X are considered. For example, in Figure 1.2, $Z = \{X_2\}$ satisfies the back-door criterion relative to (X_3, X_5) while $Z = \{X_4\}$ does not. This is perhaps consistent with the collider bias rule.

Theorem 1.2.1 *If a set Z satisfies the back-door criterion, the causal effect of X on Y is identifiable and is given by*

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z) \cdot P(Z = z)$$

The proof of Theorem 1.2.1 encodes the idea that if Z satisfies the back-door criterion relative to (X, Y), the intervention has the same effect of conditioning on $X = x$.

Then if Z satisfies Theorem 1.2.1, conditioning on Z will block the back-door pathway from X to Y. Open paths transmit association. Let us consider Figure 1.6a where X potentially causes Y and Z confounds the exposure-outcome relation. If we do not adjust for Z, the causal effect will be biased by the association between X and Y through Z. Blocking the pathway $X \leftarrow Z \rightarrow Y$, the remaining effect is completely causal. On the other hand, Figure 1.6b represents a situation where X and Y are associated but the exposure does not cause the outcome. If we do not adjust for Z, we would interpret this association as causation.

(a) (b)

Figure 1.6: (a) DAG illustrating associational and causal pathway between three variables (b) DAG illustrating only associational pathway between three variables

1.2.3 Front-door Criterion

The first condition in Definition 1.2.2 precludes situations in which back-door pathways are not feasible. This is perhaps one of the purpose of this dissertation.

Definition 1.2.3 (Front-door Criterion [51]) *Given a DAG G, a set of variable Z satisfies the front-door criterion relative to (X, Y) in G, if Z satisfies the following conditions*

- *Z intercepts all direct path from X to Y;*

- *there is no back-door path from X to Z;*
- *all back-door paths from Z to Y are blocked by X.*

Thus in Figure 1.2, $Z = \{X_4\}$ satisfies the front-door criterion for (X_1, X_5) while $Z = \{X_3\}$ or $Z = \{X_2\}$ does not.

Theorem 1.2.2 *If a set Z satisfies the front-door criterion and if $P(X = x, Z = z) > 0$, the causal effect of X on Y is identifiable and is given by*

$$P(Y = y|do(X = x)) = \sum_z P(Z = z|X = x) \sum_{x'} P(Y = y|X = x', Z = z) \cdot P(X = x').$$

In the next section we will introduce a combined method between back and front-door criterion.

From Definition 1.2.2 and Definition 1.2.3 follow immediately that, in order to produce unbiased causal effects, backdoor paths have to be blocked, while frontdoor paths have to be opened.

1.2.4 *do*-Calculus

Pearl (2009) in [51] defines a set of rules called *do*-Calculus, which are capable of identifying the causal effect from a graph G. Let us denote with X, Y, Z and W a set of four disjoined nodes in G. With $G_{\overline{X}}$ we will denote the graph obtained from G intervening on X that is, a graph where all arrows pointing to X are deleted. With $G_{\underline{X}}$ we will denote a graph obtained from G deleting all arrows coming out from X. For simplicity we will denote $P(Y = y)$ as $P(y)$ and with $P(y|\hat{x})$ the effect of the intervention $do(X = x)$.

Theorem 1.2.3 *For any disjoint set of variables X, Y, Z and W we have the following rules*

Rule 1 (Insertion/deletion of observations)

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$$

Rule 2 (Action/observation exchange)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$$

Rule 3 (Insertion/deletion of actions)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z(W)}}}$$

where $Z(W)$ is a set of nodes Z that are not ancestors of any node W in $G_{\overline{X}}$

Corollary 1.2.1 (Pearl 2000) *A causal effect* $q = P(y_1, \ldots, y_k|\hat{x}_1, \ldots, \hat{x}_k)$ *is identifiable in a model characterized by a graph* G *if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 1.2.3, that reduces* q *into a standard (*i.e. *hat-free) probability expression involving observed quantities.*

For example, let us consider the simplest mechanism of mediation where a graph G contains a chain. As described by 1.7a, the independent variable affects a third one, called *Mediator*, which then affects the outcome Y. From (1.7), given there is no back-door path between X and Y, we can simply equate $P(y|\hat{x}) = P(y|x)$. On the other hand, considering Theorem 1.2.3 we have

$$\begin{aligned} P(y|\hat{x}) &= \sum_m P(y, m|\hat{x}) = \sum_m P(y|\hat{x}, m) \cdot P(m|\hat{x}) \\ &= \sum_m P(y|\hat{x}, m) \cdot P(m|x) \text{ if } (M \perp\!\!\!\perp X)_{G_{\underline{X}}} \\ &= \sum_m P(y|x, m) \cdot P(m|x) \text{ if } (Y \perp\!\!\!\perp X|M)_{G_{\underline{X}}} \\ &= \sum_m P(y|x, m) \cdot P(m|x) = P(y|x). \end{aligned} \tag{1.8}$$

Then the causal effect of X on Y in Figure 1.7 is fully identifiable and is given by (1.8).

(a) (b)

Figure 1.7: (a) DAG G illustrating a Mediation Mechanism (b) DAG obtained by mutilating G by all arrows coming out from X

1.3 Counterfactuals

The intuitive idea beyond causation is represented by changes in the outcome due to changes in the exposure. To properly measure causality we should then compare what would have happen to the outcome for different settings of the exposure. The problem is that we can observe, for each subject, only one results from the exposure. In this section we introduce the counterfactual framework first defined by Neyman (1923) in his master thesis (see [72] revisited) and then extended by

Rubin (1974) in [65].

Let us consider an individual, called Ann, that might be subjected to some exposure X, for example, a drug, and might develop some outcome Y, for example, mortality. We will denote with $X_A = \{0, 1\}$ the value of Ann's exposure and $Y_A = \{0, 1\}$ the value of Ann's outcome. We are interesting in knowing what would happen to Ann if she were to take the drug or what would happen to Ann if she were not to take it. This can be achieved defining a new type of variable called potential.

Definition 1.3.1 (Potential Variable) *Let us consider X the exposure of interest and Y the outcome, $Y(x)$ is the potential value that Y would take if X had been set to x.*

Then $Y_A(0)$ is the outcome we could expect if she had not taken the drug and $Y_A(1)$ the outcome we could expect if she had taken it. The potential outcome $Y(x)$ is treated as an ordinary random variable with a distribution and consistent with the usual axioms of probability and independence. It is connected to the real outcome by a necessary condition for meaningful causal inference

Consistency condition
The potential value $Y(x)$ must be equal to the real outcome when the exposure is observed. Formally the consistency rule states that $Y(x) = Y$ if $X = x$. In terms of probability we have $P(Y(x) = y|X = x) = P(Y = y|X = x)$. For binary exposure this condition states that $Y_{obs} = X \cdot Y(1) + (1 - X)Y(0)$.

Let us suppose further that, unfortunately, Ann took a drug and passed away. With this additional information, and for the consistency condition, $Y_A(1)$ is observable and is equal to one while $Y_A(0)$ is unknown and counterfactual.

In general, for every subject i in a population where we collected information on a binary exposure X and a binary outcome Y, we can construct the Table 1.1.

For every subject i, we can define two potential outcomes $\{Y_i(0), Y_i(1)\}$, one observable and one counterfactual (represented by a question mark). In an ideal world, knowing that the first subject had been exposed to X and had developed the outcome, we could answer the question: "What would have happened to him if he had not been exposed?" but in the real world we can only guess it from the data. In this section we will introduce situations in which observed data can be used to answer this query.

Regarding Ann, if $Y_A(0) = 0$, we could conclude that Ann's disappearance was caused by the drug because she would not have died if she had not taken it. On

	X_i	Y_i	$Y_i(0)$	$Y_i(1)$
1	1	1	?	1
2	0	1	1	?
3	0	0	0	?
...	...	...	...	...
n	1	0	?	0

Table 1.1: Example of a realization of the vector of binary variables $(X, Y, Y(0), Y(1))$ in a population with n subjects. Question marks correspond to counterfactual values

the other hand, if $Y_A(0) = 1$, she would have died anyway. If for example, different to the fact situation, $Y_A(1) = 0$ and $Y_A(0) = 0$ we could conclude that the drug has not effect on Ann. This will lead immediately to the following definition.

In the counterfactual framework we say that there is an individual causal effect of X on Y when the potential value of Y changes with X.

Definition 1.3.2 (Individual Causal Effect) *The individual causal effect of X on Y is defined as the difference between the outcome of the unit i under level of the exposure x, $Y_i(x)$, and the outcome of the same unit under a different level of the exposure $Y_i(\tilde{x})$*

$$Y_i(x) - Y_i(\tilde{x}). \tag{1.9}$$

If X is binary $ICE = Y_i(1) - Y_i(0)$.

However, the ICE cannot be completely measured for the same subject i as we saw for Ann's case. More interesting and feasible is the average causal effect in the whole population.

Definition 1.3.3 (Average Causal Effect) *The Average Causal Effect of X on Y is:*

$$ACE(x, \tilde{x}) = E[Y(x)] - E[Y(\tilde{x})] \tag{1.10}$$

When X is binary $ACE = P(Y(1) = 1) - P(Y(0) = 0)$. Hereafter we will consider only binary variables.

But under which conditions we can infer the above quantities from the data? And which ones will be the best substitutes? The best candidate is a measure of the exposure-outcome association; the expected value of Y among those who actually had the exposure x, $P(Y = y|X = x)$. This raises the question: when does association reflect causation?

1.4 Experimental Studies, Nonexperimental studies and Exchangeability

As we discussed in the previous sections, a DAG implicitly underlines assumptions. For example, in Figure 1.2, the arrow $X_1 \to X_2$ encodes the idea that the former *may* cause the latter. On the other hand, the absence of an arrow from X_1 to X_5 reflects a prior (*subjective*) knowledge of no direct association.

If we believe in these assumptions, DAG such as Figure 1.2 or Figure 1.7 represents fully observable pathways in which we can infer causation from nonexperimental observations. However, these assumptions cannot generally be tested in nonexperimental studies.
In the light of these considerations, experimental studies (such as *Clinical Trials*), are considered the best method to infer causation. They are usually designed to test whether a treatment affects an outcome. The main idea is randomization where participants are randomly assigned to treatments. This intervention allows the deletion of all arrows pointing to the exposure and hence, theoretically avoid having to adjust for exposure-outcome confounding. In ideal randomised experiments (if there are no issues of measurement error or loss to follow up and if they are double bind) when studying the effect of a treatment in a population, association reflects causation [26]. Furthermore, patients exposed and unexposed are actually exchangeable. Exchangeability means that the effect of X on Y does not differ with respect to the distribution of exposure-outcome confounding. A perfect randomization ensures exchangeability.

In the counterfactual framework, the exchangeability condition states that $\{Y(0), Y(1)\} \perp\!\!\!\perp X$. The potential value that Y would take under different levels of the exposure does not depend on the observed treatment. On the other hand, if the assignment is random, changes in Y are due only to changes in X and not from other causes. This is why, in experimental studies, association is causation.
If the exchangeability condition holds then $P(Y(x) = y) \equiv P(Y = y|X = x)$, *i.e.* the counterfactual probability under exposure level x equals the observed probability among those who actually received treatment x.

Despite all this qualities, nonexperimental studies are usually more common as they are less expensive. Furthermore, randomized treatments are not always ethical or feasible. Consider for example exposures such as heart transplantation, birth weight or HIV status. For these reasons, one of the major goals of Causal Inference is to define situations where nonexperimental studies can be used to infer causation.
In this dissertation we will focus on deriving causal inference from nonexperimental design. This is the case where the exchangeability condition does not hold. However, we can still measure causal effects if, after taken the confounder in

consideration, patients are exchangeable in every strata of it.

1.4.1 Conditional Exchangeability

Let Z be an exposure-outcome confounder such as in Figure 1.6a. Two patients are said to be *conditionally exchangeable* in respect of treatment X if, within the strata of Z, they are exchangeable. Restating this proposition in Dawid's counterfactual notation we get $Y(x) \perp\!\!\!\perp X|Z$.

If a variable Z satisfies the conditional exchangeability condition then:

$$\begin{aligned} P(Y(x) = y) &= \sum_c P(Y(x) = y|Z = z)P(Z = z) \\ &= \sum_c P(Y(x) = y|X = x, Z = z)P(Z = z) \quad Y(x) \perp\!\!\!\perp X|Z \\ &= \sum_c P(Y = y|X = x, Z = z)P(Z = z) \quad \textit{Consistency} \end{aligned} \tag{1.11}$$

Since a confounder Z satisfies the back-door criterion, the counterfactual causal effect embodied by equation (1.11) correspond exactly to Theorem 1.2.1.
Conditions such as exchangeability and conditional exchangeability in terms of counterfactual cannot easily be read from complicated DAGs. Various authors illustrated methods capable of representing potential outcomes in a graph [51] [68] [57]. In section subsection 3.2.1 we will examine one of these methods.

1.5 G-methods

If a variable Z confounds the exposure-outcome relationship, the effect of X on Y differs with respect to the distribution of Z. If Z is sufficient to adjust for confounding, stratified measure such as Equation (1.11), called the *standardization formula*, are preferable. In a simple case, defining a parametric model for every endogenous variable in Equation (1.11), will lead to a correct estimation of the causal effect. These estimates are correct if models are correctly specified and if Z is sufficient to adjust for confounding. Robins (1986) in [60] provides a generalization of the standardization formula for time-dependent variables (exposures, confounders and outcomes) called *g-formula*. An alternative method is Inverse Probability weighting (IPW) [31] which creates a re-weighted population in which exposed and unexposed are then exchangeable given Z. Hernan (2010) and Robins (1986) in [27, 60], called this two approaches, IPW and g-formula, the *g-methods*. The term "g" is referred to "generalized" because, unlike regression analysis, they usually

address many situations including time-varying variables (exposures, confounders and outcomes).

A completely different approach is stratification which is basically a regression based method. It permits to estimate the causal effect of X on Y specifying a regression model for the outcome on the exposure and covariates Z. If the set Z is a sufficient set of confounding variables, the regression coefficient of X will estimate the causal effect of the exposure on the outcome.

1.6 Decision Theory

Causal inference in the counterfactual framework requires the definition of Potential Variables. However, as we discussed in § 1.3, these outcomes are not completely identifiable given that we can observe only one of them for each subject. Starting from P. Dawid [15], many statisticians reasonable believe that proper causal inference should not depend on unobservable quantities and un-testable assumptions. Dawid (2014) in [17] addresses the question of Causality as a Decision problem. He considered, for example, X a decision variable denoting whether or not to take an aspirin and Y the log-time it takes for the headache disappearance. Let us consider a new subject $u_\emptyset$ which suffers from headache and has to choose whether or not to take the drug. Although X is a decision variable, we can define the probability of Y given the decision on X, *i.e.* $P_0 = P(Y = y|X = 0)$ and $P_1 = P(Y = y|X = 1)$. These distributions are all that is needed to answer the causal query. Only comparing P_0 and P_1 we are able to choose the decision that will improve my outcome. Different comparisons will take into account different aims and perspectives. For example, we can simply take the difference $P_1 - P_0$. If we assume these distributions to follow two normal probability density functions, we can even compare the means or the variances of these two distributions. These comparisons can all be considered as the causal effect arising after the choice $X = 1$ rather than $X = 0$.

This method can be even formalized defining a *Loss function* $L(Y)$ such that $L(y)$ will be the loss that this new subject $u_\emptyset$ will suffer if his headache lasts y minutes.

The decision tree associated to this situation is described by Figure 1.8. At the node $v_\emptyset$, the subject $u_\emptyset$ can choose between to take the drug (upper branch) or not to take the drug (lower branch). At node v_1, the outcome Y will be distributed according to P_1 while at v_0, the outcome Y will be distributed according to P_0. For every possible value y of Y and nodes v_0 and v_1, we will have a node $L(y)$ corresponding to the loss associated at the different decision. Given each nodes v_0 and v_1, we can calculate the expected loss $E_{P_0}\{L(Y)\}$ and $E_{P_1}\{L(Y)\}$. The idea beyond this method requires that, at the decision note $v_\emptyset$, this new subject will

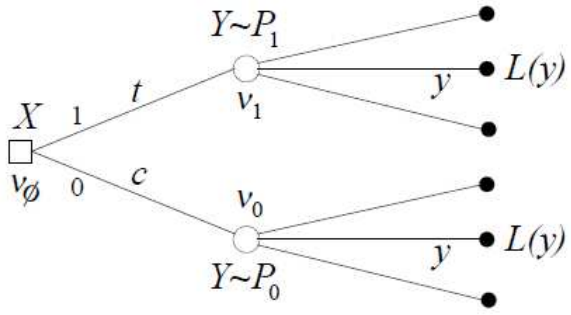


Figure 1.8: Decision Tree

choose the treatment leading to the smaller expected loss. Whatever loss function is used, this method only invoke the distributions P_0 or P_1 but any counterfactual entities.

Chapter 2

Different Type of Causal Questions

Establishing if an exposure potentially causes an outcome is becoming more and more important in real life situations. Nowadays, most of epidemiological, econometrics and psychological problems can be viewed into the causal inference lens. However, behind the realistic need to confirm causes, the first and basic issue consists in formulate a mathematical definition for causality. In order to solve the problem the researcher has to perfectly identify the causal question of interest. Let us consider again the following causal questions introduced in § 1.2:

Example 2.0.1 *Knowing that Ann took a drug (exposure) and passed away (outcome), how likely is that she would not have died if she had not taken the drug?*

Example 2.0.2 *Does a particular diet (exposure) influence the relapse of breast cancer (outcome)?*

Example 2.0.3 *Does the measles-mumps-rubella (MMR) vaccine (exposure) cause autism (outcome)?*

A general causal query can be categorized in one of the following main classes: question on the *causes of observed effects* and question on the *effects of observed causes*. This basic distinction, barely familiar in causal inference literature, is fundamental to identify the correct definition of causation.

Let us consider the situation described in the Example 2.0.1. An individual, called Ann, had been subjected to some exposure X, and has developed some outcome Y. For simplicity we will refer to X as a binary decision variable denoting whether or not Ann takes the drug and Y the outcome variable coded as 1 if she dies and 0 if not. We will denote with $X_A = \{0, 1\}$ the value of Ann's exposure and $Y_A = \{0, 1\}$ the value of Ann's outcome.

This two causal questions, regarding the situation represented in Example 2.0.1, can be simply identified as:

Effects of Causes (EoC) Ann has not taken the drug yet. What happens to Ann if she decides to take the drug? What happens to Ann if she decides to not take it?

Causes of Effects (CoE) Unfortunately, Ann took the drug and passed away, how likely is that she would not have died if she had not taken the drug?

The above causal questions point out two different perspectives. The first is a perfect decision problem: what decision will improve Ann's survival probability? What is the effect of X on Y?

A more tricky situation is described by the CoE causal question because it hides an *individual* problematic. Dawid *et al.* (2015) in [20] argue that this approach can be considered Bayesian in several aspects mostly because of the subjective nature of CoE questions. In this dissertation we will discuss how, this subjected query, can be solved using the information coming from real data.

It is surprising to discover that one of the first attempt to formalize the distinction between CoE and EoC quetions goes back to 1774 and is due to Laplace ([34] and translate in English by Stingler (1986) in [35]). In his memoir, Laplace defined the uncertainty of human knowledge as concerned with causes and events. Considering an urn containing black and white balls, he defined a cause as the ratio of white and black balls and an event with the usual probability notation "drawn a white ball by chance". If an event E can be produced by two causes C_1 and C_2, he defined the uncertainty on the causes C_1 of the effect E, as the probability $P(C_1|E)$ where the cause is unknown and the event is given. On the other hand, he defined the uncertainty on the effect of causes as the probability of the event when the cause is given $P(E|C_1)$. Is even more surprising to discover that the well known formulation of Bayes' Theorem is actually due to Laplace in 1774. In fact, Bayes' first formulation of his famous theorem was developed to compute the distribution (rather than talking about events) for the probability parameter of a binomial distribution conditioning on the observations.
From Laplace's definition, we might be tempted to relate EoC and CoE questions via Laplace-Bayes' Theorem. However, the distinction between Ann's causal queries point out a more complex situation.

Suppose that a good experimental study, in which subjects were randomly assigned to be either exposed ($X = 1$) or unexposed ($X = 0$), tested the same drug that Ann might take, and produced the data reported in Table 2.1.

	Die	Live	Total
Exposed	30	70	100
Unexposed	12	88	100

Table 2.1: Deaths in individuals exposed and unexposed to the same drug that Ann might take

Since our analysis here is not concerned with purely statistical variation due to small sample sizes, we take proportions computed from this table as accurate estimates of the corresponding population probabilities (but see Dawid [20] for issues related to the use of small-sample data for making causal inferences).
Thus we have

$$P(Y = 1 \mid X \leftarrow 1) = 0.30 \quad (2.1)$$
$$P(Y = 1 \mid X \leftarrow 0) = 0.12. \quad (2.2)$$

We see that, in the experimental population, individuals exposed to the drug ($X \leftarrow 1$) were more likely to die than those unexposed ($X \leftarrow 0$), by 18 percentage points. Throughout this section, we will use situations similar to those reproduced in Table 2.1 to try to answer both EoC and CoE questions.

2.1 Effects of Causes

Questions on the effects of observed causes, named "EoC", identify much of classical statistical design and analysis as, for example, randomized clinical trials. In the EoC framework we would be interested in asking: "What would happen to Ann if she were to take the drug?" or "What would happen to Ann if she were not to take the drug?". Let us consider the information encoded in Table 2.1 where, in a experimental population, individuals exposed to the drug were 18% more likely to die than those unexposed. According to this results, if Ann can be considered comparable with the individual in the experiment, taking the drug will not be the preferable decision.

In particular, the EoC framework is interested, rather than to an individual-level causal effect, to a population-level causal effect. Knowing the answer of both individual EoC questions for every subject in a population, we can answer to the more general query: "Death is effectively caused by the drug?". In this case, Dawid (2015) [20], the EoC causal inference is based on a simple contrast between the two distributions $P_1 = P(Y = 1|X = 1)$ and $P_0 = P(Y = 1|X = 0)$, *i.e.* the probability distributions of Y ensuing when X is set to the value 1 and 0 respectively. According to Dawid, as we have mentioned in § 1.6, assessing the

effects of causes can be achieved in straightforward fashion using a framework based on probabilistic prediction and statistical decision theory where the two distributions, P_1 and P_0, are all that is needed to address EoC queries. He formulated this situation as a perfect decision problem: we can compare these two different distributions for Y, decide which one we prefer, and take the associated decision. The perfect tool to address this type of queries can be simply defined as the difference $P(Y = 1|X = 1) - P(Y = 0|X = 0)$. In § 1 we introduced causality using more complex definitions such as interventions § 1.2 and counterfactuals § 1.3. However, in simple cases, these methods are completely equivalent. In fact, this difference $P_1 - P_0$ coincides with the Average Causal Effects defined in Definition 1.3.3 in the counterfactual framework when the exchangeability condition holds.

In the rest of this thesis we will abandon the decision theory approach to the counterfactual framework. In fact, this thesis is focused on studying the causal effect of an exposure on an outcome when a third variable is involved in the pathway as a mediator. Even if important results have been found to study mediation in a non-counterfactual framework [22, 23], most of statistical methods and softwares have been implemented within the counterfactual approach.

2.2 Causes of Effects

A more tricky situation is described by questions on the cause of observed effects CoE "Unfortunately, Ann took the drug and passed away, how likely is that she would not have died if she had not taken the drug?". They hide an individual problematic that we want to address using statistical data.

This kind of queries are common in a Court of Law, when we want to asses legal (usually individual) responsibility. For instance, considering the Example 2.0.1, where we supposed that Ann has developed the outcome after being exposed. A typical question will be "Knowing that Ann did take the drug and passed away, how likely she would not have died if she had not taken the drug?". The problem is that we know the real value of exposure and outcome for Ann and then we can not longer base our answer on the difference $P_1 - P_0$.
In fact we can not base our approach purely on the probability distribution of Y and X conditioned on known facts. We know the values of both variables ($Y = 1$, $X = 1$), and after conditioning on that knowledge, there is no probabilistic uncertainty left to work with. Nevertheless we want an answer. To answer the CoE question: "Did the drug cause her death?" we should known what would have happened had she not taken the drug. This circumstance is actually the perfect situation to introduce "potential variables", usually apply to CoE questions instead. As described in § 1.3, in the potential notation, at any individual i, we can associate two pair of variables $\boldsymbol{Y}_i := (Y_i(0), Y_i(1))$ where $Y(x)$ denotes

the potential value that Y would take if X had been set to x. Both potential responses are regarded as existing, simultaneously, prior to the choice of X. For any subject, only one of them will be observable, the other is called *counterfactual*. In the case of Ann, we can only observe $Y_A(1) = 1$ while $Y_A(0)$ will be unknown.

In this paper we aim to investigate causation for CoE queries by using the formulation of *Probability of Causation*, as given by Dawid (2011) in [16] and also named by Pearl as *Probability of Necessity* [48].
In terms of the triple $(X_A, Y_A(0), Y_A(1))$, the Probability of Causation in Ann's case is defined as:

Definition 2.2.1

$$PC_A = P_A(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1)$$

where P_A denotes the probability distribution over attributes of Ann different from (2.1) and (2.2) which denote population probabilities.

Given the fact that Ann actually took the drug and passed away, the probability of causation defined above will provide an answer to the question: how likely she would not have died if she had not taken the drug?

Let us consider again the data provided in Table 2.1 regarding the same drug taken by Ann. Can the court infer that it was Ann's taking the drug that caused her death? More generally: Is it correct to use such experimental results, concerning a population, to say something about a single individual? This "Group-to-individual" (G2i) issue is discussed by Dawid (2013) in [18] in relation to the question "When can Science be relied upon to answer factual disputes in litigation?". In the next sessions we will discuss how to incorporate those information when calculating the Probability of Causation.

The bivariate potential distribution underlined in Definition 2.2.1, prohibits a point estimation for PC_A but permits useful information in the form of bounds. In this dissertation we review the current literature of bounding the probability of causation in a set of framework: the simple analysis of an exposure acting on an outcome, when additional information about a pre-treatment covariate is available and when unmeasured confounding affects the exposure-outcome relation. In addition we we will introduce new bounds for PC_A in the case of complete mediation between exposure and outcome and in the most real situation of partial mediation.

Chapter 3

Mediation as EoC: methods and historical background

Several factors can influence the presence of an outcome in a population. In § 1.2 we described one of the most popular approach to infer causation when a directed acyclic graph, connecting the exposure to the outcome, is established. A particular mechanisms that may link X to Y is the one arising by a chain involving a third variable called *Mediator*.

In this thesis we focus on the study of the *Mediation analysis* approach which investigates the mechanisms depicted in the DAG of Figure 3.1. In addition to a direct effect of X on Y, Figure 3.1 suggests a (presumed) chain of causes where the independent variable X affects a third one, the *Mediator*, which then affects the outcome Y.

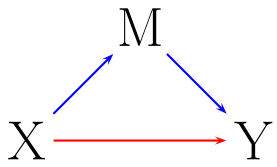


Figure 3.1: DAG illustrating a Mediation Mechanism between an exposure X, an outcome Y and a mediator M

This setting is crucial in Epidemiology when the researcher wants to quantify how much the total causal effect of X on Y is purely direct and how much is mediated by M. Rather than suppose a unique effect from X on Y, the mediation model disentangles the pathway in two different effects. The main goal of Mediation analysis is to quantify how much the total causal effect of X on Y is mediated by M. In this dissertation we will discuss different methodologies capable of measuring distinctly this effects.

Outline an historical background for Mediation analysis is very complex and

confusing. The major problem is the large overlaps between sources. According to MacKinnon (2008) in [37], on of the first example of mediation hails from Aristotle in the 3rd century BC. He identified *four causes* as all possible explanations of a change, classifying the four possible answers to the fundamental question "why?". These four causes include: material cause, formal cause, efficient cause and final cause. In particular, he defined an efficient cause as an agent of the change or the thing that brings something about. For example, a father is the efficient cause of a child and a carpenter is the efficient cause of a table. Investigating on an efficient cause of a change is actually the major goal of mediation analysis.

More recently, several authors adopt mediation to analyze the potential causal relation between exposure and outcome. One of the first researcher that studied mediation is MacKinnon (2009) in [38]. In this paper he presented a psychology model for the chain arising after a stimulus: it has an effect on the organism that will, in turn, produce a response. This is an example of a *complete mediation model* because underlines a unique indirect chain where the organism mediates completely the effect of the stimulus on the response. It can be described as the mediation mechanism in Figure 3.1 in the presence of the blue arrow alone, *i.e.* the stimulus does not affect the response directly. Other examples of complete mediation can be found in MacKinnon (2009) in [38]: a tobacco prevention program reduces cigarette smoking by changing the social norms for tobacco use; exposure to negative life events affects blood pressure through the mediation of cognitive attributions to stress. Once a true mediating process is identified, he pointed out the importance of mediation as it can develop treatment effect improving the mediation mechanism.

In the next sections we will introduce and discuss other historical and modern approach to Mediation analysis.

3.1 Model based approach to Mediation analysis

In this section we will introduce several methods capable of measuring mediation effects via model based computations while in the next section we will introduced model free definitions. In addition, we will show how, in simple and linear framework, these methods will lead to equivalent results.

3.1.1 Path Analysis

Several authors proposed distinct methodologies to Mediation according to the different fields to which they had been applied. From a Philosophical to a Statistical point of view, one historical contribution to mediation analysis came from Sewall Wright (1921) in [82]. With his *path analysis*, Wright proposed to quantify causal effects linking a regression coefficient with every path in a diagram. He then defined

the overall effect of the exposure on the outcome as simply the product of the two single effects of X on M and of M on X.

An equivalent approach was proposed by Baron & Kelly (1986) in [4]. Let us consider the triangle of causes depicted in Figure 3.2 where the symbols in each arrow correspond to a coefficient in two different models

$$M = i_1 + aX + e_1 \tag{3.1}$$
$$Y = i_2 + cX + bM + e_2. \tag{3.2}$$

Let us consider also the model for Y regressed on X alone

$$Y = i_3 + c'X + e_3 \tag{3.3}$$

where the terms e_1, e_2 and e_3 are the residuals uncorrelated with each other and with X.

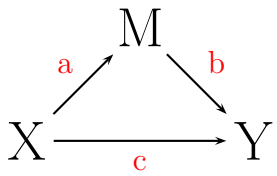


Figure 3.2: DAG illustrating a Mediation Mechanism and its relation to the product method

Baron and Kenny defined four steps that have to be satisfied to identify a true mediating process.

Baron & Kenny assumptions

1. the exposure should be significantly associated to the mediator, *i.e.* the estimate $\hat{a}$ of a in (3.1) should be statistically significant;

2. the mediator should be significantly associated with the outcome when both exposure and mediator are controlled for, *i.e.* the estimate $\hat{b}$ of b in (3.2) should be statistically significant;

3. the exposure should be significantly associated with the outcome, *i.e.* the estimate $\hat{c}'$ of c' in (3.3) should be statistically significant;

4. the estimate of the coefficient relating the outcome to the exposure only, must be bigger (in absolute value) than the estimate of the coefficient relating the exposure to the outcome in the model with the mediator, *i.e.* $|\hat{c}'| > |\hat{c}|$.

Baron and Kenny (1986) in [4], as suggested by Judd and Kenny (1981) in [33], propose two different methods to calculate mediation effects, the product and the difference methods. The first approach consists in calculating mediation effects including equation (3.1) in (3.2)

$$Y = (i_2 + i_1 b) + (c + ab)X + (e_2 + e_2 b).$$

We can then estimate the *total effect* of X on Y as $\hat{c} + \hat{b}\hat{a}$. Baron and Kenny proposed to estimate the *direct effect* of X on Y as $\hat{c}$ and $\hat{b}\hat{a}$ as the estimated *indirect effect* from which the name "Product Method" comes from. On the other hand, comparing (3.2) to (3.3), we have that $\hat{c}' = \hat{c} + \hat{b}\hat{a}$ and then $\hat{c}' - \hat{c} = \hat{b}\hat{a}$ where $\hat{c}' - \hat{c}$ is the indirect effect estimated by the "Difference method". The equivalence of these two methods, in terms of ordinary least square and maximum likelihood, was shown by MacKinnon *et al.* (1995) in [39].

In the light of these considerations, condition (4) of Baron and Kenny states that M can be considered a mediator if the product (or the difference) method gives raise to an indirect effects different from zero. The situation is indeed much more complicated then the four steps introduced by Baron and Kenny. Furthermore, if direct and indirect effects have opposite signs, the total effect of X on Y ($\hat{c}'$) could not be statistically significant even in the case that M is a true mediator (see [37]). This phenomenon is called *inconsistent mediation* and is common in multiple mediators model. In the next section we will show how, an estimated null indirect effect, cannot be considered alone to conclude the absence of a mediation mechanism.

3.1.2 Linear Structural Equation Modelling

The original Baron and Kenny approach did not have covariates, the *Structural equation model* approach (SEM) generalizes path analysis defining a statistical model for every endogenous variable in a DAG [7].

Nowadays, the SEM approach include several methods designed to cover a large set of statistical problems such as: confirmatory factor analysis, latent variable models and path analysis. They are frequent in psychology, widely used to summarize latent information. This knowledge is then linked to measured variables via multivariate models. The main idea is to test whether these hypothetical models are consistents with the observed data. For example, we are not able to measure concepts such as human intelligence or stress directly. Using the SEM approach we can collect information on a set of variables capable of summarizing these latent constructs. Let us suppose to be interested in evaluating the association between intelligence and income. To measure human intelligence we can construct an index based on measured variables such as the academic performance and/or the results

of an IQ test. The SEM approach is the perfect tool to investigate these associations.

The name "Structural equation model" comes from the dual nature of this multivariate model: a first latent model called *measurement model* able to define an index for the latent variable and a *structural regression model* that depicts the causal dependencies between latent and observed variables.

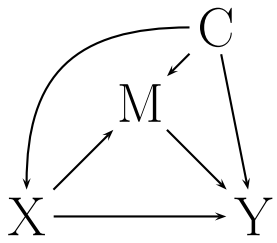


Figure 3.3: DAG illustrating a Mediation Mechanism with confounder C

In the simplistic case of only continuous measured variables, the SEM approach for the diagram in Figure 3.3 will lead to the following structural regression models [21]

$$M = \beta_0 + \beta_x X + \beta_c C + \epsilon_m \tag{3.4}$$
$$Y = \theta_0 + \theta_x X + \theta_m M + \theta_c C + \epsilon_y. \tag{3.5}$$

The error terms ϵ_M and ϵ_Y have zero means and are supposed not to be correlated with each other and with X and C. As the path analysis, including (3.4) in (3.5) we have

$$Y = (\theta_0 + \beta_0\theta_m) + (\theta_x + \beta_x\theta_m)X + (\theta_c + \beta_c\theta_m)C + (\theta_m\epsilon_m + \epsilon_y) \tag{3.6}$$

where $\theta_x + \beta_x\theta_m$ represents the total effect of X on Y composed by a pure direct effect θ_x and an indirect effect $\beta_x\theta_m$. This is exactly a generalization of path analysis in the presence of a set of confounders C.

Let us consider another situation where an additional variable L, called "intermediate confounder", confounds the mediator-outcome relation and is affected by the exposure [21].

In the case of continuous measured variables, the SEM approach for the diagram in Figure 3.4 will lead to the following structural regression models

$$L = \alpha_0 + \alpha_x X + \epsilon_l \tag{3.7}$$

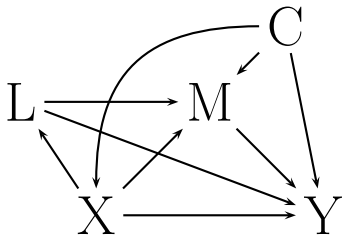


Figure 3.4: DAG illustrating a Mediation Mechanism with confounder C and intermediate confounder L

$$M = \beta_0 + \beta_x X + \beta_c C + \beta_l L + \epsilon_m \tag{3.8}$$
$$Y = \theta_0 + \theta_x X + \theta_m M + \theta_c C + \theta_l L + \epsilon_y. \tag{3.9}$$

where additionally ϵ_l is supposed not to be correlated with the other error terms and both X and C. As the path analysis, including (3.7) in (3.8) and then in (3.9) we have

$$\begin{aligned} Y = (\theta_0 + \theta_m\beta_0 + \theta_m\beta_l\alpha_0 + \theta_l\alpha_0) + (\theta_x + \theta_m\beta_x + \theta_m\beta_l\alpha_x + \theta_l\alpha_x)X + \\ (\theta_c + \theta_m\beta_c)C + (\theta_m\beta_l\epsilon_l + \theta_m\epsilon_m + \theta_l\epsilon_l + \epsilon_y) \end{aligned} \tag{3.10}$$

The term $\theta_x + \theta_m\beta_x + \theta_m\beta_l\alpha_x + \theta_l\alpha_x$ can then be used to estimate the total effect of X on Y. The direct effect can be estimated using $\theta_x + \theta_l\alpha_x$ and the indirect effect using $\theta_m\beta_x + \theta_m\beta_l\alpha_x$. Linking every path to a coefficient, we can see that indirect is considered every (directed) path that go through the mediator, even the one passing over L and M consecutively. On the other hand, direct is considered every remaining pathway not passing trough the mediator. Then, in the simple case of continuous variables for models that do not involve interactions and nonlinearities, to measure both direct and indirect effects we have to multiply consecutive pathways and sum over them.

This procedure is as simple as the DAG representing the mechanisms for which the exposure acts on the outcome. MacKinnon (2008) in [37], extends these models to more complex situations with more than one independent variable, mediator and outcome (*path analysis models*). Hayes *et al.* (2010) in [24], generalize this approach defining an instantaneous indirect effect for any non-linear models that are linear in the parameters. Muthen (2011) in [43] derives formulas to handle mediation by a nominal variable. Traditional approach as Path Analysis may produce flawed results when more complicated DAG, involving non-linearities and interactions, are required. In fact, path analysis cannot be use in complicated models since we cannot define an arrow to represent non-linearities and interactions.

3.2 Counterfactual approach to Mediation

In § 1.3 we introduced counterfactual variables able to represent changes in the outcome that we might not be able to observe. However, they can potentially be measured from the data if the exposure is randomized or if the exchangeability condition holds in every strata of the confounding. We defined the potential variable related to Y as

Definition 3.2.1 (Potential Variable) *Let us consider X the exposure of interest and Y the outcome, $Y(x)$ is the potential value that Y would take if X had been set to x.*

This new variable $Y(x)$ is treated as an ordinary random variable with a distribution that is consistent with the usual axioms of probability and independence. It is connected to the real outcome by the consistency condition $Y_{obs} = X \cdot Y(1) + (1 - X)Y(0)$ for binary exposure.
In the counterfactual notation, Mediation analysis requires the definition of a potential variable for every endogenous entry in the DAG.

For the simple mediation mechanism in Figure 3.1 the potential variables are:

i) $M(x)$ the potential value that M would take if X had been set to x;

ii) $Y(x, m)$ the potential value that Y would take if X had been set to x and M to m;

iii) $Y(x, M(\tilde{x}))$ the nested counterfactual value that Y would take if X had been set to x and M to $M(\tilde{x})$ that is when M arises naturally after setting X to $\tilde{x}$.

According to the research question of interest, mediation analysis provides useful tools to address different type of causal questions (here we are only referring to EoC causal questions).

If the researcher is interested in the effect of the drug in the population for every existing pathways of drug use, the Total Causal Effect is the target. This causal measure is the easiest to interpret, define and estimate [50]. This is also called "Average Treatment Effect", widely known in Clinical Studies where the exposure is randomized and the clinician wants to calculate how being or not exposed to the treatment will change the outcome.

In this thesis we will use the definition of mediation effects as given by Pearl (2000) in [50].

Definition 3.2.2 *The Total Causal Effect of X on Y, for every $x, \tilde{x}$ is:*

$$TCE(x, \tilde{x}) = E[Y(x)] - E[Y(\tilde{x})]. \tag{3.11}$$

The total causal effect is exactly the average causal effect defined in § 1.3. When X is binary $TCE = P[Y(1) = 1] - P[Y(0) = 0]$.

The TCE captures the real comprehensive effect of the exposure on the outcome because it contrasts two hypothetical worlds, one where all subjects are exposed to the drug and one where all subjects are not exposed to the drug.

In general rule we have that $E[Y(x)] \neq E[Y|X = x]$ but, under some assumptions, we can identify the TCE from the observed data. If C is a confounding variable affecting all exposure, outcome and mediator, those assumptions states:

TCE identifiability assumptions

1. No Interference: it assumes no interference between units (subjects) on their relatives outcome, *i.e.* no infectious diseases;

2. Consistency: the potential outcome must be equal to the real outcome when the exposure is observed, $Y(x) = Y$ if $X = x$. This assumption permits to estimate the potential outcome's average from the observed data $E[Y(x)|X = x] = E[Y|X = x]$;

3. Conditional Exchangeability (**CE**):

   a) no unmeasured confounding on the exposure-outcome relation $Y(x) \perp\!\!\!\perp X|C \ \ \forall x$. It means that control for C is enough to remove the $X - Y$ confounding, *i.e.* the subject in the population are conditionally exchangeable

It is important to notice that we can never test the CE assumption. It is related to the potential value of Y if all individuals were set to exposed and unexposed in the same time ($Y(0) \perp\!\!\!\perp X|C$ and $Y(1) \perp\!\!\!\perp X|C$). Given that only one exposure level is observable for each individual, this is actually a missing data problem.

If the above assumptions hold, for category C, we can estimate the TCE form the data as

$$\begin{aligned} TCE(x, \tilde{x}) &= E[Y(x)] - E[Y(\tilde{x})] \\ &= \sum_c \{E[Y(x)|C = c] - E[Y(\tilde{x})|C = c]\} \, P(C = c) \\ &= \sum_c \{E[Y(x)|X = x, C = c] - E[Y(\tilde{x})|X = \tilde{x}, C = c]\} \, P(C = c) \quad \mathbf{CE(a)} \end{aligned}$$

$$= \sum_c \{E[Y|X=x, C=c] - E[Y|X=\tilde{x}, C=c]\}\, P(C=c) \text{ consistency of } \mathbf{Y(x)}$$

where with **CE(a)** we mean the use of assumption $3a$. As we can see, for estimating the total causal effect, the above equation simply requires a correct measure of the associational effects $E[Y|X=x, C=c]$ and distributions $P(C=c)$.

An interesting example by Pearl (2001) in [50], is the case where a clinician is testing the efficacy of a drug treatment on a disease. Let has suppose that one possible side effect of this drug is headache. A possible mediator between the treatment and the disease can be the use of aspirin. In fact, subjects exposed to the drug will likely take the aspirin which in turn might have an effect on the disease. To determine how beneficial the drug is to the population as a whole, under existing patterns of aspirin usage, the TCE will be the right measure [50]. On the other hand, it would be interesting to see if encourage or discourage the use of aspirin during the treatment will affect the outcome. We might be interested in knowing what would be the effect of the treatment on the headache if a dose of aspirin was administrated to each patient.

This concept is measured by the following effect:

Definition 3.2.3 *The Controlled Direct Effect of X on Y when M is controlled at m is defined as:*

$$CDE(x, \tilde{x}, m) = E[Y(x, m)] - E[Y(\tilde{x}, m)] \tag{3.12}$$

This measure is able to quantify the sensitivity of Y to changes in X while all other factors (M) are controlled. Keeping the mediator fixed to a particular level m, the CDE is capable of measuring a direct effect of the exposure on the outcome. It is important to notice that the CDE depends on m, *i.e.* if the mediator is a variable defined by five different categories, we could define five different $CDEs$.

As for the TCE, the Controlled Direct Effect requires some additional assumptions to be measured in nonexperimental studies:

CDE identifiability assumptions

1. No Interference: between mediator and outcome;
2. Consistency: $Y(x, m) = Y$ when $X = x$ and $M = m$;
3. Conditional Exchangeability (**CE**):
   - **a)** no unmeasured confounding on the exposure-outcome relation $Y(x) \perp\!\!\!\perp X|C \ \forall x$;

b) no unmeasured confounding on the mediator-outcome relation $Y(x,m) \perp\!\!\!\perp M|C, X \ \forall x, m$.

If the above assumptions hold, for category C, we can estimate the CDE as

$$\begin{aligned}
CDE(x,\tilde{x},m) &= E[Y(x,m)] - E[Y(\tilde{x},m)] \\
&= \sum_c \{E[Y(x,m)|C=c] - E[Y(\tilde{x},m)|C=c]\}\, P(C=c) \\
&= \sum_c \{E[Y(x,m)|X=x, C=c] - E[Y(\tilde{x},m)|X=\tilde{x}, C=c]\}\, P(C=c) \qquad \mathbf{CE(a)} \\
&= \sum_c \{E[Y(x,m)|X=x, M=m, C=c] - E[Y(\tilde{x},m)|X=\tilde{x}, M=m, C=c]\}\, P(C=c) \qquad \mathbf{CE(b)} \\
&= \sum_c \{E[Y|X=x, M=m, C=c] - E[Y|X=\tilde{x}, M=m, C=c]\}\, P(C=c) \qquad \textbf{consistency of } \mathbf{Y(x,m)}
\end{aligned}$$

where with $\mathbf{CE(a)}$ and $\mathbf{CE(b)}$ we denote the use of assumptions $3a$ and $3b$. In § 5, using the SWIGs method defined in [57], we will prove that the assumptions $\mathbf{CE(a)}$ and $\mathbf{CE(b)}$ imply $Y(x,m) \perp\!\!\!\perp X|C$ end then the third equation in the estimation above.

However, there are situations where neither the TCE or CDE do not adequately represent the target of investigation. Let us consider and example proposed by Pearl (2001) in [50]. He considered the effect of a birth-control pill suspects of producing thrombosis [29]. It may be claimed that the pill, reducing the number of pregnancies, is an indirect protection for thrombosis (pregnancy is a known risk factor for thrombosis). To investigate the beneficial effect of the pill on the thrombosis via pregnancies rather than its direct effect, the TCE will not be exhaustive.

The idea is to define a measure sensitive to the direct effect of the exposure on the outcome and a measure sensitive to its effect via other consecutive pathways. Choice among mediation effects must depend on the research question of interest. In the situations described above, neither the TCE or the CDE are adequates. In fact, the CDE is a measure of the direct effect of X on Y for a particular level of the mediator. In contrast, here we would like to measure the overall direct effect rather than fixing M to a particular level. Several authors proposed different distinctions for direct and indirect effects [61],[50].

Reviewing Pearl (2001) in [50], the effect on the outcome directly attributable to the exposure is achieved by the following measure:

Definition 3.2.4 *The Pure Natural Direct Effect of X on Y is:*

$$PNDE(x,\tilde{x}) = E[Y(x, M(\tilde{x}))] - E[Y(\tilde{x}, M(\tilde{x}))] \tag{3.13}$$

According to Pearl (2001) in [50], Definition 3.2.4 permits to measure the direct effect comparing the potential composite variable $Y(x, M(\tilde{x}))$ with $Y(\tilde{x}, M(\tilde{x}))$ where they both set $M(x)$ as arising naturally on the reference value $\tilde{x}$. Usually

when X is binary, $\tilde{x} = 0$ and $x = 1$. For binary exposure, the pure natural direct effect capture a direct measure of X on Y comparing the distribution of the potential outcomes $Y(1, M(0))$ and $Y(0, M(0))$, when X change from exposed to unexposed subject and the mediator is fixed to the value $M(0)$. In the next section (METTI QUALE) we will describes mediation effects where $\tilde{x} = 1$.

For simplicity of interpretation, let us consider a binary exposure X where $x = 1$ and $\tilde{x} = 0$. Holding $M(0)$ in both terms of Definition 3.2.4, the $PNDE$ would capture the additional change in the outcome due only to the exposition at the drug (in respect to the unexposed).

To estimate the Pure Natural Direct Effect from nonexperimental data we need additional assumptions [79],[54],[50]:

PNDE identifiability assumptions

1. No Interference: between exposure and mediator;
2. Consistency on $M(x)$, $Y(x, m)$ and $Y(x, M(x))$: $M(x) = m$ when $X = x$, $Y(x, m) = Y$ if $X = x$ and $M = m$ and $Y(x, M(x)) = Y$ when $X = x$ and $M(x) = M$;
3. Conditional Exchangeability (**CE**):
   - **a)** no unmeasured confounding on the exposure-outcome relation $Y(x) \perp\!\!\!\perp X|C \ \forall x$;
   - **b)** no unmeasured confounding on the mediator-outcome relation $Y(x, m) \perp\!\!\!\perp M|C, X \ \forall x, m$;
   - **c)** no unmeasured confounding on the exposure-mediator relation $M(x) \perp\!\!\!\perp X|C \ \forall x$;
   - **d)** $Y(x, m) \perp\!\!\!\perp M(\tilde{x})|C \ \forall x, m$.

3.2.1 Different identifiably assumption for Natural Direct Effects

Condition **CE**(**d**) is one of the most discussed assumptions in Causal Inference. Here we use the identifiability assumption proposed by Pearl (2001) in [50]. It states that, for binary exposure and within level of confounding C, the individual counterfactual outcome $Y(1, m) \ \forall m$, does not depend on the counterfactual variable $M(0)$ that is the potential value of M if every subject were set to unexposed. In other words, within level of confounding C, the counterfactual outcome $Y(1, m)$ is a function of only the treatment, the mediator level and some errors $Y(1, m) = f(1, m, e)$ but not of $M(0)$. In the next section we will describe

the *SWIG* method [57] able to represent assumptions such as **CE**(**a**), **CE**(**b**) and **CE**(**c**) in a DAG. On the other hand, the assumption $Y(x,m) \perp\!\!\!\perp M(\tilde{x})|C$ is referred as a *cross world assumption* because it relates two different hypothetical worlds that cannot coexist together and hence, cannot be represented in a Single World Intervention DAG. Another way to represent potential variables in a graph is described by Pearl (2009) in [51] with the name of *twin network* or *triple network*. Pearl (2001) in [50] proves that **CE**(**d**) assumption can be easily read from a graph as $(Y \perp\!\!\!\perp M|X,C))_{G_{\underline{XM}}}$ where $G_{\underline{XM}}$ is the graph obtained mutilating all arrows coming out from X and M as described in § 1.2.4. It is not so easy to deduce this connection from Pearl's twin network method. In this section we will try to illustrate this method using the notations defined above.

According to Shpitser *et al.* (2008) in [69], if the query is to represent the value $E[Y(x,m)|M(\tilde{x}) = m, C = c]$ defined in Definition 3.2.4, the twin network will be composed by 3 different models, each connected to the others by latent mechanisms U (see the triple network in Figure 3.5). The first DAG at left in Figure 3.5 is the original model M where no intervention is made. The second DAG is the submodel M_{xm} where we intervene setting X to x and M to m and the last DAG at right in Figure 3.5 is the submodel $M_{\tilde{x}}$ where we intervene setting X to $\tilde{x}$. As described by the network in Figure 3.5, there are four different latent variables connecting each submodel: U_c connecting each confounding node C in every submodels, U_m connecting any mediator nodes, U_y connecting any outcome nodes and U_x connecting any exposure nodes. From Figure 3.5, we can notice that U_x is connected only with X in the original model M and not with the exposure intervention node in M_{xm} and $M_{\tilde{x}}$. This is because, according to what described in § 1.2.4, intervention removes all incoming arrows to the node.

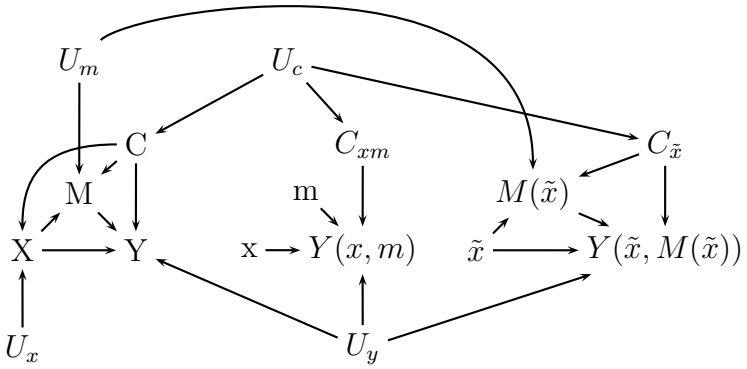


Figure 3.5: Triple network illustrating three different models connected by latent variables: the original model M at left where no intervention is made; the central submodel M_{xm} where the we intervene setting X to x and M to m; the submodel $M_{\tilde{x}}$ at right where we intervene setting X to $\tilde{x}$.

To verify the **CE**(**d**) assumption we should list all back-door path linking $Y(x,m)$ to $M(\tilde{x})$. These are:

1. $Y(x,m) \leftarrow C_{xm} \leftarrow U_c \rightarrow C_{\tilde{x}} \rightarrow M(\tilde{x})$
2. $Y(x,m) \leftarrow C_{xm} \leftarrow U_c \rightarrow C \rightarrow M \leftarrow U_m \rightarrow M(\tilde{x})$
3. $Y(x,m) \leftarrow C_{xm} \leftarrow U_c \rightarrow C \rightarrow X \rightarrow Y \leftarrow U_y \rightarrow Y(\tilde{x}, M(\tilde{x})) \leftarrow M(\tilde{x})$
4. $Y(x,m) \leftarrow C_{xm} \leftarrow U_c \rightarrow C \rightarrow Y \leftarrow U_y \rightarrow Y(\tilde{x}, M(\tilde{x})) \leftarrow M(\tilde{x})$

The first two back-door paths are blocked conditioning on C and the second two are already blocked because of the colliders. In [51] Pearl (2009) explained that the license to replace C with $C_{\tilde{x}}$ or C_{xm} is obtained from the third rule in Theorem 1.2.3. Then we can easily conclude that **CE**(**d**) holds in the original model M.
However **CE**(**d**) does not hold if there is a confounding variable between mediator and outcome that is affected by the exposure (see Figure 3.4).

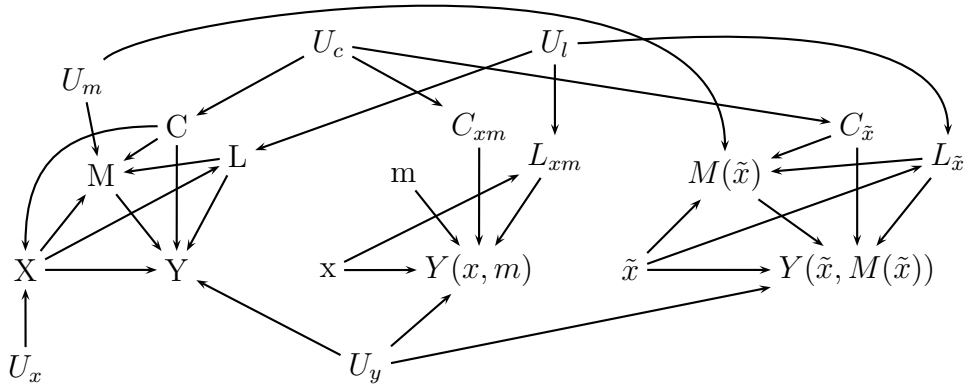


Figure 3.6: Triple network illustrating three different models connected by latent variables and intermediate confounding L: the original model M at left where no intervention is made; the central submodel M_{xm} where the we intervene setting X to x and M to m; the submodel $M_{\tilde{x}}$ at right where we intervene setting X to $\tilde{x}$.

In the situation described by the submodel at left in Figure 3.6, listing all back-door path from $Y(x,m)$ to $M(\tilde{x})$ we can see that C is not a sufficient set of confounding variables.
In conclusion, we can replace Pearl's assumption $Y(x,m) \perp\!\!\!\perp M(\tilde{x})|X, C$ with no unmeasured intermediate confounding between mediator and outcome affected by exposure. In the case that the no unmeasured intermediate confounding assumption is not possible we have to refer to different conditions in order to measure direct and indirect effects. Petersen *et al.* (2006) in [54], assume the following weaker assumption to identify natural direct effects:

$$E[Y(x,m) - Y(0,m)|M(0) = m, C] = E[Y(x,m) - Y(0,m)|C]. \tag{3.14}$$

Proving that Pearl's assumption $Y(x,m) \perp\!\!\!\perp M(\tilde{x})|C$ implies (3.14) is straightforward. Petersen *et al.* argued that, $Y(0,z)$ explains a lot of the variation of

$Y(x,z)$, suggesting that Pearl's assumption is less reasonable.
Robins and Greenland (1992) in [61], suggest an alternative assumption on the absence of interaction between exposure and mediator on the outcome. As we will see later, it implies to identify the CDE with the $PNDE$. Especially in Epidemiology, this assumption seems very unrealistic in many situations.

If the above assumptions hold, for category C, we can estimate the $E[Y(x,M(\tilde{x}))]$ form the data as

$$\begin{aligned}
E\{Y(x,M(\tilde{x}))\} &= \sum_m \sum_c E[Y(x,M(\tilde{x}))|M(\tilde{x})=m, C=c]P(M(\tilde{x})=m|C=c)P(C=c) \\
&= \sum_m \sum_c E[Y(x,m)|M(\tilde{x})=m), C=c]P(M(\tilde{x})=m|C=c)P(C=c) \ \textbf{consistency of } \boldsymbol{Y(x,M(x))} \\
&= \sum_m \sum_c E[Y(x,m)|C=c)]P(M(\tilde{x})=m|C=c)P(C=c) \quad \textbf{CE(d)} \\
&= \sum_m \sum_c E[Y(x,m)|X=x, C=c)]P(M(\tilde{x})=m|C=c)P(C=c) \ \textbf{CE(a)} \rightarrow \mathbf{Y(x,m) \perp\!\!\!\perp X|C} \\
&= \sum_m \sum_c E[Y(x,m)|X=x, M=m, C=c)]P(M(\tilde{x})=m|X=\tilde{x}, C=c)P(C=c) \ \textbf{CE(b) and CE(c)} \\
&= \sum_m \sum_c E[Y|X=x, M=m, C=c)]P(M=m|X=\tilde{x}, C=c)P(C=c) \ \textbf{Consitency of Y(x,m) and M(x)}
\end{aligned} \tag{3.15}$$

where with **CE(a)**, **CE(b)**, **CE(c)** and **CE(d)** we mean the use of assumptions $3a$ and $3b$, $3c$ and $3d$.

And

$$\begin{aligned}
E\{Y(\tilde{x},M(\tilde{x}))\} &= \sum_m \sum_c E[Y(\tilde{x},M(\tilde{x}))|M(\tilde{x})=m, C=c]P(M(\tilde{x})=m|C=c)P(C=c) \\
&= \sum_m \sum_c E[Y(\tilde{x},m)|M(\tilde{x})=m), C=c]P(M(\tilde{x})=m|C=c)P(C=c) \ \textbf{consistency of } \boldsymbol{Y(x,M(x))} \\
&= \sum_m \sum_c E[Y(\tilde{x},m)|C=c)]P(M(\tilde{x})=m|C=c)P(C=c) \quad \textbf{CE(d)} \\
&= \sum_m \sum_c E[Y(\tilde{x},m)|X=x, C=c)]P(M(\tilde{x})=m|C=c)P(C=c) \ \textbf{CE(a)} \rightarrow \mathbf{Y(x,m) \perp\!\!\!\perp X|C} \\
&= \sum_m \sum_c E[Y(\tilde{x},m)|X=x, M=m, C=c)]P(M(\tilde{x})=m|X=\tilde{x}, C=c)P(C=c) \ \textbf{CE(b) and CE(c)} \\
&= \sum_m \sum_c E[Y|X=\tilde{x}, M=m, C=c)]P(M=m|X=\tilde{x}, C=c)P(C=c) \ \textbf{Consitency of Y(x,m) and M(x).}
\end{aligned}$$

On the other hand, a measure of the indirect effect can be defined as:

Definition 3.2.5 *The Total Natural Indirect Effect of X on Y through M:*

$$TNIE = E[Y(x,M(x))] - E[Y(x,M(\tilde{x}))] \tag{3.16}$$

For binary exposure, the $TNIE$ is a comparison between the distribution of the two potential outcomes $Y(1,M(1))$ and $Y(1,M(0))$ where the exposure is held to $x=1$ and the mediator is allowed to change from $M(0)$ to $M(1)$. Holding the level

of X fixed and varying only the mediator, the $TNIE$ captures, by definition, the indirect effect via M.

The suffix *Total* refers to the decomposition of the TCE as a sum of the $PNDE$ and the $TNIE$:

Theorem 3.2.1 *The Total Natural Indirect Effect of X on Y:*

$$TNIE = TCE - PNDE \tag{3.17}$$

Proof 3.2.1

$$\begin{aligned} TNIE &= TCE - PNDE = E[Y(x)] - E[Y(\tilde{x})] - \{E[Y(x, M(\tilde{x}))] - E[Y(\tilde{x}, M(\tilde{x}))]\} \\ &= E[Y(x, M(x))] - E[Y(\tilde{x}, M(\tilde{x}))] - \{E[Y(x, M(\tilde{x}))] - E[Y(\tilde{x}, M(\tilde{x}))]\} \\ &= E[Y(x, M(x))] - E[Y(x, M(\tilde{x}))] \end{aligned}$$

This is a fundamental result in Mediation analysis an in particular in Epidemiology. It permits to disentangle the total effect such that we can always decide which pathway donates the major contribute. Estimation of $TNIE$ requires the same assumption for $PNDE$ and identifiability is similar to what we previously calculated.

3.2.2 Controlled Direct Effect vs Natural Direct Effect

In the previous section we illustrated how different mediation effects correspond to different research questions. Regarding the example of Pearl (2001) in [50], concerning the use of aspirin as a mediator between drug treatment and headache, we said that if we are interesting of knowing what would be the effect of the treatment on the headache if a dose of aspirin was administrated to each patient, the CDE will be the target. On the other hand, if the goal is to measure the overall direct effect of drug treatment on headache without fixing the mediator to any particular level, the $PNDE$ will provide an answer. Both these definitions are capable of measuring the direct effect of the exposure on the outcome but they share some differences. Even the assumptions needed to estimate both CDE and $PNDE$ are very different. The latter requires no intermediate confounding and no unmeasured confounding on the exposure-mediator relation in addition to no unmeasured confounding on the exposure-outcome relation and between mediator and outcome that are required for CDE to be estimated.

Pearl (2001) in [50] demonstrates that, if there is not unmeasured confounding between exposure and mediator, the $PNDE$ can be simply obtained as a weighted average of the $CDEs$ for different values of the mediator:

$$PNDE(x, \tilde{x}) = \sum_m \{E[Y(x, m)] - E[Y(\tilde{x}, m)]]P(M = m|\tilde{x}) \tag{3.18}$$

with weight given by $P(M = m|\tilde{x})$ for every level of M.

In the case of no interaction between exposure and mediator on the outcome, the difference $E[Y(x,m)] - E[Y(\tilde{x},m)]$ will be the same for different values of m and hence $PNDE$ would coincide with CDE from (3.18).

In the section § 3.3 we will show how, comparing the SEM approach to the Counterfactual, this equality will be much easier to see.

3.2.3 Alternative scales

According to the different types of variables, to the different types of study design and even purpose of the study, we can define mediation effects in alternative scales: differences, risk ratio, odds ratio and hazard ratio. In terms of odds ratio the mediation effects are

$$TCE = \frac{E[Y(1)]/1 - E[Y(1)]}{E[Y(0)]/1 - E[Y(0)]} \tag{3.19}$$

$$CDE(m) = \frac{E[Y(1,m)]/1 - E[Y(1,m)]}{E[Y(0,m)]/1 - E[Y(0,m)]}. \tag{3.20}$$

$$PNDE = \frac{E[Y(1,M(0))]/1 - E[Y(1,M(0))]}{E[Y(0,M(0))]/1 - E[Y(0,M(0))]} \tag{3.21}$$

$$TNIE = \frac{E[Y(1,M(1))]/1 - E[Y(1,M(1))]}{E[Y(1,M(0))]/1 - E[Y(1,M(0))]} \tag{3.22}$$

such that $TCE = PNDE \cdot TNIE$. Interpretation of Equations (3.19) to (3.20) are quite different from the effect defined in § 3.2 for the difference scale. Odds Ratios are associational measure commonly use to measure the relation between two variables, widely used in case-control studies [74]. An Odds Ratio (OR) relating Y to X will measure the odds for developing the outcome in subject exposed to X, compared to the odds for developing the outcome in subject unexposed to X. An OR equals to one will sustain no association between exposure and outcome. An OR bigger than one will imply that the exposure is associated with higher odds of the outcome and smaller than one when the exposure is associated with lower odds of the outcome. Let us consider again the example studied by Pearl (2001) in [50] where he reflected on the effect of a birth-control pill on thrombosis, potentially mediated by the number of pregnancies. The TCE, in the OR scale, will then measure the odds for developing thrombosis comparing women exposed and unexposed to the pill. A $CDE(m)$, in the OR scale, will measure the odds for developing thrombosis comparing women exposed and unexposed to the pill if all

subject were set to have m pregnancies. The odds ratio $PNDE$ will instead measure the odds for developing thrombosis comparing women exposed and unexposed to the pill, setting the number of pregnancies at the value that they would have had if they were not exposed to the pill. The odds ratio $TNIE$ will capture the odds for developing thrombosis in women exposed to the drug, comparing the number of pregnancies set at the value that they would have had if they were exposed with the number of pregnancies set at the value that they would have had if they were not exposed.

We can even express mediation effects in terms of conditional risk ratio, *i.e.* within the observed level $C = c$

$$TCE = \frac{P[Y(1) = 1|\mathbf{C} = \mathbf{c}]}{P[Y(0) = 1|\mathbf{C} = \mathbf{c}]} \tag{3.23}$$

$$CDE(m) = \frac{P[Y(1, m) = 1|\mathbf{C} = \mathbf{c}]}{P[Y(0, m) = 1|\mathbf{C} = \mathbf{c}]} \tag{3.24}$$

$$PNDE = \frac{P[Y(1, M(0)) = 1|\mathbf{C} = \mathbf{c}]}{P[Y(0, M(0)) = 1|\mathbf{C} = \mathbf{c}]} \tag{3.25}$$

$$TNIE_{\mathbf{c}} = \frac{P[Y(1, M(1)) = 1|\mathbf{C} = \mathbf{c}]}{P[Y(1, M(0)) = 1|\mathbf{C} = \mathbf{c}]} \tag{3.26}$$

such that

$$TCE_{\mathbf{c}} = PNDE_{\mathbf{c}} \cdot TNIE_{\mathbf{c}}. \tag{3.27}$$

Different values of $\mathbf{C}$ will lead to different meaning of mediation effects. Interpretation of Equations (3.23) to (3.24) is quite different from the effects defined in § 3.2 for the difference scale. First of all, Equations (3.23) to (3.24) are conditional effects while definitions introduced in § 3.2 are marginal. However, identification of mediation effects in terms of risk ratio requires the same assumptions of the difference scale. By definition of conditional exchangeability given C, the simple exchangeability assumption will hold in every strata of the confounder, *i.e.* the conditional effects above are well defined and identifiable $\forall\, \mathbf{c}$. In particular, the $TCE_{\mathbf{c}}$, in the risks ratio scale defined by (3.23), will measure the risk (in stratum $\mathbf{C} = \mathbf{c}$) to develop the outcome if all subject have been exposed to the treatment compared to a situation where all subject have not been exposed. A risk ratio bigger than one will imply a harmful causal effect of the exposure on the outcome (in stratum $\mathbf{C} = \mathbf{c}$) while a risk ratio smaller than one will imply a protective effect (in stratum $\mathbf{C} = \mathbf{c}$). A risk ratio of one will imply no association. A $PNDE_{\mathbf{c}}$ bigger than one (smaller than one) will imply a harmful direct effect (a protective direct effect) of the exposure on the outcome in stratum $\mathbf{C} = \mathbf{c}$. A $PNDE_{\mathbf{c}}$ equals to one supports evidence of no direct effect of the exposure on the outcome in stratum $\mathbf{C} = \mathbf{c}$. Interpretation is similar for $NIE_{\mathbf{c}}$ and $CDE_{\mathbf{c}}$.

Under the assumptions stated in § 3.2, we can estimate the conditional mediation effects from observed data as:

$$TCE_{\mathbf{c}} = \frac{P[Y(1)=1|\mathbf{C}=\mathbf{c}]}{P[Y(0)=1|\mathbf{C}=\mathbf{c}]} = \frac{P[Y=1|X=1,\mathbf{C}=\mathbf{c}]}{P[Y=1|X=0,\mathbf{C}=\mathbf{c}]}, \tag{3.28}$$

$$CDE_{\mathbf{c}}(m) = \frac{P[Y(1,m)=1|\mathbf{C}=\mathbf{c}]}{P[Y(0,m)=1|\mathbf{C}=\mathbf{c}]} = \frac{P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})}{P(Y=1|X=0,M=m,\mathbf{C}=\mathbf{c})}, \tag{3.29}$$

$$\begin{aligned} PNDE_{\mathbf{c}} &= \frac{P[Y(1,M(0))=1|\mathbf{C}=\mathbf{c}]}{P[Y(0,M(0))=1|\mathbf{C}=\mathbf{c}]} \\ &= \frac{\sum_m P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=0,\mathbf{C}=\mathbf{c})}{\sum_m P(Y=1|X=0,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=0,\mathbf{C}=\mathbf{c})}, \end{aligned} \tag{3.30}$$

and

$$\begin{aligned} TNIE_{\mathbf{c}} &= \frac{P[Y(1,M(1))=1|\mathbf{C}=\mathbf{c}]}{P[Y(1,M(0))=1|\mathbf{C}=\mathbf{c}]} \\ &= \frac{\sum_m P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=1,\mathbf{C}=\mathbf{c})}{\sum_m P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=0,\mathbf{C}=\mathbf{c})}. \end{aligned} \tag{3.31}$$

3.2.4 Mediated interactive effect

The counterfactual definitions of mediation effects latently englobe an interaction between exposure and mediator on the outcome. Depending on how this interaction is considered, we can define supplementary definitions of mediation effects and, hence, obtain different decomposition of the total effect. In § 3.2, we choose to measure indirect effects comparing two hypothetical worlds: one were the mediator arises naturally if all subject have been set to exposed and one were the mediator arises naturally if all subject have been set to unexposed, both were the exposure has been set to one. A different way to define the indirect effect would be setting all subject to be unexposed.

Robins and Greenland defined this new effect as the "Pure Natural Indirect Effect":

$$PNIE = E[Y(0,M(1))] - E[Y(0,M(0))] \tag{3.32}$$

Thus, exactly as we did in § 3.2, they defined the "Total Natural Direct Effect" as

$$TNDE = TCE - PNIE = E[Y(1, M(1))] - E[Y(0, M(1))], \tag{3.33}$$

where we are comparing two hypothetical worlds: one were all subject have been exposed and one where all subject have been unexposed, both with the mediator set to the value that would have take if all subject have been set to exposed. In the light of this considerations, the terms "total" or "pure" are referred to the different decomposition of direct and indirect effects. According to VanderWeele (2013) in [79], this different terminology arises on how the interaction was accounted for.

VanderWeele (2013) in [79] suggests a three-way decomposition of the total effect into a purely direct effect, a purely indirect effect and a third effect called "mediated interactive effect":

$$TCE = PNDE + PNIE + E[Y(1, M(1)) - Y(1, M(0)) - Y(0, M(1)) + Y(0, M(0))]\, E[M(1) - M(0)] \tag{3.34}$$

such that

$$\begin{aligned} TNDE &= PNDE + E[Y(1, M(1)) - Y(1, M(0)) - Y(0, M(1)) + Y(0, M(0))]E[M(1) - M(0)] \\ TNIE &= PNIE + E[Y(1, M(1)) - Y(1, M(0)) - Y(0, M(1)) + Y(0, M(0))]E[M(1) - M(0)]. \end{aligned}$$

Accounting for the mediated interactive effect in the indirect effect will lead to the usual decomposition of $TCE = TNIE + PNDE$.

In terms of conditional relative risks we have

$$\begin{aligned} PNIE_{\mathbf{c}} &= \frac{E[Y(0, M(1))|\mathbf{C} = \mathbf{c}]}{E[Y(0, M(0))|\mathbf{C} = \mathbf{c}]} \\ TNDE_{\mathbf{c}} &= \frac{TCE}{PNIE} = \frac{E[Y(1, M(1))|\mathbf{C} = \mathbf{c}]}{E[Y(0, M(1))|\mathbf{C} = \mathbf{c}]}. \end{aligned}$$

VanderWeele suggested a three-way additive decomposition of the total effect (in terms of conditional excess relative risk) into a purely direct excess, a purely indirect excess and a measure of "mediated excess relative risk due to interaction":

$$TCE_{\mathbf{c}} - 1 = (PNDE_{\mathbf{c}} - 1) + (PNIE_{\mathbf{c}} - 1) + RERI_{mediated} \tag{3.35}$$

where

$$RERI_{mediated} = \frac{E[Y(1, M(1))|c]}{E[Y(0, M(0))|c]} - \frac{E[Y(1, M(0))|c]}{E[Y(0, M(0))|c]} - \frac{E[Y(0, M(1))|c]}{E[Y(0, M(0))|c]} + 1. \tag{3.36}$$

Equation (3.35) consents to evaluate the proportion of causal effect attributable to the direct effect, a proportion of causal effect attributable to the indirect effect and a quantity attributable to the mediated interaction.

In terms of conditional risk ratio (multiplicative scale) we have found

$$TCE_{\mathbf{c}} = PNDE_{\mathbf{c}} \cdot PNIE_{\mathbf{c}} \cdot K_{\mathbf{c}} \tag{3.37}$$

where

$$K_{\mathbf{c}} = \frac{TNDE_{\mathbf{c}}}{PNDE_{\mathbf{c}}} = \frac{TNIE_{\mathbf{c}}}{PNIE_{\mathbf{c}}} = \frac{RR_{11}}{RR_{10} \cdot RR_{01}} \tag{3.38}$$

and $RR_{ij} = E[Y(i, M(j))|c]/E[Y(0, M(0))|c]$. The term K is a measure of the interaction in a multiplicative scale defined as the amount to which the effect of exposure and mediator together, exceeds the effect of each considered individually (Chapter 15 and 18 in [63]). Identifiability of $PNIE$ and $TNDE$ requires the same assumptions as for $PNDE$ and $TNIE$.

3.2.5 G-Computation in Mediation

Equation (3.15) correspond, in mediation analysis, to the standardization formula defined in (1.11) for the simple case of an exposure on the outcome. Pearl in [53, 50] called the first equation with the name of *mediation formula*. It additionally requires integrating over M to obtain direct and indirect effects. In the next section we will face a real mediation problem by using two different methods: g-computation and a counterfactual regression based approach. In the first scenario, we estimated mediation effects using the fully parametric implementation of Pearl's mediation formula which is performed in the *gformula* command implemented in Stata 13 [12]. Instead of integrating analytically over M, in this package, the *gformula* procedure estimates causal effects by the g-computation procedure using Monte Carlo Simulations [60].

3.3 Counterfactual vs linear SEM

The two approach described above, of Counterfactual causal inference and Path Analysis, aim both to measure causal effects but have some important differences.

SEM approach is more intuitive and requires only our knowledge of regression models. However, its estimands can be defined only for simple models which are linear in the parameters. In fact, Sewall's idea to multiply consecutive pathways and sum over them, cannot be applied in presence of non-linearities

and interactions. Furthermore, they are model based definitions and require a correct specification of parameters. On the other hand, the counterfactual approach requires the specification of a new type of variables, the Potential variables, that are not completely observable and sometimes difficult to understand. Furthermore, mediation counterfactual effects can be estimated if and only if some un-testable assumptions are met. These issues are some of the reasons why, the Counterfactual approach, is highly criticized in the literature. However, these assumptions are not so different from SEM's conditions: first the models have to be correctly specified and second, all possible variables have to be measured to properly interpret the parameters. In addition, the counterfactual approach leads to model free definition that can be applied to any type of variable.

Despite these differences, they both aim to the same target and, in simple situations, they produce the same results.

Let us consider the associational relation represented in the DAG in Figure 3.3 for which we can define the following linear models:

$$\begin{aligned} M &= \beta_0 + \beta_x X + \beta_c C + \epsilon_m \\ Y &= \theta_0 + \theta_x X + \theta_m M + \theta_c C + \epsilon_y. \end{aligned}$$

See [43] for more general situations. The SEM approach identifies $\theta_x + \beta_x\theta_m$ as the total effect of X on Y composed by a pure direct effect θ_x and an indirect effect $\beta_x\theta_m$.
Applying the counterfactual definitions to the model above we have, for category C:

$$\begin{aligned} CDE(x,\tilde{x},m) &= E[Y(x,m)] - E[Y(\tilde{x},m)] = \\ &= \sum_c \{E(Y|X=x,C=c,M=m) - E(Y|X=\tilde{x},C=c,M=m)\}\, P(C=c) \\ &= \sum_c \{[\theta_0 + \theta_x x + \theta_m m + \theta_c c] - [\theta_0 + \theta_x \tilde{x} + \theta_m m + \theta_c c]\}\; P(C=c) \\ &= \sum_c \theta_x(x-\tilde{x})\, P(C=c) \\ &= \theta_x(x-\tilde{x}) \sum_c P(C=c) \\ &= \theta_x(x-\tilde{x}) \end{aligned}$$

where we can directly estimate the CDE because the variables in Figure 3.3 satisfy $CE(a)$ and $CE(b)$. Even for direct and indirect effects, the DAG in Figure 3.3 satisfies every CE assumption required for estimation:

$$PNDE(x,\tilde{x}) = E[Y(x,M(\tilde{x}))] - E[Y(\tilde{x},M(\tilde{x}))] =$$

$$
\begin{aligned}
&= \sum_c \int_m \{E(Y|X=x, C=c, M=m) - E(Y|X=\tilde{x}, C=c, M=m)\} \\
&\quad f_M(m|X=\tilde{x}, C=c)\, dm\, P(C=c) \\
&= \sum_c \int_m [(\theta_0 + \theta_x x + \theta_m m + \theta_c c) - (\theta_0 + \theta_x \tilde{x} + \theta_m m + \theta_c c)] \\
&\quad f_M(m|X=\tilde{x}, C=c)\, dm\, P(C=c) \\
&= \theta_x(x-\tilde{x}) \sum_c \int_m f_M(m|X=\tilde{x}, C=c)\, dm\, P(C=c) \\
&= \theta_x(x-\tilde{x})
\end{aligned}
$$

where, for continuous mediator, we replaced the sum with an integral over M and $P(M=m|X=\tilde{x}, C=c)$ with $f_M(m|X=\tilde{x}, C=c)$. Then, in the case of no interaction between exposure and mediator on the outcome, the CDE is equal to the $PNDE$.

For the natural indirect effect we have:

$$
\begin{aligned}
NIE(x,\tilde{x}) &= E[Y(x, M(x))] - E[Y(x, M(\tilde{x}))] = \\
&= \sum_c \int_m E(Y|X=x, M=m, C=c)\{f_M(m|X=x, C=c) - f_M(m|X=\tilde{x}, C=c)\} dm\, P(c) \\
&= \sum_c \int_m (\theta_0 + \theta_x x + \theta_m m + \theta_c c)\, \{f_M(m|X=x, C=c) - f_M(m|X=\tilde{x}, C=c)\}\, dm\, P(c) \\
&= \theta_m \sum_c \int_m m\{f_M(m|X=x, C=c) - f_M(m|X=\tilde{x}, C=c)\}\, dm\, P(c) \\
&= \theta_m \sum_c \{E(M|X=x, C=c) - E(M|X=\tilde{x}, C=c)\}\, P(c) \\
&= \theta_m \sum_c [(\beta_0 + \beta_x x + \beta_c c) - (\beta_0 + \beta_x \tilde{x} + \beta_c c)]\, P(c) \\
&= \theta_m \beta_x (x-\tilde{x})
\end{aligned}
$$

where $P(c) = P(C=c)$. Thus the Total Causal Effect will be simply the sum $\theta_x + \theta_m \beta_x$ exactly as the SEM approach.

Chapter 4

Mediation as EoC: applications to real problems

Questions on the effects of observed causes, named "EoC", identify much of classical statistical design and analysis as, for example, randomized clinical trials. A typical EoC question, regarding the Example 2.0.1, could be: "What would happen to Ann if she were to take the drug?" or "What would happen to Ann if she were not to take the drug?". To a population level a EoC question will be "Is death effectively caused by the drug?". EoC queries are usually involved in Epidemiology where we want to assess the effect of a relevant treatment on a outcome or a disease. Perfect randomized Clinical Trials (if there are no issues of measurement error or loss to follow up and if they are double bind) are the best tools to infer causation for EoC queries. Patients enrolled in a Clinical Trial are considered exchangeable in respect to the treatment.

In section § 2 we illustrated EoC questions as merely decision problem: we can compare the distribution of Y in subject exposed and non exposed to the treatment and take the associated decision. However, the situation is much more complex when dealing with real life situations. Even if in § 2.2 we assigned the counterfactual-based approach to CoE questions, counterfactual-based definition such as § 3.2 are naturally build for EoC queries.

In this chapter we will measure mediation effects for EoC queries using the methods defined in § 3.2. Furthermore we will discuss the problems arising from the interpretability of these measures.

4.0.1 NINFEA dataset

In this thesis we investigate two potential mediating mechanisms using data from a birth cohort called Ninfea [58]. Ninfea is an Italian web-based birth cohort of 6445 pregnant women. This study started in 2005 in Turin and was successively

extended to the rest of the country from December 2007. The main goal of the project is to investigate the effects of several pre-natal and post-natal exposures on later life events collecting a range of information: demographics about both parents and child, some maternal disease before and after pregnancy, occupational factors and some other prenatal and postnatal exposures.
Recruitment begins voluntarily during pregnancy with a first web administered questionnaire [1], with other follow up questionnaires planned at 6 and 18 months after delivery and when the children are 4 and 7 years old.
The eligibility criteria are to know the Italian language, have access to the Internet and know about the study([55, 56]).

4.1 Conditioning on a mediator

In § 1.2 we made aware the reader on selecting the appropriate set of covariates to adjust for. In subsection 1.2 we saw that, adjusting for a collider in the presence of uncontrolled confounding will open a non-causal path from exposure to outcome, biasing the resulting estimated causal effect. When discussing mediation we have to be very careful in order to avoid this type of bias. In fact, as we saw in § 3.2, identifying mediation effects usually requires to condition on covariates (to obtain total effects) and mediator (to obtain direct and indirect effects).

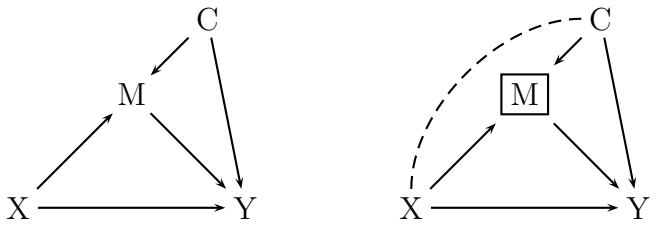


Figure 4.1: DAG illustrating a Mediation Mechanism affected by mediator-outcome confounding (a) with a spurious path arising from conditioning on M

Let us consider the DAG in Figure 4.1 where the variable C is a confounder of the mediator-outcome relation. In this DAG we can see that, conditioning on the mediator M, will open a spurious path from the confounder C to the exposure that could however be blocked by conditioning on C (in the model for Y that includes M). Problems arise when the covariate C is unmeasured and we cannot condition on it. When designing a study, the gold standard should be to draw a DAG with all possible confounders of the exposure-mediator, exposure-outcome and mediator-outcome relationships and include them in the data collection. However, there are latent phenomena which cannot be measured or identified.

One of the most recognized problem due to unmeasured confounding is the one arising when Birth Weight partially mediates the effect of a prenatal exposure on some outcomes. Hernndez-Diaz *et al.* (2006) in [28] discuss that, conditioning on birth weight, could lead to paradoxical results. They examined how infants born to smokers have higher risk of LBW (less than 2500g) and infant mortality than infants born to non-smokers, but in the LBW stratum maternal smoking appears not to be harmful for infant mortality relatively to non-smoking. When studying neonatal Epidemiology, birth weight is often considered as a strong predictor of infant mortality. Hernndez-Diaz *et al.* justify this choice because birth weight is one of the most collected hospital data and researchers often stratify on birth weight. They argue that this stratification is responsible of a crossover of the birth-weight-specific mortality curves. This phenomenon, known as the *Birth Weight paradox*, has been explained as a consequence of the presence of unmeasured confounding between birth weight and infant mortality.

VanderWeele (2012) in [81] proposes birth defects, which were not controlled for in [28], as the latent risk factor of both birth weight and infant mortality capable of explaining this apparent paradox. In fact, for smoking mothers with LBW children, LBW can be a consequence of both smoking or birth defects. On the other hand, LBW infants born from non smoking mothers should be affected by some other causes because LBW cannot be a reaction of smoking. Results obtained without controlling for birth defects will be biased.

Moreover, these paradoxical results are not limited to the effect of smoking on mortality. In fact, any mediation mechanism affected by unmeasured mediator-outcome confounder might produce similar problems.
The easiest solution would be to not condition on the mediator. However, if the goal is to measure mediation effects, we can not avoid it.

4.1.1 How to deal with the paradox

VanderWeele *et al.* (2012) in [81] proposed three different approaches to deal with this phenomenon. They are: conditioning on the estimated risk of being LBW instead of the mediator itself, conditioning on the mediator with sensitivity analysis and conditioning on the principal stratum. In this thesis we will focus only on the first two approaches. In particular, in the subsection 4.1.1 we will describe the first approach which consists in conditioning on the estimated risk of being LBW. It will be described here because it can be applied to both rare and regular outcome. The second approach is defined for rare outcomes and will be described in subsection 4.2.3. The third approach, conditioning on the principal stratum, involves assessing the effect of the exposure on the outcome among the subpopulation for whom the intermediate would be present irrespective of exposure status [81].

Given the aims of this thesis, that is assessing mediation analysis in a counterfactual framework, we will not further describe this method.

Conditioning on the Risk of an Intermediate

The first approach proposed by VanderWeele consists of conditioning on the estimated risk of being LBW predicted by baseline covariates denoted by **C** and mediator determinants **Det** as described in figure Figure 4.2. With mediator determinants we mean any factor or variable that can affect the frequency of the mediator. In fact, these variables can predict the mediator and can be used as proxy of a latent construct.

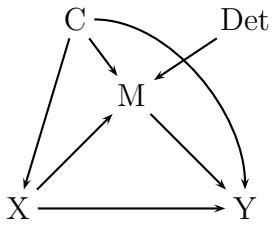


Figure 4.2: DAG illustrating a Mediation Mechanism including confounders and mediator determinants

VanderWeele suggested to predict individuals probabilities by the following logistic model for M

$$logit[P(M = 1|\mathbf{C} = \mathbf{c}, \mathbf{Det} = \mathbf{d})] = \gamma_0 + \gamma_1'\mathbf{c} + \gamma_2'\mathbf{d}.$$

The above model is then estimated by maximum likelihood and the parameters $\{\hat{\gamma}_0, \hat{\gamma}_1', \hat{\gamma}_2'\}$ are used to calculate the predicted probabilities of being LBW $\forall\ c, d$. He then defined a new variable H such that $H = 1$ for children who have predicted probabilities (of being LBW) above the 95th percentile, and zero otherwise:

$$H = \begin{cases} 1 & \text{if } logit^{-1}(\hat{\gamma}_0 + \hat{\gamma}_1'\mathbf{c} + \hat{\gamma}_2'\mathbf{d}) \geq 0.95 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Other cutoff can be used for different targets. He noticed that these new variable H is a function only of baseline covariates and mediator's determinants. Conditioning on it, does not imply conditioning on the mediator and hence generating collider bias. To assess whether the exposure is protective or harmful among the group of infants who have a high risk to be LBW, a logistic regression model can be fitted for the outcome on the exposure X, high risk status H, confounding variables C and an interaction term XH:

$$logit[P(Y = 1|x, h, \mathbf{c})] = \lambda_0 + \lambda_1 x + \lambda_2 h + \lambda_3 xh + \lambda_4'\mathbf{c} \tag{4.2}$$

Then e^{λ_1} will be the odds ratio for the outcome comparing exposed and unexposed low risk infants while $e^{\lambda_1+\lambda_3}$ the odds ratio for the outcome comparing exposed and unexposed high risk infants.

However, by this approach, we will not be able to estimate direct and indirect effects. In fact, the above measure are only estimates of the total effect of the exposure on the outcome for children at high and low risk to be LBW. Furthermore, this methods is highly affected by the choice of conditioning variables because a set of covariates with bigger predictive power will produce different results. In addition, if the mediator is rare, he noticed that "these measure may not be an accurate reflection of the effect of the exposure for whom the intermediate will in fact develop".

4.2 Rare Outcome

It has long been recognized that several factors can influence the presence of asthma in childhood such as environmental, genetic or demographic factors. Neil Pearce *et al.* (1998) in [46] described methods for measuring some of the major risk factors for asthma including parity or birth order. Shaw *et al.* (1994) [67] showed, in Kawerau children aged 8-13 years, a protective effect on current wheeze in those with older children living in the same house (for 2 or more older children in the same household, OR = 0.5, 95% CI 0.2-1.0). However is not common to find studies on the possible causal relation of birth order on asthma in earlier childhood. One way to evaluate when such a risk factor is causal is to identify all possible mechanisms that take place in and around this relation. A particular role is played by Birth Weight that should be always considered as a fundamental mechanism between birth order on asthma. In fact, in this causal relation, birth weight usually plays the role of mediator because it disentangles the pathway between exposure and outcome.

In the next sections we will describe the dataset and the variables included the final DAG considered. Then, after partitioning the causal effect in a purely direct effect from parity to recurrent wheezing and an indirect effect via birth weight, we evaluate the compatibility of these results with the phenomenon of the *Birth Weight paradox*. Furthermore we provide a plausible graphical explanation and explore the magnitude of the potential bias with ad hoc sensitivity analysis.

Dataset description

To discuss the presence of such paradox we used the Italian NINFEA web-based birth cohort described in § 4.0.1.

Asthma is a common disorder characterized by a various range of symptoms

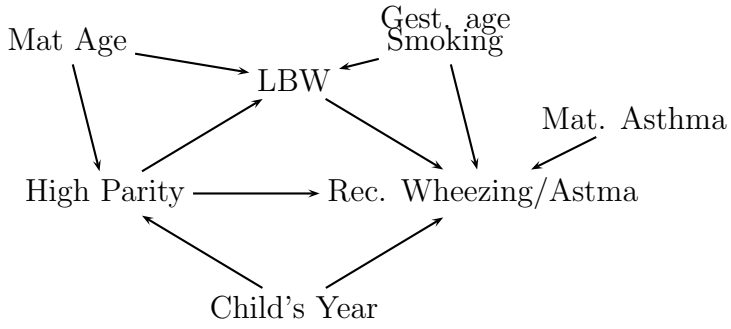


Figure 4.3: DAG representing the relational assumptions between high parity ($\geq$1), low birth weight (<2500g), recurrent wheezing or asthma up to 18 months, potential confounders: child's year of birth , maternal age, gestational age and maternal smoking. Maternal asthma was considered as a risk factor for childhood asthma

such as wheezing, chest tightness, dyspnea, and/or cough. In this section we considered recurrent wheezing up to 18 months or asthma (diagnosed by doctor) as an appropriate outcome for childhood asthma. The outcome was defined as recurrent wheezing (at least two episodes) or asthma up to 18 months of age. The exposure was parity dichotomized to be zero for first child and one otherwise and the mediator was birth weight dichotomized to be one for low birth weight infants (birth weight less than 2500g) zero otherwise.
In order to adjust for confounding we selected a set of potential baseline variables for the exposure-mediator, mediator-outcome and exposure-outcome associations. These are maternal age as exposure-mediator confounder centered at 33 years old. Child's year of birth was considered as an exposure-outcome confounder (centered at 2009) as proxy of the specific year's situation such as pollution, economics *etc.*.

There is not information about a causal relationship among parity and gestational age but significative associations are well documented [25]. This led us to consider gestational age as a baseline mediator-outcome confounder, centered at 37 weeks to consider term and preterm infants. We further consider smoking as a mediator-outcome confounder. Furthermore, it is well established that maternal asthma status is significantly associated with childhood asthma [45]. To take into account this association we considered maternal asthma as a risk factor for recurrent wheezing and estimating a crude Odds Ratio of 1.64 (CI 95% 1.03-2.61). All these decisions led to the final DAG represented in Figure 4.3.

Among the 4124 children participating to NINFEA (at May 2013) we selected 3392 children with complete records (regarding exposure, outcome, mediator and covariates): 75,47% were first born while 4.70% were under weight (<2500g). The prevalence of recurrent wheezing or asthma in the dataset was very low (5.48%). Demographic information for this sample are encoded in Table 4.1.

		Birth Weight		Parity	
	Overall (n=3392) No. (%)	$<$ 2500gr (n=160) No. (%)	$\geq$ 2500gr (n=3232) No. (%)	1° Child (n=2560) No. (%)	2° or more (n=832) No. (%)
Rec. Wheezing or Asthma					
Yes	186(5.48)	12(7.5)	174(5.38)	99(3.87)	87(10.46)
No	3206(94.52)	148(92.50)	3058(94.62)	2461(96.13)	745(89.54)
Child Year of Birth					
2005-2006	437(12.88)	21(13.13)	416(12.87)	345(13.48)	92(11.06)
2007-2008	879(25.91)	43(26.88)	836(25.87)	691(26.99)	188(22.60)
2009-2010	987(29.10)	43(26.88)	944(29.21)	743(29.02)	244(29.33)
2011-2012	1089(32.10)	53(33.13)	1036(32.05)	781(30.51)	308(37.02)
Smoke					
Yes	263(7.75)	10(6.25)	253(7.83)	222(8.67)	41(4.93)
No	3129(92.25)	150(93.75)	2979(92.17)	2338(91.33)	791(95.07)
Gestational age (weeks)					
19-23	12(0.35)	2(1.25)	10(0.31)	8 (0.31)	4(0.48)
24-28	11(0.32)	2(1.25)	9(0.28)	7(0.27)	4(0.48)
29-33	28(0.83)	15(9.38)	13(0.40)	21(0.82)	7(0.84)
34-38	602(17.75)	92(57.50)	510(15.78)	428(16.72)	174(20.91)
39-43	2737(80.69)	49(30.63)	2688(83.17)	2095(81.84)	642(77.16)
$>$43	2(0.06)	0(0.00)	2(0.06)	1(0.04)	1(0.12)
Maternal age (years)					
$<$20	4(0.12)	1(0.63)	3(0.09)	4(0.16)	0(0.00)
20-24	82(2.42)	3(1.88)	79(2.44)	73(2.85)	9(1.08)
25-29	595(17.54)	23(14.37)	572(17.70)	521(20.35)	74(8.89)
30-34	1503(44.31)	66(41.25)	1437(44.46)	1180(46.09)	323(38.82)
35-39	1004(29.60)	55(34.38)	946(29.36)	653(25.51)	351(42.19)
40-44	198(5.84)	12(7.50)	186(5.75)	123(4.80)	75(9.01)
$>$44	6(0.18)	0(0.00)	6(0.19)	6(0.23)	0(0.00)

Table 4.1: Distribution of recurrent wheezing or asthma, child year of birth, smoke, maternal age at birth and gestational age by birth weight and parity in the NINFEA sample

4.2.1 Methods

To study the effect of high parity on recurrent wheezing or asthma when low birth weight plays the role of mediator we used the Mediation analysis approach in the counterfactual framework defined in § 3.2. Let us denote X the exposure high parity, Y the outcome regarding the occurrence of recurrent wheezing or asthma and M the mediator LBW. With $\mathbf{c}_1$ we meant maternal age, with $\mathbf{C}_2$ we meant gestational age and smoking and with C_3 we meant child's year of birth. With A we meant maternal asthma. Given that the outcome considered was binary, we could choose between different scales of effect measures subsection 3.2.3. In the following we will use risk ratios. One way to measure these conditional mediation effects consists in estimating a regression model for every endogenous variable in the DAG [2]. For example, for the DAG in Figure 4.3 we have the following associational models:

$$logit\ P(M = 1|x, c_1, \mathbf{c}_2) = \beta_0 + \beta_x x + \beta_1 c_1 + \beta_2' \mathbf{c}_2 \tag{4.3}$$

$$logit\ P(Y = 1|x, m, \mathbf{c}_2, c_3, a) = \theta_0 + \theta_x x + \theta_m m + \theta_{xm} xm + \theta_2' \mathbf{c}_2 + \theta_3 c_3 + \theta_4 a \tag{4.4}$$

If the outcome is rare in all strata of exposure, mediator and confounders, then risk ratios approximate odds ratios [81, 78]. If the models above are correctly specified and the assumptions stated in § 3.2 hold, we can identify the $CDEs$ from the data as

$$\begin{aligned} CDE_{\mathbf{c}}(m) &= \frac{P(Y = 1|X = 1, M = m, \mathbf{c}_2, c_3, a)}{P(Y = 1|X = 0, M = m, s\mathbf{c}_2, c_3, a)} \\ &= \frac{e^{\theta_0+\theta_x+\theta_m m+\theta_{xm} m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}}{1 + e^{\theta_0+\theta_x+\theta_m m+\theta_{xm} m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}} \cdot \frac{1 + e^{\theta_0+\theta_m m++\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}}{e^{\theta_0+\theta_m m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}} \\ &\cong \frac{e^{\theta_0+\theta_x+\theta_m m+\theta_{xm} m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}}{e^{\theta_0+\theta_m m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}} \\ &= e^{\theta_x+\theta_{xm} m}. \end{aligned} \tag{4.5}$$

Assuming a rare outcome [2], the above identification is obtained adopting the following approximation

$$\frac{e^{\theta_0+\theta_x+\theta_m m+\theta_{xm} m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}}{1 + e^{\theta_0+\theta_x+\theta_m m+\theta_{xm} m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}} \cong e^{\theta_0+\theta_x+\theta_m m+\theta_{xm} m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}.$$

For simplicity of notation, hereafter, we will denote
$P(Y = y|X = x, M = m, \mathbf{C}_2 = \mathbf{c}_2, C_3 = c_3, A = a)$ as $P(Y = y|x, m, \mathbf{c}_2, c_3, a)$ and $P(M = m|X = x, C_1 = c_1, \mathbf{C}_2 = \mathbf{c}_2)$ as $P(M = m|x, c_1, \mathbf{c}_2)$. Under the direct and indirect identifiability assumptions and assuming a rare outcome, we can identify the $PNDE$ and the $TNIE$ as

$$
\begin{aligned}
PNDE_{\mathbf{c}} &= \frac{\sum_m P(Y=1|1,m,\mathbf{c}_2,c_3,a)\ P(M=m|0,c_1,\mathbf{c}_2)}{\sum_m P(Y=1|0,m,\mathbf{c}_2,c_3,a)\ P(M=m|0,c_1,\mathbf{c}_2)} \\
&= \frac{\sum_m \frac{e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}{1+e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}\ P(M=m|0,c_1,\mathbf{c}_2)}{\sum_m \frac{e^{\theta_0+\theta_m m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}{1+e^{\theta_0+\theta_m m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}\ P(M=m|0,c_1,\mathbf{c}_2)} \\
&\cong \frac{\sum_m e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=m|0,c_1,\mathbf{c}_2)}{\sum_m e^{\theta_0+\theta_m m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=m|0,c_1,\mathbf{c}_2)} \\
&= \frac{e^{\theta_0+\theta_x+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=0|0,c_1,\mathbf{c}_2)+e^{\theta_0+\theta_x+\theta_m+\theta_{xm}+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=1|0,c_1,\mathbf{c}_2)}{e^{\theta_0+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=0|0,c_1,\mathbf{c}_2)+e^{\theta_0+\theta_m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=1|0,c_1,\mathbf{c}_2)} \\
&= \frac{e^{\theta_0+\theta_x+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ \frac{1}{1+e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}+e^{\theta_0+\theta_x+\theta_m+\theta_{xm}+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ \frac{e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}{1+e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}}{e^{\theta_0+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ \frac{1}{1+e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}+e^{\theta_0+\theta_m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ \frac{e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}{1+e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}} \\
&= \frac{e^{\theta_0+\theta_x+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}+e^{\theta_0+\theta_x+\theta_m+\theta_{xm}+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}{e^{\theta_0+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}+e^{\theta_0+\theta_m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}} \\
&= e^{\theta_x}\ \frac{1+e^{\theta_m+\theta_{xm}+\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}{1+e^{\theta_m+\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}} \qquad (4.6)
\end{aligned}
$$

and

$$
\begin{aligned}
TNIE_{\mathbf{c}} &= \frac{\sum_m P(Y=1|1,m,\mathbf{c}_2,c_3,a)P(M=m|1,c_1,\mathbf{c}_2)}{\sum_m P(Y=1|1,m,\mathbf{c}_2,c_3,a)P(M=m|0,c_1,\mathbf{c}_2)} \\
&= \frac{\sum_m \frac{e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}{1+e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}\ P(M=m|1,c_1,\mathbf{c}_2)}{\sum_m \frac{e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}{1+e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}}\ P(M=m|0,c_1,\mathbf{c}_2)} \\
&\cong \frac{\sum_m e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=m|1,c_1,\mathbf{c}_2)}{\sum_m e^{\theta_0+\theta_x+\theta_m m+\theta_{xm}m+\theta_2'\mathbf{c}_2+\theta_3c_3+\theta_4a}\ P(M=m|0,c_1,\mathbf{c}_2)} \\
&= \frac{\frac{1}{1+e^{\beta_0+\beta_x+\beta_1c_1+\beta_2'\mathbf{c}_2}}+e^{\theta_m+\theta_{xm}}\ \frac{e^{\beta_0+\beta_x+\beta_1c_1+\beta_2'\mathbf{c}_2}}{1+e^{\beta_0+\beta_x+\beta_1c_1+\beta_2'\mathbf{c}_2}}}{\frac{1}{1+e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}+e^{\theta_m+\theta_{xm}}\ \frac{e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}{1+e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}} \\
&= \frac{1+e^{\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}}{1+e^{\beta_0+\beta_x+\beta_1c_1+\beta_2'\mathbf{c}_2}}\cdot\frac{1+e^{\theta_m+\theta_{xm}+\beta_0+\beta_x+\beta_1c_1+\beta_2'\mathbf{c}_2}}{1+e^{\theta_m+\theta_{xm}+\beta_0+\beta_1c_1+\beta_2'\mathbf{c}_2}} \qquad (4.7)
\end{aligned}
$$

When the interaction term θ_{xm} is zero, the Pure natural direct effect will be equal to the controlled direct effect

$$
CDE_{\mathbf{c}}(m) = e^{\theta_x}
$$

$$PNDE_{\mathbf{c}} = e^{\theta_x} \frac{1 + e^{\theta_m+\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1 + e^{\theta_m+\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}} = e^{\theta_x}$$

as we intuitively described in subsection 3.2.2.

4.2.2 Mediated interactive effect

In this section we will identify the alternative mediation effects defined in § 4.2.2 for the DAG in Figure 4.3. Let us suppose that the identifiability assumptions for natural effects are verified and that the models (4.3) and (4.4) are correctly specified, for the three-way decomposition in (3.37) we calculated:

$$\begin{aligned}
TNDE_{\mathbf{c}} &= \frac{\sum_m P(Y=1|1,m,\mathbf{c}_2,c_3,a)\ P(M=m|1,c_1,\mathbf{c}_2)}{\sum_m P(Y=1|0,m,\mathbf{c}_2,c_3,a)\ P(M=m|1,c_1,\mathbf{c}_2)} \\
&\cong \frac{\sum_m e^{\theta_0+\theta_x+\theta_m m+\theta_{xm} m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ P(M=m|1,c_1,\mathbf{c}_2)}{\sum_m e^{\theta_0+\theta_m m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ P(M=m|1,c_1,\mathbf{c}_2)} \\
&= \frac{e^{\theta_0+\theta_x+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{1}{1+e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}} + e^{\theta_0+\theta_x+\theta_m+\theta_{xm}+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1+e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}}{e^{\theta_0+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{1}{1+e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}} + e^{\theta_0+\theta_m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1+e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}} \\
&= \frac{e^{\theta_0+\theta_x+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a} + e^{\theta_0+\theta_x+\theta_m+\theta_{xm}+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{e^{\theta_0+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a} + e^{\theta_0+\theta_m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}} \\
&= e^{\theta_x} \frac{1 + e^{\theta_m+\theta_{xm}+\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1 + e^{\theta_m+\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}
\end{aligned} \tag{4.8}$$

and

$$\begin{aligned}
PNIE_{\mathbf{c}} &= \frac{\sum_m P(Y=1|0,m,\mathbf{c}_2,c_3,a)P(M=m|1,c_1,\mathbf{c}_2)}{\sum_m P(Y=1|0,m,\mathbf{c}_2,c_3,a)P(M=m|0,c_1,\mathbf{c}_2)} \\
&\cong \frac{e^{\theta_0+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{1}{1+e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}} + e^{\theta_0+\theta_m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1+e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}}{e^{\theta_0+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{1}{1+e^{\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}} + e^{\theta_0+\theta_m+\theta_2' \mathbf{c}_2+\theta_3 c_3+\theta_4 a}\ \frac{e^{\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1+e^{\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}}} \\
&= \frac{1 + e^{\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1 + e^{\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}} \cdot \frac{1 + e^{\theta_m+\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1 + e^{\theta_m+\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}}.
\end{aligned} \tag{4.9}$$

The term K, measure of interaction in a multiplicative scale, will then be:

$$K_{\mathbf{c}} = \frac{TNDE_{\mathbf{c}}}{PNDE_{\mathbf{c}}} = \frac{1 + e^{\theta_m+\theta_{xm}+\beta_0+\beta_x\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1 + e^{\theta_m+\theta_{xm}+\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}} \cdot \frac{1 + e^{\theta_m+\beta_0+\beta_1 c_1+\beta_2' \mathbf{c}_2}}{1 + e^{\theta_m+\beta_0+\beta_x+\beta_1 c_1+\beta_2' \mathbf{c}_2}} \tag{4.10}$$

4.2.3 Results

The estimated odds ratio relating recurrent wheezing or asthma to high parity adjusted for maternal age, gestational age, maternal smoking, maternal asthma and child's year of birth was 3.22 (CI 95% 2.35-4.40). The estimated odds ratio for recurrent wheezing or asthma on low birth weight adjusted for parity, maternal age, gestage and smoking was 1.38 (CI 95% 0.72-2.64). On the other hand the estimated adjusted odds ratio (adjusted for maternal age) relating low birth weight on high parity was 0.59 (CI 95% 0.39-0.90).

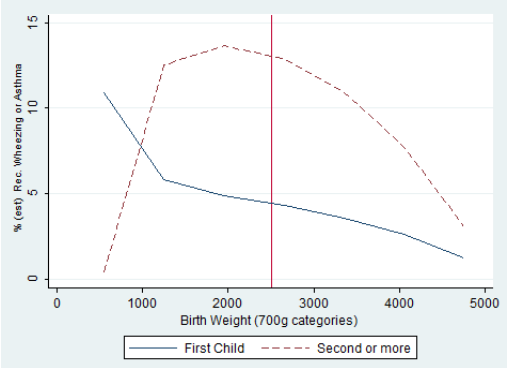


Figure 4.4: Estimated percentages of recurrent wheezing or asthma by birth weight (700g categories) and parity, NINFEA sample Italy May 2013.

Figure 4.4 represents a graphical intersection between birth weight and high parity on wheezing in the NINFEA sample: for very low birth weight children, the outcome's proportion (estimated by a polynomial regression model) of high parity is smaller than for first born children.

	OR (95% CI)
Normal Birth Weight (95.3%)	3.44 (2.49-4.76)
Low Birth Weight (4.7%)	1.06 (0.22-5.21)
p-value for interaction	0.154

Table 4.2: Estimated adjusted Odds Ratios of wheezing on high parity by birth weight category

Adjusted odds ratio of recurrent wheezing or asthma for high parity stratified by LBW are reported in Table 4.2. Interaction between high parity and low birth weight was also considered to take in account the evidence presented in Figure 4.4. Given that the general harmful association of high parity on recurrent wheezing or asthma, adjusted for confounders was 3.22 (CI 95% 2.35-4.40), the results in Table 4.6 point towards an harmful association just in the normal birth weight group. As stressed by VanderWeele *et al.* (2012) in [81], even if not significant, this apparent paradoxical difference can be interpreted as evidence of interaction

	Estimate (95% CI)
$CDE_{\mathbf{c}}(0)$	3.32 (2.43-4.53)
$CDE_{\mathbf{c}}(1)$	1.02 (0.27-3.68)
$PNDE_{\mathbf{c}}$	3.11 (2.26-4.17)
$TNIE_{\mathbf{c}}$	1.01 (0.98-1.03)
$TCE_{\mathbf{c}}$	3.14 (2.28-4.2)

Table 4.3: Mediation effects estimates for a mean individual (a child born in 2009 of 39 gestational weeks from a 33 years old mother non smoker and non asthmatic) NINFEA sample May 2013. The effects $CDE(0)$ and $CDE(1)$ refer to controlled direct effects when birth weight is set to normal birth weight and low birth weight respectively.

between exposure and mediator which is likely to be a consequence of unmeasured mediator-outcome confounding.

Partitioning the causal effect

In order to estimate mediation effects defined in § 4.2.1 we used the approximate analytical formulas of Ananth and VanderWeele (2011) applied to the framework of logistic regression when the outcome is rare [2]. The conditional mediation effects defined in the method section are referred to a mean individual that is a child born in 2009 of 39 gestational weeks from a 33 years old mother non smoker and non asthmatic.
The mediation effects defined in § 4.2.1 are estimated for the NINFEA sample by maximum likelihood estimation from the models for M and Y defined in (4.3) and (4.4). The relevant estimates were then inserted in the formulas (3.28)-(3.31). Standard errors were obtained via 10000 bootstrap sampling using the bias-corrected methods since there was evidence of non-normality. Results are displayed in Table 4.3.

According to Table 4.3, there is evidence of an harmful causal association of parity on recurrent wheezing or asthma with a risk ratio of being born as second or more child equal to 3.14 (CI 95% 2.28-4.2). The total effect is almost entirely attributed to the direct path, while the total natural indirect effect shows no mediated effect via birth weight with $TNIE = 1.01$ (CI 95% 0.98-1.03). From $CDEs$, setting each child to be normal birth weight, the direct effect of parity on recurrent wheezing or asthma will be strong and significant. On the other hand, setting each child to be low birth weight and hence, more at risk, the exposure does not seem associated directly with the outcome: the exposure seems to act as a risk factor just in the normal birth weight intervention group, what is meant to be the least at risk. Moreover, the results shown in Table 4.3, point towards evidence of no indirect effect form the exposure to the outcome thorough the mediator. However, an indirect effect different from one is not required for the birth weight paradox to

	Estimate (95% CI)
$TNDE_{\mathbf{c}}$	3.22 (2.32;4.26)
$PNIE_{\mathbf{c}}$	0.98 (0.92;1.00)
$K_{\mathbf{c}}$	1.04 (1.00;1.11)

Table 4.4: Estimation of the causal interaction between high parity and LBW on recurrent wheezing or asthma and mediation effects estimates accounting for a mediated interactive effect as part of the direct effect.

arise. In fact, Hernndez-Diaz *et al.* (2006) in [28], exhibit different scenarios able to exhibit these odd results. One of these is a situation where no indirect effect is present, *i.e.* no arrow from the mediator to the outcome. In this dissertation we will describe situations where the unmeasured confounding, capable of producing these paradoxical results, might be also responsible for the indirect effect to disappear.

Another plausible explanation might be the way in which the interaction was accounted for. In order to test this hypothesis we also decomposed the TCE as in Equation (3.37). According to Table 4.4, there is not evidence to support this hypothesis: not accounting for interaction in the indirect effect does not affect the meaning of the $PNIE$ which remains not significantly different from one.

VanderWeele's approaches to the paradox

This practice of considering birth weight as a mediator in the causal pathway between two variables has come under critique by various authors because it often produces paradoxical results such as in Table 4.2 and Table 4.3. In § 4.1 we discussed one method proposed by VanderWeele that might be capable of dealing with this phenomenon. In this section we will apply this method to the NINFEA dataset for the case of a rare outcome.

According to statistical (backward stepwise selection) and biological association we selected the following set of low birth weight's determinants (DET): foreign status, child's sex, maternal eclampsia, maternal height and weight (before pregnancy), maternal hay fever and finally maternal hypertension or preeclampsia before or during pregnancy. The variable foreign status was coded as zero if the country of delivery corresponded to maternal country, one otherwise. In Table 4.5 we reported the regression estimated parameters of the logistic model for LBW on determinants $DETs$ and confounders maternal age, smoking status and gestational age.

The first approach proposed by VanderWeele consists of conditioning on the estimated risk of being LBW predicted by the baseline covariates and mediator's determinants reported in Table 4.5. After fitting the model for M, we predicted the probabilities of being low-birth-weight for every child in the study. Using a cut-point

Variable	OR	95% CI
Foreign status	1.98	$0.94 - 4.16$
Sex (female)	1.96	$1.37 - 2.81$
Eclampsia	2.58	$1.15 - 5.8$
Height (cm)	0.95	$0.93 - 0.98$
Weight (kg)	0.98	$0.96 - 1.00$
Hay fever	0.5	$0.26 - 0.95$
Hypertension/preeclampsia	3.12	$1.77 - 5.51$
$PseudoR^2$		0.1923

Table 4.5: Logistic regression estimated parameters for Low Birth Weight. Foreign status was coded as zero if the country of delivery corresponded to maternal country, one otherwise. Sex was coded as 0 for males and 1 for females.

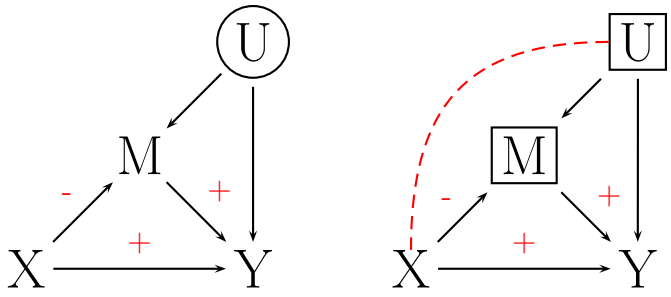
corresponding to its 95th centile, we classified as "at risk of LBW" (denoted H) those above this cut-point. Among the whole population, we identified 169 infants as high risk of being LBW where 72% were first born, 36% were under weight and 7.7% were wheezing. According to subsection 4.1.1, for low risk infants $H = 0$, the adjusted odds ratio relating recurrent wheezing or asthma to high parity was 3.26 (CI 95% 2.38-4.47) while, for high-risk-infants ($H = 1$), we obtained 1.22 (CI 95% 0.35-4.22), again not statistically different from one. As described in methods, these results are not affected by collider bias (unlike those reported in Table 4.2). Nevertheless, they point out the same conclusions. However, as described by VanderWeele, this method is highly affected by the predictive strength of the model given that another set of covariates, with bigger predictive power, will produce different results. Furthermore, by this method, we will not be able to distinguish between direct and indirect effects which could be done with the second method proposed by VanderWeele.

A possible explanation to the paradox

The previous section supports the presence of unmeasured confounding between mediator and outcome as an explanation of the apparent no associated effect of high parity on recurrent wheezing or asthma in the low birth weight group, the group the should be the most at risk. It has been suggested [81, 28] that this unmeasured common cause between mediator and outcome might likely be birth defects or malnutrition. There variables are in fact difficult to measure and analyses are usually not controlled for them. In the example of maternal smoking as a risk factor for infant mortality mediated by birth weight, VanderWeele (2012) in [81] suggests that, for smoker mothers with LBW infants, low birth weight might be a consequence of either smoking or of a birth defect. For non smoker mothers who give birth to LBW infants, some other causes have to be present. In fact, if LBW is not a consequence of the mother's smoking status, some worst risk factors have

to operate leading the unmeasured variable U being more common in unexposed subject. The same concept can be adopted for the causal relation of high parity on recurrent wheezing or asthma partially mediated by LBW. Overall, high parity is strongly and harmfully associated with recurrent wheezing or asthma. On the other hand, stratum specific effects did show an harmful effect only in the group that should be the less at risk while they are not associated for low birth weight infants. This can be explained by the presence of an unmeasured variable U, birth defects or malnutrition, more common in high parous children.

In this subsection we will describe graphically, the relational assumptions between exposure, outcome, mediator and unmeasured confounder U capable of masking the indirect effect. The following considerations will apply only if there are no interactions (between exposure, outcome, mediator and unmeasured confounder) and if all relational assumptions are linear in the coefficients.



(a) (b)

Figure 4.5: a) DAG figuring the unmeasured confounding U that might affect both mediator and outcome and the conditional associations found in the NINFEA sample; b) spurious path opened after conditioning on a collider M and after the potential condition on U

According to the associations found in § 4.2.3, we can draw the conditional relations encoded in 4.5a that are: high parity is negatively associated with low birth weight and positively associated with recurrent wheezing or asthma. The effect of LBW on recurrent wheezing or asthma of 1.38 (expressed as an OR) given in § 4.2.3, will be biased due to the unmeasured variable U. However, according to the literature [8][66], low birth weight should be strongly and harmfully associated with the outcome.

As we previously saw in subsection 1.2, conditioning on the mediator will open a spurious path from the exposure to the outcome through U. If we could condition on U, this bias path will be blocked. The indirect effect will then be negative as given by the hypothetical product of the negative effect of X on M and the positive effect of M on X (§ 3.1.1).

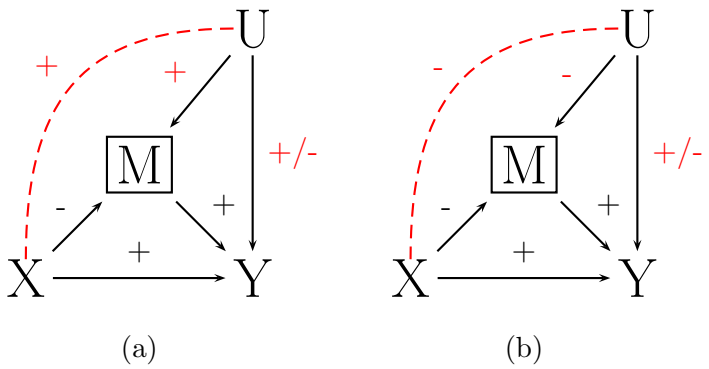


Figure 4.6: DAG figuring the spurious path opened after conditioning on a collider M when: a) U positively affects M and b) U negatively affects M

However in reality, we cannot condition on U and hence, we cannot block this new pathway arose after conditioning on M. Let us suppose that U is positively associated with the mediator as shown in 4.6a. From the collider bias rules, conditioning on M will create a positive association between X and U, independently from the effect of U on Y. We can define the bias affecting the indirect effect as the positive effect arising from the product of the consecutive pathways effects $X - U \rightarrow M \rightarrow Y$ (given that indirect is considered any pathway that goes through the mediator). Thus, the positive unbiased indirect effect should be summed to the negative bias obtained via the spurious pathways $X - U \rightarrow M \rightarrow Y$. To mask the indirect effect the negative bias should be as big as the unbiased indirect effect is different from one. Let us suppose that the real effect of LBW on the outcome is 1.5, closer to those obtained in [8][66]. In a hypothetical situation, involving only linearities and non interaction, the unbiased indirect effect of X on Y should be closer to 0.89 leading to a bias closer to 0.13.
On the other hand, assuming a negative association between U and M such as in 4.6b, conditioning on M will open a spurious negative path from X to U. The resulting bias affecting the indirect effect will be again positive as given by the hypothetical product of the consecutive pathways effects $X - U \rightarrow M \rightarrow Y$. Again, to mask the indirect effect the bias arising after conditioning on M should be closer to 0.13. However, by this method, we are not able to distinguish between 4.6a and 4.6b but we can at least investigate on the magnitude of the bias.

It is important to notice that this type of analysis is usual in linear model but ideally it can also work in logistic models for rare outcomes and mediator with no interactions between any variable in the model and if all relational assumptions are linear in the coefficients.

Sensitivity Analysis: Rare Outcome

Conditioning on the risk of being LBW instead of the mediator itself will lead to estimate the total effect of the exposure on the outcome for high and low risk infants. However, if we still wish to estimate direct and indirect effects, conditioning on M will not be unavoidable. Differently from subsection 4.2.3, here we will not assume the absence of non-linearities and interactions between any variable in the model. This section examines the situation in which an unmeasured variable, that affects both mediator and outcome, may have to affect both of them to invalidate the qualitative conclusions made on controlled and natural effects. In particular, in this section we will describe the bias formulas for sensitivity analysis for mediation effects introduced by VanderWeele in [78, 81]. In the first paper VanderWeele argues that, the estimates (4.5) to (4.7), will be biased if the assumption of no unmeasured-confounding between mediator and outcome does not hold. If the outcome is rare in every strata of exposure X, mediator M, covariates and if we can further assume that the outcome is rare in every strata of the unmeasured confounder, then we can define the bias affecting the mediation effects as the ratio between the estimand and the true effect [78]:

$$Bias[CDE_{\mathbf{c}}(m)] = \frac{\frac{P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})}{P(Y=1|X=0,M=m,\mathbf{C}=\mathbf{c})}}{\frac{E[Y(1,m)|\mathbf{c}]}{E[Y(0,m)|\mathbf{c}]}}$$

$$Bias[PNDE_{\mathbf{c}}] = \frac{\frac{\sum_m P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=0,\mathbf{C}=\mathbf{c})}{\sum_m P(Y=1|X=0,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=0,\mathbf{C}=\mathbf{c})}}{\frac{E[Y(1,M(0))|\mathbf{c}]}{E[Y(0,M(0))|\mathbf{c}]}}$$

$$Bias[TNIE_{\mathbf{c}}] = \frac{\frac{\sum_m P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=1,\mathbf{C}=\mathbf{c})}{\sum_m P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})P(Y=1|X=1,M=m,\mathbf{C}=\mathbf{c})P(M=m|X=0,\mathbf{C}=\mathbf{c})}}{\frac{E[Y(1,M(1))|\mathbf{c}]}{E[Y(1,M(0))|\mathbf{c}]}}$$

Let U be a binary variable and let us suppose that the effect of U on Y is constant across strata of X, i.e. $P(Y|x,m,\mathbf{c},U=1)/P(Y|x,m,\mathbf{c},U=0)=\gamma$. Furthermore, let us suppose that there is not any other confounder between X and Y except $\mathbf{C}$ and between Y and M except U, then:

$$Bias[CDE_{\mathbf{c}}(m)] = \frac{1+(\gamma-1)P(U=1|1,m,\mathbf{c})}{1+(\gamma-1)P(U=1|0,m,\mathbf{c})} \tag{4.11}$$

where we will call $pi_{0m} = P(U=1|0,m,\mathbf{c}) = P(U=1|X=0,M=m,\mathbf{c})$ and $pi_{1m} = P(U=1|1,m,\mathbf{c}) = P(U=1|X=1,M=m,\mathbf{c})$. The parameter γ can be informally interpreted as the direct effect of U on Y. The hypothesis regarding $P(Y|x,m,\mathbf{c},U=1)/P(Y|x,m,\mathbf{c},U=0)=\gamma$ being constant across strata of X

would hold if the unmeasured variable U would not interact (on the multiplicative scale) with the effect of the exposure on the outcome. A further assumption, already required for natural effect to be identified, is the absence of intermediate confounding between mediator and outcome affected by exposure.

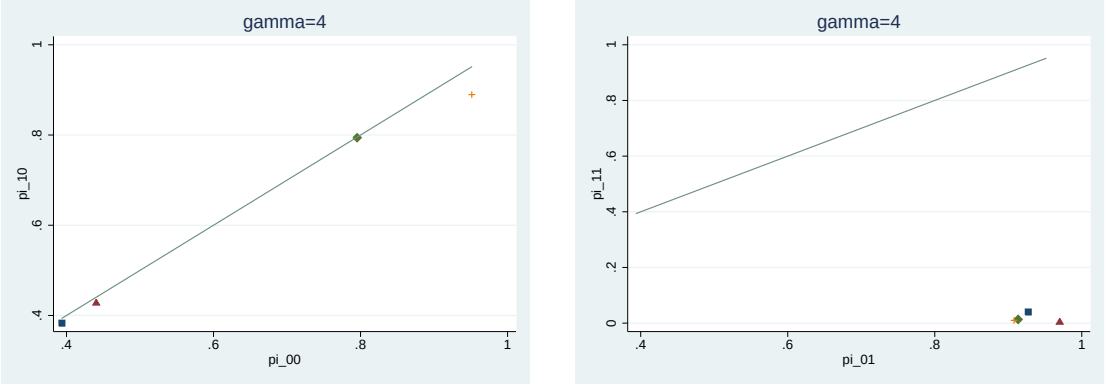
Once we specify the parameter γ and the hypothetical prevalence of U in every strata of exposure, mediator and confounders, we can simply obtain an unbiased estimator for CDE by dividing the potentially confounded estimate in Table 4.3 by (4.11).

Here we will focus on the following research question: if we could control for U, what degree of confounding is capable of showing a $CDE(1)$ bigger (or at least equal) than $CDE(0)$? And what will be the relative indirect effect?

For example, let us consider a severe confounding scenario where $\gamma = 4$ that is if $U = 1$ were to conditionally increases the probability of recurrent wheezing or asthma by a factor of four. The sensitivity analysis was performed setting the prevalence of U among the population to: $pi_{00} = 0.4$, $pi_{10} = 0.38$, $pi_{01} = 0.93$ and $pi_{11} = 0.05$ (square dots in 4.7a and 4.7b). In line with the earlier discussion of the likely unmeasured confounder, the prevalence of U among unexposed was set to be bigger than the prevalence of U among exposed (both in normal and low birth weight infants). The correct (possibly unbiased if the assumed values for these sensitivity parameters are suitable) controlled direct effect of high parity on recurrent wheezing or asthma if we could intervene setting each child to be normal birth weight will then be 3.37 while setting each child to be low birth weight will be 3.44. This severe degree of confounding is consistent with the results found by Basso *et al.* in [6, 5]. In 4.7a and 4.7b we plotted four possible combinations of the probabilities $pi_{00}, pi_{10}, pi_{01}, pi_{11}$ for $\gamma = 4$ capable of showing a corrected $CDE(0)$ smaller or at least equal than $CDE(1)$. Out of 10000 simulations, only the four combinations described in 4.7a and 4.7b correspond to an unbiased $CDE(0) \leq CDE(1)$. Here and later we will call respectively CDE^b and CDE^{unb} the biased and the corrected (unbiased) controlled direct effect. The number of results such that $CDE^{unb}(0) \leq CDE^{unb}(1)$ increased with the degree of the unmeasured confounding γ (see 4.8a, 4.8b for $\gamma = 5$ and 4.8c, 4.8d for $\gamma = 7$). However, the simulations with γ smaller than four, did not produce any solution such as $CDE(0) \leq CDE(1)$.

Sensitivity formulas can be obtained also for $PNDE$ and $TNIE$ by assuming further that $U \perp\!\!\!\perp X|C$:

$$Bias(PNDE_{\mathbf{c}}) = \frac{\sum_m [1 + (\gamma - 1)pi_{1m}]v_m P(m|0, \mathbf{c})}{\sum_m [1 + (\gamma - 1)pi_{0m}]v_m P(m|0, \mathbf{c})} \tag{4.12}$$



(a) (b)

Figure 4.7: Plot of the four quartet of values pi_{ij} with $i = 0, 1$ and $j = 0, 1$ capable of showing a $CDE(1)$ bigger than $CDE(0)$ for $\gamma = 4$ where in a) we have the values of pi_{00} and pi_{10} and in b) we have the values of pi_{01} and pi_{11}. Similar symbols correspond to the same quartet. Only values of pi_{ij} (with $i = 0, 1$ and $j = 0, 1$) that lie below the red line will be considered (for which $pi_{0j} > pi_{1j}$ with $j = 0, 1$)

$$Bias(TNIE_{\mathbf{c}}) = \frac{1}{Bias(PNDE_{\mathbf{c}})} \tag{4.13}$$

where $pi_{xm} = P(U = 1|X = x, m, \mathbf{c})$ and $v_m = \frac{E[Y|x,m,\mathbf{c},U=0]}{E[Y|x,\tilde{m},\mathbf{c},U=0]}$.

The corrected natural risk ratios can be obtained dividing the estimated risk ratios by the bias factor (4.12) and (4.13). From Equations (4.12) and (4.13) we can notice two important details: 1) the bias for the natural indirect effect is simply obtained as one over the bias for the natural direct effect, *i.e.* once e calculate the first we can easily obtain the latter; 2) the total causal effect is unconfounded by U

$$\begin{aligned} TCE_{\mathbf{c}} &= PNDE^{b}_{\mathbf{c}} \cdot TNIE^{b}_{\mathbf{c}} \\ &= PNDE^{unb}_{\mathbf{c}} \cdot Bias(PNDE_{\mathbf{c}}) \cdot TNIE^{unb}_{\mathbf{c}} \cdot Bias(TNIE_{\mathbf{c}}) \\ &= PNDE^{unb}_{\mathbf{c}} \cdot TNIE^{unb}_{\mathbf{c}} \end{aligned}$$

If $v_m = 1$ for all m, then

$$\begin{aligned} Bias(PNDE_{\mathbf{c}}) &= \frac{\sum_m [1 + (\gamma - 1)\, pi_{1m}]\, P(m|0, \mathbf{c})}{\sum_m [1 + (\gamma - 1)\, pi_{0m}]\, P(m|0, \mathbf{c})} \\ Bias(TNIE_{\mathbf{c}}) &= \frac{1}{Bias(PNDE_{\mathbf{c}})} \end{aligned}$$

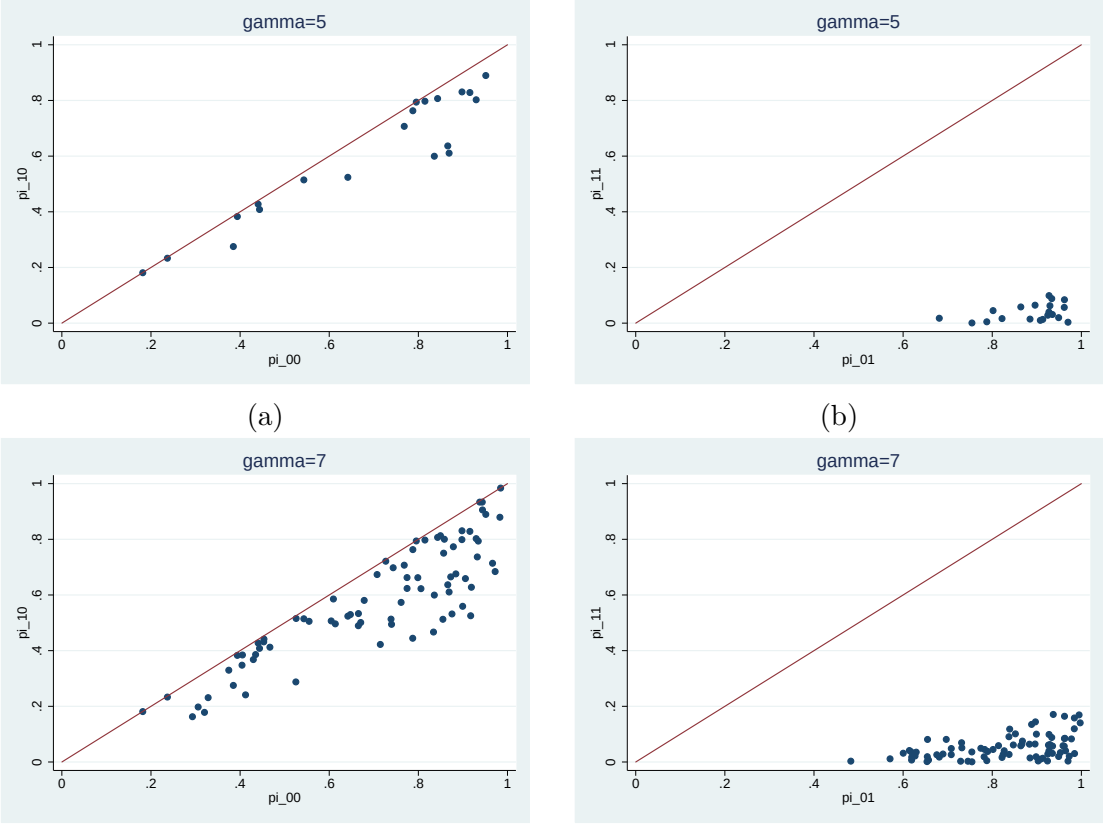


Figure 4.8: Plot of the combinations of values pi_{ij} with $i = 0, 1$ and $j = 0, 1$ capable of showing a $CDE(1)$ bigger than $CDE(0)$ for different values of γ where in a) we have the values of pi_{00} and pi_{10} and in b) we have the values of pi_{01} and pi_{11}. Only values of pi_{ij} (with $i = 0, 1$ and $j = 0, 1$) that lie below the red line will be considered (for which $pi_{0j} > pi_{1j}$ with $j = 0, 1$)

where $P(m|0, \mathbf{c})$ is the probability of the mediator among unexposed in the NINFEA dataset [80]. Considering the model in (4.3), the bias affecting natural direct and indirect effects will then be

$$Bias(PNDE_{\mathbf{c}}) = \frac{[1 + (\gamma - 1)\, pi_{10}] + [1 + (\gamma - 1)\, pi_{11}]\, e^{\beta_0 + \beta_1 c_1 + \beta_2' \mathbf{c}_2}}{[1 + (\gamma - 1)\, pi_{00}] + [1 + (\gamma - 1)\, pi_{01}]\, e^{\beta_0 + \beta_1 c_1 + \beta_2' \mathbf{c}_2}}$$
$$Bias(TNIE_{\mathbf{c}}) = \frac{1}{Bias(PNDE_{\mathbf{c}})}.$$

For $\gamma = 4$, $pi_{00} = 0.4$, $pi_{10} = 0.38$, $pi_{01} = 0.93$ and $pi_{11} = 0.05$, the corrected direct and indirect effects for a mean individual will be

$$PNDE^{unb}_{\mathbf{c}} = 3.11/\frac{[1+(4-1)\,0.38]+[1+(4-1)\,0.05]\,0.05}{[1+(4-1)\,0.4]+[1+(4-1)\,0.93]\,0.05} = 3.42$$
$$TNIE^{unb}_{\mathbf{c}} = 1.01 * Bias(PNDE_{\mathbf{c}}) = 0.92. \quad (4.14)$$

where $e^{\beta_0+\beta_1 c_1+\beta_2'\mathbf{c}_2} = 0.05$ for a mean individual that is a child born of 39 gestational weeks from a non smoking mother of 33 years old.

As described in subsection 4.2.3, for the case of models with only linear relationships, the indirect effect of high parity on recurrent wheezing or asthma through LBW (if we could condition on U such that $\gamma = 4$ and $pi_{00} = 0.4$, $pi_{10} = 0.38$, $pi_{01} = 0.93$ and $pi_{11} = 0.05$) should be protective.

4.3 Regular Outcome

Several factors can influence the occurrence of infants wheezing. One that has been investigated in relation to asthma is maternal parity as an indicator of both biological and environmental exposures [46]. However, there are no studies investigating the association between wheezing and high parity in younger children. Several authors adopt mediation to analyze the potential causal relation between exposure and outcome [38]. One possible key is via birth weight as this usually increases with maternal parity and is associated with lung function.

However, as we saw in the § 4.2, studying mediating effects via birth weight is complex. Here we will refer to a regular outcome as a not rare outcome.

In this section we investigate the potential mediating effect of birth weight in the relation between birth order and wheezing [42] in young children using data from the birth cohort Ninfea described in § 4.0.1.

4.3.1 Methods

We aim to estimate various measures of direct and indirect effects of an exposure, maternal parity, onto the outcome. Wheezing was assessed at the 6-month and the 18-month questionnaire, asking whether the child had episodes of wheezing or whistling in the chest in the first 6 months or between 6 and 18 months of life. The exposure was parity dichotomized to be zero for the first child and one otherwise. The mediator, reported by the mother at the 6-month questionnaire, was birth weight dichotomized to be one for low birth weight infants (birth weight less than 2500g) zero otherwise. Various potential confounders were considered: maternal age for the exposure-mediator relationship, gestational age and maternal smoking for the mediator-outcome relationship and child's year of birth as exposure-outcome

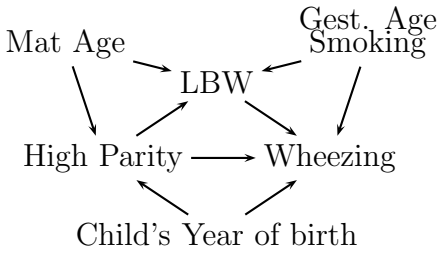


Figure 4.9: DAG representing the relational assumptions between high parity (child $\geq$1), low birth weight (birth weight <2500g), wheezing up to 18 months of age and potential confounders: child's year of birth , maternal age, gestational age and maternal smoking.

confounder as illustrated in Figure 4.9. Gestational age, also reported by the mother at the 6-month questionnaire, was centered at 37 weeks, maternal age at 33 years old and child's year of birth at 2009.

Mediation analysis aims to disentangle the causal effect of an exposure on an outcome in two different portions: the indirect and the direct effects. Differently from § 4.2, the outcome defined as at least one episode of wheezing up to 18 months of age is not rare. Throughout this section we will define mediation effects in the counterfactual framework as defined in subsection 3.2.3 in terms of odds ratio. In the light of these definitions, the $PNDE$ will measure the direct effect attributable to parity on wheezing via pathways not involving LBW while $TNIE$ will capture the effect of parity on wheezing through LBW and $CDE(1)$ is the controlled direct effect of high parity on wheezing if we could intervene setting each child to be low birth weight.

Differently from § 4.2.1, the definitions (3.19) to (3.20) are marginal definitions of mediation effects. These effects require the same identifiability assumptions defined in § 3.2. In order to estimate these mediation effects, we used fully parametric implementation of Pearl's mediation formula [53, 51] which is performed in the *gformula* command implemented in Stata 12 [12]. This command estimates causal effects by the g-computation procedure using Monte Carlo Simulation [60].

Comments

This practice of considering birth weight as a mediator in perinatal epidemiology, has come under critique by various authors because is likely to be affected by unmeasured confounding. In particular, it has been suggested that paradoxical results might arise as a consequence of collider bias when unmeasured confounding affects the mediator-outcome relationship (see Figure 4.10). In fact, if an unmeasured variable U influences both mediator and outcome, conditioning on the first will open a spurious path from the exposure to the outcome. VanderWeele *et al.* [81, 78] proposed different approaches to deal with this phenomenon: conditioning on the estimated risk of being LBW instead of the mediator itself as described in subsection 4.1.1 and conditioning on the mediator in combination with sensitivity analysis

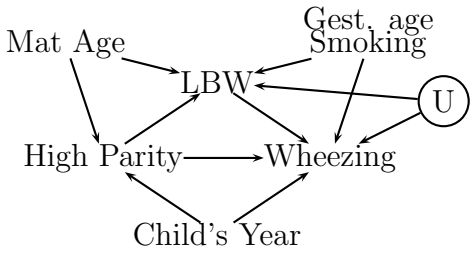


Figure 4.10: DAG representing the relational assumptions between high parity, low birth weight, wheezing, confounders and a potential unmeasured confounder U.

to examine the robustness of mediation estimands.

The first approach consists of conditioning on the estimated risk of being LBW when fitting a model for the outcome and hence obtaining unbiased estimates of $CDE(m)$. These estimated risks are predicted by baseline covariates $\mathbf{C}$ and mediator determinants that we will call Det.

VanderWeele suggested to predict individuals probabilities by the logistic model for M, $P(M = 1|\ \mathbf{C} = \mathbf{c}, Det = d)$ and then define a new variable H such that it is one for children who have predicted probabilities (of being LBW) above the 95th percentile, and zero otherwise. Because H is a function of determinants Det and confounders $\mathbf{C}$, conditioning on it does not imply conditioning on the mediator and hence generating collider bias. However, this methods is highly affected by the choice of conditioning variables and by the prevalence of the mediator. Furthermore, it does not allow to evaluate mediation effects but only the effect of high parity on wheezing among infants at low or high risk of being low birth weight.

The second approach addresses the problem via sensitivity analysis. VanderWeele in [81] and [78] defines formulas for the bias affecting the controlled direct, the indirect and the direct effects of X in the presence of an unmeasured mediator-outcome confounder such as U. Since these formulas are designed for rare outcomes, they will not be discussed in this paper.

Instead, we propose Monte Carlo simulations of two different situations capable of illustrating the bias.

4.3.2 Results

The analyses below involve 3,392 NINFEA children (out of 4124 children included in the NINFEA database version 2013.03) who have complete information on all relevant variables. Among them, 75.5% were first born while 4.7% were under weight (<2500g). The prevalence of wheezing was 16.6%. The odds ratio relating wheezing to high parity adjusted for maternal age, gestational age, maternal smoking and child's year of birth was 2.32 (CI 95% 1.89-2.83). The odds ratio relating wheezing to low birth weight adjusted for parity, maternal age, gestational age and smoking was 0.94 (CI 95% 0.60-1.47). On the other hand the adjusted odds ratio (adjusted

for maternal age) relating low birth weight to high parity was 0.59 (CI 95% 0.39-0.90). In Table 4.6 we report the adjusted odds ratio relating wheezing to high parity stratified by birth weight. Statistical interaction between high parity and low birth weight was also considered.

	OR (95% CI)
Normal Birth Weight (95.3%)	2.35 (1.91-2.88)
Low Birth Weight (4.7%)	1.56 (0.55-4.37)
p-value for interaction	0.4

Table 4.6: Adjusted Odds Ratios relating wheezing to high parity stratified by birth weight category

As discussed in [81], parity seems to have a harmful effect on wheezing only for normal-birth-weight infants. On the other hand, although there was no evidence of effect modification in the multiplicative scale, this association was weaker and not statistically different from one for low-birth weight infants. Basso *et al.* in [6, 5] suggest that this apparent difference between $CDEs$ may likely be a consequence of unmeasured $M - Y$ confounding. This situation will be analyzed in the Sensitivity analysis section.

The estimated controlled direct effects shown in Table 4.7 point out similar discrepancies: if each child were set to have a normal weight, the direct effect of parity on wheezing will be particularly harmful. On the other hand, if we could set each child to be low-birth-weight the direct effect will be towards the null. Finally the total causal effect seems to be entirely attributed to the direct path according to the $PNDE$ estimate.

In this section we investigate different scenarios that might be responsible to the apparent null association between parity and wheezing in low-birth-weight infants and also capable of masking the indirect effect.

Furthermore, mediation estimands rely on the assumptions stated in the methods section. In particular the assumption of no unmeasured mediator-outcome confounding is the most relevant here because its violation seems responsible for this paradoxical results.
To avoid the collider bias induced by such unmeasured mediator-outcome confounding we selected, in the NINFEA dataset, all potential birth weight's determinants not directly attributable to parity, able to produce the best predictive model for M. The predictive strength of this model was $R^2 = 23\%$. After fitting the model for M, we predicted the probabilities of being low-birth-weight for every child in the study. Using a cut-point corresponding to its 95th centile, we classified as "at risk of LBW" (denoted H) those above this cut-point. Among the whole

	OR (95% CI)
TCE	2.28 (1.86-2.8)
PNDE	2.27 (1.84-2.8)
TNIE	1.00 (0.98-1.05)
CDE(0)	2.29 (1.87-2.81)
CDE(1)	1.53 (0.5-4.72)

Table 4.7: Mediation effects estimated by g-computation, NINFEA sample May 2013.

population, we identified 169 infants as high risk of being LBW where 72% were first born, 37% were under weight and 18% had wheezing. Repeating the analyses stratifying by H we found that, for low-risk-infants ($H = 0$), the adjusted odds ratio relating wheezing to high parity was 2.39 (CI 95% 1.95-2.93) while for high-risk-infants ($H = 1$) we obtained 1.20 (CI 95% 0.50-2.89), again closer to one. As described in methods, these results are not affected by collider bias (unlike those reported in Table 4.6). Nevertheless, they point towards the same conclusions. However, as described by VanderWeele, this method is highly affected by the predictive strength of the model which in this case is poor. Furthermore, if the mediator is rare, he states that "these measure may not be an accurate reflection of the effect of the exposure for whom the intermediate will in fact develop".

Another possible situation might be the presence of unmeasured confounding between LBW and wheezing that is affected by high parity. In the next session we will introduce some sensitivity analysis able to examine this situation.

4.3.3 Sensitivity analysis

In this section we propose Monte Carlo simulations of two alternative settings capable of illustrating the type of bias that may affect our results. The first scenario is described by Figure 4.11 where unmeasured confounding affects the mediator-outcome relationship.

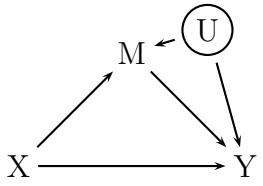


Figure 4.11: DAG illustrating a Mediation Mechanism affected by unmeasured mediator-outcome confounding

We generated X and U as binary independent random variables with prevalence 0.25 for X (as in the NINFEA sample) and 0.5 for U in order to investigate more

general scenarios. The mediator M and the outcome Y were generated from the following logistic regression models

$$logit(M = 1) = \gamma_0 + \gamma_X X + \gamma_U U;$$
$$logit(Y = 1) = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM + \beta_U U.$$

In order to investigate the sensitivity to the prevalence of M, γ_0 was allowed to vary in the set $\{-2, -1.4, 0.4, 2.2\}$. Here and after we will call p_M the baseline prevalence of M, *i.e.* the $exp(\gamma_0)/(1 + exp(\gamma_0))$. The adjusted odds ratio relating birth weight to high parity ($\exp(\gamma_x)$) was set to 0.4 to resemble the relational assumptions in the dataset. The coefficients γ_U and β_U were allowed to vary in $\{-0.7, 1, 2, 3\}$ in order to investigate both moderate and severe confounding bias settings. The baseline prevalence of Y was maintained at 0.2 setting β_0 to -1.4. The adjusted odds ratio relating wheezing to high parity ($\exp(\beta_x)$) was set to 2.2. According to [8] and [66], β_M was set to 1 to investigate a harmful association between LBW and wheezing. The term β_{XM} was set to zero in order to study whether the interaction shown in Table 4.6 is due only to unmeasured confounding and not to effect modification.

For every combination of parameters $\gamma_0, \gamma_U, \beta_U$, we simulated 1000 Monte Carlo datasets each composed of 10000 observations. For each combination we estimated the mediation effects in two different situations: including or not U in the g-computation formula. We called $TNIE_b$ the biased natural indirect effect estimated not including U and $TNIE_{unb}$ the unbiased natural indirect effect obtained including U in the g-computation formula. The same terminology will be used for pure natural direct effects and controlled direct effects.

In order to quantify the confounding bias problem we defined a new parameter $S := \gamma_u \cdot \beta_u$ as a crude measure of the hypothetical association between U and both M and Y. To represent how the confounding bias may affect the apparent differences between $CDEs$ we choose to plot two different lines, one for the mean of $CDE_b(0) - CDE_b(1)$ and one for the mean of $CDE_{unb}(0) - CDE_{unb}(1)$, for different values of S. We choose to plot the difference $CDE(0) - CDE(1)$ to better display and interpret the discrepancy between $CDEs$ due to the paradox. The value of zero will be the hypothetical cutoff, values lying over zero will corroborate the observed paradoxical results. On the other hand, negative differences will support the theory where the direct effect of high parity on wheezing will be bigger if we could intervene setting each child to be LBW instead of normal birth weight. From Figure 4.13a we can see that, for almost every degree of confounding, the dashed unbiased line describes only null or negative differences for which $CDE_{unb}(0) \leq CDE_{unb}(1)$. On the other hand, the solid line increases with the degree of confounding. It is interesting to notice that, the difference

found in the NINFEA sample ($2.28 - 1.53 = 0.76$) corresponds to an extreme degree of confounding. This is consistent with the results found by Basso *et al.* in [6, 5].

To evaluate how the confounding bias may mask the indirect effect, we choose to plot the ratio between $TNIE_{unb}$ and $TNIE_b$ as a measure of the bias affecting natural indirect effects. The reason for displaying differences between $CDEs$ and ratios for $TNIEs$ is to better display the paradox due to unmeasured confounding in the first case and to display the real bias affecting indirect effects in the latter. Figure 4.13b and 4.13c display the bias affecting natural indirect effects for different degrees of confounding and different values of p_M. As we can see from Figure 4.13b and 4.13c, the bias affecting natural indirect effects has a quadratic shape, it increases considerably with lower level of p_M and decreases for $p_M > 0.6$. According to Figure 4.13c, for low mediator prevalence and extreme unmeasured-confounding, $TNIE_{unb}$ should exceed $TNIE_b$ by more than 20%. From these simulations we can see that it would be possible to have an indirect effect of parity via birth weight under certain unmeasured confounding scenarios.
The Figure Figure 4.12 deals with a more complex setting that involves an intermediate confounder, *i.e.* a confounder that lies on the causal path from X to Y. For this setting identification of natural mediation effects is complex and required for example the additional assumption of no average intermediate confounder mediator interaction in their effect on Y, conditional on X and $\mathbf{C}$ [75].

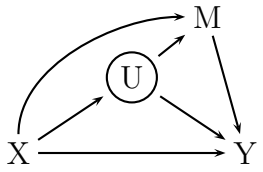


Figure 4.12: DAG illustrating a mediation mechanism with two mediators M and U, the first observed and the second unobserved

The binary variable X was generated as in the previous scenario. The mediator M, the outcome Y and the unmeasured mediator U were generated from the following logistic regression models

$$
\begin{aligned}
&logit(U = 1) = \alpha_0 + \alpha_X X; \\
&logit(M = 1) = \gamma_0 + \gamma_X X + \gamma_U U; \\
&logit(Y = 1) = \beta_0 + \beta_X X + \beta_M M + \beta_U U.
\end{aligned}
$$

The parameter α_x was allowed to vary between $\{-0.7, 1\}$ to investigate different $X - U$ associations. We choose to maintain the other parameters equal to the previous scenario.

For every combination of parameters $\gamma_0, \gamma_U, \beta_U, \alpha_X$, we simulated 100 Monte Carlo datasets each composed by 1000 observations. For each combination we estimated the mediation effects in two different situations: including or not U in the g-computation formula. We called NIE_b the biased natural indirect effect estimated not including U in the computation. As described in [11], we defined NIE_U, NIE_M, NIE_{UM} the unbiased indirect effects through U alone, M alone and through both U and M respectively. We further defined the unbiased controlled direct effect as the odds ratio

$$CDE(m) = \frac{E[Y(1, U(1), m)]/1 - E[Y(1, U(1), m)]}{E[Y(0, U(0), m)]/1 - E[Y(0, U(0), m)]}$$

that is the direct effect of X on Y when M is controlled to a specific level m and U arises naturally after setting X to x (one for the numerator and zero for the denominator).

As the previous scenario, Figure 4.13d and 4.13e show the biased and unbiased differences between controlled direct effects for different values of S.
As the case of one mediator, the difference between $CDEs$ shown in the NINFEA sample (0.76) corresponds to an extreme degree of confounding (for each combination of α_x considered). The only combination compatible with these results and capable of showing an indirect effect equal to one is $\alpha_x = 1$. Figure 4.13f shows the indirect effects NIE_b, NIE_U, NIE_M, NIE_{UM} by p_M for $\alpha_X = 1$, $\gamma_u = 3$ and $\beta_u = 3$. We can see that, for every prevalence p_M, the biased natural indirect effect seems closer to one than NIE_M. In our opinion it means that, in the case of two mediators, the real indirect effect via M alone should be smaller (and then protective) than the biased indirect effect shown in Table 4.7. On the other hand, for every prevalence p_M, NIE_U is always bigger than NIE_M, *i.e.* the unmeasured mediator U seems a stronger mediator than LBW alone.

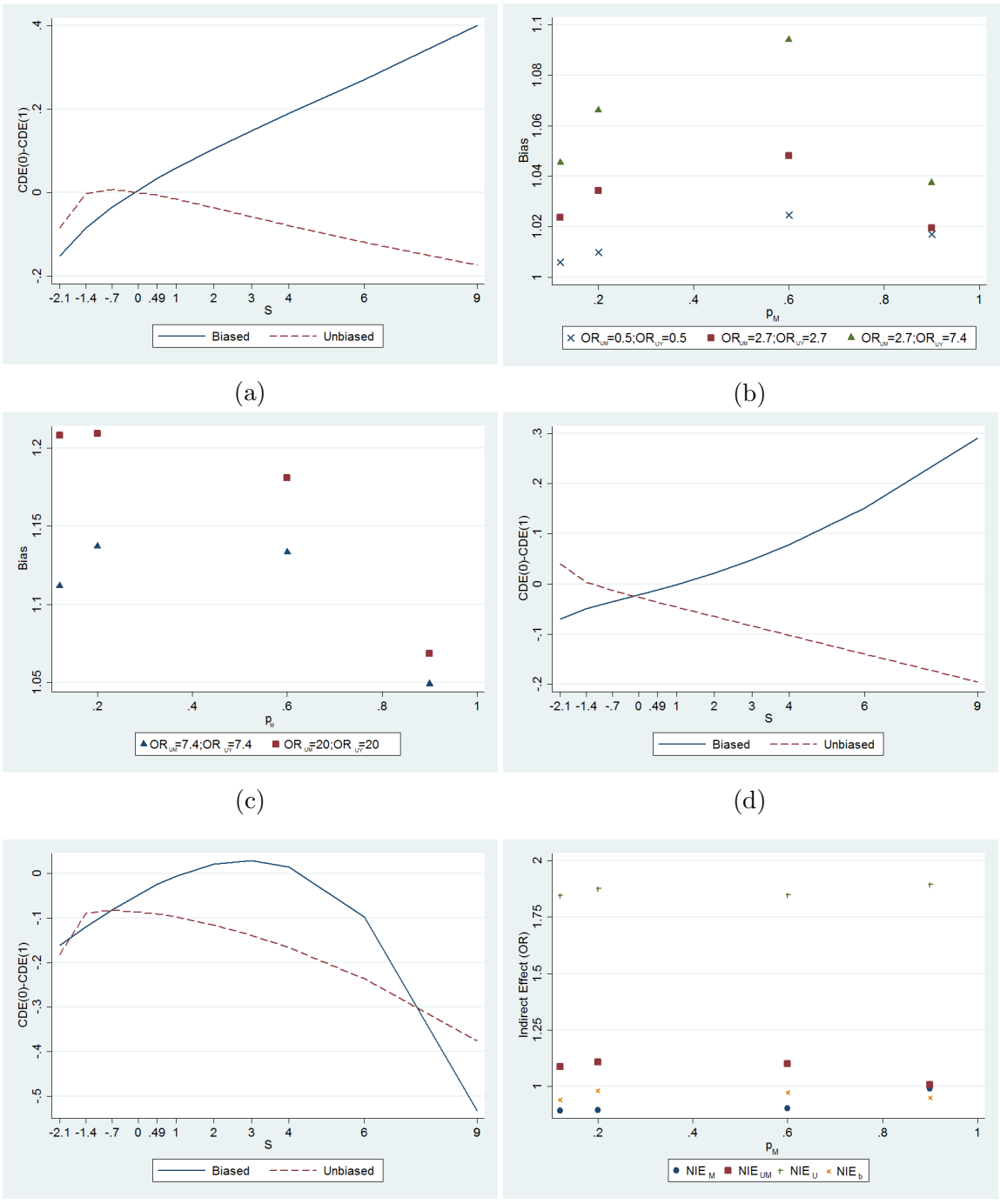


Figure 4.13: (a) Controlled direct effect difference (biased and unbiased) by unmeasured confounding strength S; (b) Bias affected indirect effect by p_M for different value of OR_{UM} and OR_{UY}; (c) Bias affected indirect effect by p_M for different value of OR_{UM} and OR_{UY}; (d) Controlled direct effect difference (biased and unbiased) by unmeasured confounding strength (S) for $\alpha_X = -0.7$; (e) Controlled direct effect difference (biased and unbiased) by unmeasured confounding strength (S) for $\alpha_X = 1$; (f) Indirect effects NIE_b, NIE_U, NIE_M, NIE_{UM} by p_M $\alpha_X = 1$, $\gamma_u = 3$ and $\beta_u = 3$

Chapter 5

Mediation as CoE

Many statistical analyses aim at a causal explanation of the data. In particular in epidemiology many studies have been conducted to understand when and if an exposure will cause a particular disease. Even in a court of law when we want to assess legal responsibility we usually refer to causality. But when we discuss about this topic is fundamental to specify the exact query we want to talk about. In § 2 we mentioned the differences between questions on the causes of observed effects and questions on the effects of observed causes. In this section we will describe a novel method capable of measure the causal effect of X on Y, ascribing to CoE questions, when we have additional information on a mediator.

For example, a court can claim that was Ann's taking the drug that was the *cause* of her death such as in the Example 2.0.1. This type of question is referred on the causes of a given effects ("CoE") and is common as allocation of responsibility. As described in § 2.2, for CoE queries, the drug has already been taken and the outcome observed. In this setting we are interested in answering at the following question: given the fact that Ann actually took the drug and passed away, how likely she would not have died if she had not taken the drug?

In this dissertation, to answer the court's claim, we will use the *Probability of Causation* (PC) as given by Dawid in [16]. Given the triple $(X_A, Y_A(0), Y_A(1))$, we can define the Probability of Causation in Ann's case as:

Definition 5.0.1 (Probability of Causation)

$$PC_A = P_A(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1)$$

Where P_A denotes the probability distribution over Ann. Nowadays, several lawsuit are focused on allocation of responsibility. However, this preponderance of evidence is usually perceived as causality even without any kind of formal definition. If for example, the outcome was measured after being exposed to some treatment, in a court of law this is sometimes considered as causality. Furthermore,

in the previous section, we saw that such evidence cannot be assess using data on single individuals. For this purpose, it seems clear that we need a formal and mathematical definition such as Definition 5.0.1. In fact, PC_A is capable of answering this important causal question: Knowing that Ann did take the drug ($X_A = 1$) and the actual response was recovery ($Y_A = 1$), what is the probability that the potential response $Y_A(0)$, that would been observed had Ann not taken the drug, would have been different ($Y_A(0) = 0$)? How are we to understand this claim?

Let's suppose that a good experimental study tested the same drug taken by Ann. A possible example is reported in Table 5.1.

	Die	Live	Total
Exposed	30	70	100
Unexposed	12	88	100

Table 5.1: Deaths in individuals exposed and unexposed to the same drug taken by Ann

From Table 5.1 we can see that, in the experimental population, individuals exposed to the drug ($X \leftarrow 1$) were 18% more likely to death versus unexposed ($X \leftarrow 0$):

$$\mathrm{P}(Y = 1 \mid X \leftarrow 1) = 0.30 \tag{5.1}$$
$$\mathrm{P}(Y = 1 \mid X \leftarrow 0) = 0.12 \tag{5.2}$$

Can the court confirm that was Ann's taking the drug that caused her death? More important: is correct to use such experimental results, concerning a general population, to say something about a single individual? Basically this is the controversy discussed by Dawid *et al.* in [18] that is "when science is relied upon to answer factual disputes in litigation".

However, without any assumptions about the data generating process (monotonicity and/or exogeneity), we can not provide an exact estimate for PC_A but we can at least state useful information about its limitation. In fact, to estimate PC_A from the data, we need to assess the joint distribution of $(X_A, Y_A(0), Y_A(1))$. However, we can never observe both $Y(0)$ and $Y(1)$ for the same individual hence, we can never assess this dependence without making any further assumption. Given the important implications of the probability of causation in real life situations, it is clear that we have to focus on studying methods capable of producing more precise bounds.

As discussed in § 1.3, a more feasible target would be to measure the PC in

the whole population, instead of on a single individual. In this section we will discuss how combine population-level information with Ann's information to say something on the Probability of Causation for Ann. In particular we will show how these bounds can be improved or adapted if further information becomes available. In § 5.1 we will review the easier situation where we have information only on Ann's exposure and outcome. In § 5.2 we bound the probability of causation when we have additional information in the form of a pre-treatment covariate. Section § 5.3 illustrates the situation in which unobserved variable confounds the exposure-outcome relationship. Finally in § 5.4 and § 5.5 we will introduce a novel analysis to bounds the probability of causation in two different situations: a complete mediation mechanism and a partial mediation mechanism respectively.

5.1 Starting Point: Simple Analysis

In this section we discuss the simple situation in which we have information, as in Table 5.1, from a randomized experimental study. We need to assume that the fact of Ann's exposure, X_A, is independent of her potential responses $\boldsymbol{Y}_A$:

$$X_A \perp\!\!\!\perp \boldsymbol{Y}_A. \tag{5.3}$$

Property (5.3) parallels the "no-confounding" property $X_i \perp\!\!\!\perp \boldsymbol{Y}_i$ which holds for individuals i in the experimental study on account of randomization. We further suppose that Ann is exchangeable with the individuals in the experiment, *i.e.* she could be considered as a subject in the experimental population.

On account of (5.3) and exchangeability, the PC_A in Definition 2.2.1 reduces to $\mathrm{PC}_A = \mathrm{P}(Y(0) = 0 \mid Y(1) = 1)$, but we can not fully identify this from the data. In fact we can never observe the joint event $(Y(0) = 0; Y(1) = 1)$, since at least one of $Y(0)$ and $Y(1)$ must be counterfactual. In particular, we can never learn anything about the dependence between $Y(0)$ and $Y(1)$. However, even without making any assumptions about this dependence, we can derive the following inequalities as described by Dawid *et al.* (2015) in [20]:

$$1 - \frac{1}{\mathrm{RR}} \leq \mathrm{PC}_A \leq \frac{\mathrm{P}(Y = 0 \mid X \leftarrow 0)}{\mathrm{P}(Y = 1 \mid X \leftarrow 1)} \tag{5.4}$$

where

$$\mathrm{RR} = \frac{\mathrm{P}(Y = 1 \mid X \leftarrow 1)}{\mathrm{P}(Y = 1 \mid X \leftarrow 0)} \tag{5.5}$$

is the *experimental risk ratio* between exposed and unexposed. These bounds can be estimated from the experimental data using the population death rates computed in Equations (5.1) and (5.2).

In many cases of interest (such as Table 5.1), we have

$$\mathrm{P}(Y = 1 \mid X \leftarrow 0) < \mathrm{P}(Y = 1 \mid X \leftarrow 1) < \mathrm{P}(Y = 0 \mid X \leftarrow 0).$$

Then the lower bound in Equation (5.4) will be non-trivial, while the upper bound will exceed 1, and hence be vacuous.

We see from Equation (5.4) that, whenever RR > 2, the Probability of Causation PC_A will exceed 50%. In a civil court this is often taken as the criterion to assess legal responsibility "on the balance of probabilities" (although the converse is false: it would not be correct to infer $\mathrm{PC}_A < .5$ from the finding RR < 2). Since, in Table 5.1, the exposed are 2.5 times as likely to die as the unexposed (RR $= 30/12 = 2.5$), we have enough confidence to infer causality in Ann's case: we have $0.60 \leq \mathrm{PC}_A \leq 1$.

5.2 Additional Covariate Information

In this Section we show how we can refine the bounds of (5.4) if further information about a pre-treatment covariate S is available. For example, S might be a gene, possession of which enhances the dangerous effect of the exposure to the drug. We now take the assumptions of § 5.1 to hold after conditioning on S (indeed in cases where the original assumptions fail, it may well be possible to reinstate them by conditioning on a suitable covariate S). In particular, $X_A \perp\!\!\!\perp \boldsymbol{Y}_A \mid S_A$, and $X_i \perp\!\!\!\perp \boldsymbol{Y}_i \mid S_i$: adjusting for S is enough to control for confounding, both for Ann and in the study.

5.2.1 Fully observable

Consider first the situation where we can observe S both in the experimental data and in Ann. In this case, the PC_A should be replaced by the more specific definition

$$\mathrm{PC}_A = P_A(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1, S_A = s_A) \tag{5.6}$$

where s_A is Ann's value for S. We can apply the analysis of § 5.1, after conditioning on S, to obtain the estimable lower bound

$$1 - \frac{1}{\mathrm{RR}(s_A)} \leq \mathrm{PC}_A,$$

where

$$\mathrm{RR}(s) = \frac{\mathrm{P}(Y = 1 \mid X \leftarrow 1, S = s)}{\mathrm{P}(Y = 1 \mid X \leftarrow 0, S = s)}. \tag{5.7}$$

5.2.2 Observable in data only

Even when it is possible to observe S only in the population and not in Ann, we can sometimes refine the bounds in (5.4). Thus suppose S is binary, and from the data we infer the following probabilities (which in particular imply the same values as given in Table 5.1):

$$\begin{aligned} \mathrm{P}_A(S=1) &= 0.50 \\ \mathrm{P}_A(Y=1 \mid X \leftarrow 1, S=1) &= 0.60 \\ \mathrm{P}_A(Y=1 \mid X \leftarrow 0, S=1) &= 0 \qquad (5.8) \\ \mathrm{P}_A(Y=1 \mid X \leftarrow 1, S=0) &= 0 \qquad (5.9) \\ \mathrm{P}_A(Y=1 \mid X \leftarrow 0, S=0) &= 0.24. \end{aligned}$$

Since we know $X_A = 1$ and $Y_A = 1$, from Equation (5.9) we realize we can not have $S_A = 0$, so we must have $S_A = 1$. Then from Equation (5.8) we see that, when we set X to 0, we can not obtain $Y = 1$, so we must have $Y_A(0) = 0$. That is, in this special case we can infer causation in Ann's case—even though we have not directly observed her value for S.

More generally as described by Dawid (2011) in [16] we can refine the bounds in (5.4) as follows:

$$\frac{\Delta}{\mathrm{P}(Y=1 \mid X \leftarrow 1)} \leq \mathrm{PC} \leq 1 - \frac{\Gamma}{\mathrm{P}(Y=1 \mid X \leftarrow 1)} \qquad (5.10)$$

where

$$\Delta = \sum_s \mathrm{P}(S=s) \times \max\{0, \mathrm{P}(Y=1 \mid X \leftarrow 1, S=s) - \mathrm{P}(Y=1 \mid X \leftarrow 0, S=s)\}$$

and

$$\Gamma = \sum_s \mathrm{P}(S=s) \times \max\{0, \mathrm{P}(Y=1 \mid X \leftarrow 1, S=s) - \mathrm{P}(Y=0 \mid X \leftarrow 0, S=s)\}$$

These bounds are never wider than those obtained from (5.4), which ignores S.

5.3 Unobserved Confounding

So far we have assumed no confounding, $X \perp\!\!\!\perp \boldsymbol{Y}$ (perhaps conditionally on a suitable covariate S), both for Ann and for the study data. Now we drop this assumption for Ann. Then the experimental data can not be used, by themselves, to learn about $\mathrm{PC}_A = \mathrm{P}(Y_A(0) = 0 \mid X_A = 1, Y_A(1) = 1)$.

	Die	Live	Total
Exposed	18	82	100
Unexposed	24	76	100

Table 5.2: Observational data

We might however be able to gather additional *observational* data, where there was no possibility of experimental control over subjects' exposure, X, which might thus be related to unobserved personal aspects affecting the response Y. However—importantly—we now assume that the dependence between X and $\boldsymbol{Y}$ for subjects in the sampled population is just the same as it is for Ann. Let Q denote the joint observational distribution of (X, Y), which is estimable from such data. Tian and Pearl (2000) in [76] obtain the following bounds for PC_A, given both experimental and observational data:

$$\begin{aligned} &\max\left\{0, \frac{\mathrm{Q}(Y=1) - \mathrm{P}(Y=1 \mid X \leftarrow 0)}{\mathrm{Q}(X=1, Y=1)}\right\} \\ \leq\ &\mathrm{PC}_A \leq \min\left\{1, \frac{\mathrm{P}(Y=0 \mid X \leftarrow 0) - \mathrm{Q}(X=0, Y=0)}{\mathrm{Q}(X=1, Y=1)}\right\}. \end{aligned} \tag{5.11}$$

For example, suppose that, in addition to the data of Table 5.1, we have observational data as in Table 5.2.

Thus

$$\begin{aligned} \mathrm{Q}(Y=1) &= 0.21 \\ \mathrm{Q}(X=1, Y=1) &= 0.09 \\ \mathrm{Q}(X=0, Y=0) &= 0.38. \end{aligned}$$

Also, from Table 5.2 we have $\mathrm{P}(Y=1 \mid X \leftarrow 0) = 0.12$ (so $\mathrm{P}(Y=0 \mid X \leftarrow 0) = 1 - 0.12 = 0.88$). From Equation (5.11) we thus find $1 \leq \mathrm{PC}_A \leq 1$. We deduce that Ann would definitely have survived had she not taken the drug.

5.4 Complete Mediation

In this Section we present a novel analysis to bound the Probability of Causation for a case where a third variable, M, is involved in the causal pathway between the exposure X and the outcome Y [19, 41]. In particular, in this section, we will focus on the case of no direct effect, as intuitively described by Figure 5.1. Applications where this assumption might be plausible can be found in [38]. In this paper he presented various mechanisms of complete mediation such as: a tobacco prevention program reduces cigarette smoking by changing the social norms for tobacco use; exposure to negative life events affects blood pressure through the

mediation of cognitive attributions to stress. Another situation in which such an assumption might be plausible is in the treatment of ovarian cancer, Silber *et al.* (2007) in [70], where X represents management either by a medical oncologist or by a gynaecological oncologist, M is the intensity of chemotherapy prescribed, and Y is death within 5 years.

We shall be interested in the case that M is observed in the experimental data but is not observed for Ann, and see how this additional experimental evidence can be used to refine the bounds on PC_A.

Figure 5.1: Directed Acyclic Graph representing a mediator M, responding to exposure X and affecting response Y. There is no "direct effect", unmediated by M, of X on Y.

To formalize our assumption of "no direct effect", we introduce $M(x)$, the potential value of M for $X \leftarrow x$, and $Y^*(m)$, the potential value of Y for $M \leftarrow m$, where the irrelevance of the value x of X to Y^* encapsulates our assumption that X has no effect on Y over and above that transmitted through its influence on the mediator M. The potential value of Y for $X \leftarrow x$ (in cases where there is no intervention on M, which we here assume) is then $Y(x) := Y^*\{M(x)\}$.

In the sequel we restrict to the case that all variables are binary, and define $\boldsymbol{M} := (M(0), M(1))$, $\boldsymbol{Y}^* := (Y^*(0), Y^*(1))$, and $\boldsymbol{Y} := (Y(0), Y(1))$. In particular, we have observable variables $(X, M, Y) = (X, M(X), Y(X))$. We denote the bivariate distributions of the potential response pairs by

$$\begin{aligned} m_{ab} &:= \mathrm{P}(M(0) = a, M(1) = b) \\ y^*_{rs} &:= \mathrm{P}(Y^*(0) = r, Y^*(1) = s) \\ y_{rs} &:= \mathrm{P}(Y(0) = r, Y(1) = s). \end{aligned}$$

Then

$$\begin{aligned} m_{a+} &= \mathrm{P}(M = a \mid X \leftarrow 0) \\ m_{+b} &= \mathrm{P}(M = b \mid X \leftarrow 1) \\ y^*_{r+} &= \mathrm{P}(Y = r \mid M \leftarrow 0) \\ y^*_{+s} &= \mathrm{P}(Y = s \mid M \leftarrow 1) \\ y_{r+} &= \mathrm{P}(Y = r \mid X \leftarrow 0) \\ y_{+s} &= \mathrm{P}(Y = s \mid X \leftarrow 1), \end{aligned}$$

where m_{a+} denotes $\sum_{b=0}^{1} m_{ab}$, *etc.*

In addition to the assumptions of § 5.1 we further suppose that none of the causal mechanisms depicted in Figure 5.1 are confounded—expressed mathematically by assuming mutual independence between X, $\boldsymbol{M}$ and $\boldsymbol{Y}^*$ (both for experimental individuals, and for Ann). Then, m_{a+}, m_{+b}, y^*_{r+}, y^*_{+s}, y_{r+}, y_{+s} are all estimable from experimental data in which X is randomized, and M and Y are observed.

It is also easy to show the Markov property:

$$Y \perp\!\!\!\perp X \mid M. \tag{5.12}$$

This observable property can serve as a test of the validity of our conditions.

The assumed mutual independence implies

$$\begin{aligned} y_{rs} &= \mathrm{P}(Y^*(M(0)) = r, Y^*(M(1)) = s) \\ &= \sum_{a,b=0}^{1} \mathrm{P}(Y^*(a) = r, Y^*(b) = s)\mathrm{P}(M(0) = a, M(1) = b). \end{aligned}$$

This yields

$$\begin{aligned} y_{00} &= m_{00}y^*_{0+} + (m_{01} + m_{10})y^*_{00} + m_{11}y^*_{+0} \\ y_{01} &= m_{01}y^*_{01} + m_{10}y^*_{10} \\ y_{10} &= m_{01}y^*_{10} + m_{10}y^*_{01} \\ y_{11} &= m_{00}y^*_{1+} + (m_{01} + m_{10})y^*_{11} + m_{11}y^*_{+1}, \end{aligned} \tag{5.13}$$

and

$$y_{r+} = m_{0+}y^*_{r+} + m_{1+}y^*_{+r} \tag{5.14}$$

$$y_{+s} = m_{+0}y^*_{s+} + m_{+1}y^*_{+s}. \tag{5.15}$$

Suppose now that we observe $X_A = 1$ and $Y_A = 1$, but do not observe M_A. We have

$$\mathrm{PC}_A = \frac{y_{01}}{y_{+1}} = \frac{m_{01}y^*_{01} + m_{10}y^*_{10}}{y_{+1}}. \tag{5.16}$$

The denominator of (5.16) is $\mathrm{P}(Y = 1 \mid X \leftarrow 1)$, which is estimable from the data.

As for the numerator, this can be expressed as

$$2\mu\eta + A\mu + B\eta + AB = 2(\mu + B/2)(\eta + A/2) + AB/2 \tag{5.17}$$

with $\mu = m_{01}$, $\eta = y^*_{01}$, $A = y^*_{+0} - y^*_{0+}$, and $B = m_{+0} - m_{0+}$. Note that A and B are identified from the data, whereas for μ and η we can only obtain inequalities:

$$\begin{array}{rcccl} \max\{0,-B\} & \leq & \mu & \leq & \min\{m_{0+}, m_{+1}\} \\ \max\{0,-A\} & \leq & \eta & \leq & \min\{y^*_{0+}, y^*_{+1}\}, \end{array}$$

so that

$$\begin{array}{rcccl} |B/2| & \leq & \mu + B/2 & \leq & \min\{\frac{1}{2}(m_{0+} + m_{+0}), \frac{1}{2}(m_{1+} + m_{+1})\} \\ |A/2| & \leq & \eta + A/2 & \leq & \min\{\frac{1}{2}(y^*_{0+} + y^*_{+0}), \frac{1}{2}(y^*_{1+} + y^*_{+1})\}. \end{array} \tag{5.18}$$

The lower (respectively, upper) limit for (5.17) will be when $\mu + B/2$ and $\eta + A/2$ are both at their lower (respectively, upper) limits. In particular, the lower limit for (5.17) is $\max\{0, AB\}$. Using (5.14) and (5.15), we compute $AB = y_{+1} - y_{1+}$, which leads to the lower bound

$$\mathrm{PC}_A \geq 1 - \frac{\mathrm{P}(Y = 1 \mid X \leftarrow 0)}{\mathrm{P}(Y = 1 \mid X \leftarrow 1)} = 1 - \frac{1}{\mathrm{RR}},$$

exactly as for the case that M was not observed. Thus the possibility to observe a mediating variable in the experimental data has not improved our ability to lower bound PC_A.

We do however obtain an improved upper bound. Taking into account the various possible choices for the upper bounds in (5.18), the upper bound for the numerator of (5.16), in terms of experimentally estimable quantities, is given in Table 5.3.

	$m_{1+} + m_{+1} \geq 1$	$m_{1+} + m_{+1} < 1$
$y^*_{1+} + y^*_{+1} \geq 1$	$m_{0+}y^*_{0+} + m_{+0}y^*_{+0}$	$m_{1+}y^*_{+0} + m_{+1}y^*_{0+}$
$y^*_{1+} + y^*_{+1} < 1$	$m_{0+}y^*_{+1} + m_{+0}y^*_{1+}$	$m_{1+}y^*_{1+} + m_{+1}y^*_{+1}$

Table 5.3: Upper bound for the numerator of the PC_A in complete mediation

In § 5.6 will shown that this upper bound is never greater than that in (5.4), which ignores the mediator M, and is strictly smaller unless $y^*_{1+} + y^*_{+1} \geq 1$ and $m_{1+} + m_{+1} = 1$.

5.4.1 Identifiability under monotonicity

Without any assumption about the data generating process, in § 5.4 we calculated a lower and an upper bound for the probability of causation when a mediator completely justifies the causal relation between exposure and outcome. In this section we will show how PC_A can be identify under the monotonicity assumption.

Definition 5.4.1 (Monotonicity) *A variable Y is monotonic relative to X in a causal model, if and only if the bivariate distribution of $Y(x)$ is monotonic in x:*

$$P[Y(0) = 1, Y(1) = 0] = 0$$

For the particular case of no direct effect of X on Y, as described by the DAG in Figure 5.1, we obtained the following result:

Theorem 5.4.1 (Monotonicity in mediation) *If M is monotonic relative to X and Y is monotonic relative to M then Y is monotonic relative to X.*

Proof 5.4.1 *If M is monotonic relative to X then $m_{10} = 0$. If Y is monotonic relative to M then $y^*_{10} = 0$. Then, from (5.13), y_{10} will be zero.*

The reverse implication is not always true.
Then if M is monotonic relative to X and Y is monotonic relative to M the Probability of Causation will be

$$\mathrm{PC}_A = 1 - \frac{1}{\mathrm{RR}}.$$

Under the monotonicity assumption we obtained a result consistent with what Tian and Pearl (2000) found in [76]. This can perhaps be considered as a proof of the validity of this method.

5.4.2 Example

Let us suppose that Ann's children filled a criminal lawsuit against a pharmaceutical manufacturer claiming that Ann died after taking one of their drugs. On the other hand, the manufacturer claims that is a rare (binary) side effect of the drug that cause the death rather than the drug itself. Suppose we obtain the following values from the data:

$$\begin{aligned}
\mathrm{P}(M = 1 \mid X \leftarrow 1) &= 0.25 \\
\mathrm{P}(M = 1 \mid X \leftarrow 0) &= 0.025 \\
\mathrm{P}(Y = 1 \mid M \leftarrow 1) &= 0.9 \\
\mathrm{P}(Y = 1 \mid M \leftarrow 0) &= 0.1.
\end{aligned}$$

Again, these imply the values given in Table 5.1. According to the four possible combinations in Table 5.3, we find $0.60 \leq \mathrm{PC}_A \leq 0.76$; whereas without taking account of the mediator we would have no non-trivial upper bound. In § 5.6 we will compare these results with the simple analysis of X on Y and with the case of partial mediation.

5.5 Partial Mediation

The situation described in § 5.4 is unlikely to be true in real life situations. Directed Acyclic Graphs such Figure 3.1, that allow both direct and indirect effects, are more plausible and truthful. In addition, a complete mediation mechanism can be seen as a partial mediation mechanism when no direct effect is present. In this section we introduce new bounds for the probability of causation when a partial mediator is involved in the causal pathway [41]. In particular we will consider and compare two different settings: assuming usual exchangeability conditions and assuming new bivariate exchangeability conditions.

Sections § 5.2 and § 5.3 improve the bounds for PC_A if additional information about a pre-treatment covariate S is available. A conceivable proposal, to study PC_A in mediation analysis, would be to use (5.6) and (5.10) in every strata of the mediator instead of S. As we saw in § 1.3, where we defined mediation effects for EoC queries, conditioning on a mediator will produce a measure of the direct effect of the exposure on the outcome for a specific level of the mediator. We called this effect, controlled direct effect. Thus, using equations (5.6) and (5.10) to say something on PC_A when S is a mediator between X and Y will not produce a measure of the total effect of X on Y (as PC_A does) but a measure of the controlled direct effect between X and Y in levels of the mediator.

On the other hand, it would be interesting to study applications of Equations (5.6) and (5.10) when S is a mediator between X and Y. However, this new bounds cannot be compared with (5.4). This would be part of our future work about PC_A in mediation analysis.

Hereafter, in this section we will consider the following notation of counterfactual variables (consistent with the notation used in § 5.4)

1. $M(x)$ the potential value of M when X is set to x;

2. $Y^*(x, m)$ the potential value of Y when X is set to x and M is set to m;

3. $Y(x) = Y^*(x, M(x))$, the potential value of Y when X; is set to x and M arises naturally after setting X to x.

Let us suppose further the assumptions stated in § 5.1 that a good experimental study tested the same drug taken by Ann for which $X_i \perp\!\!\!\perp \boldsymbol{Y}_i$ and that Ann is exchangeable with the individuals in the experiment, *i.e.* she could be considered as a subject in the experimental population such that the PC_A will reduce to $\mathrm{PC}_A = \mathrm{P}(Y(0) = 0 \mid Y(1) = 1)$.

5.5.1 Disentangling the pathway for the PC

According to the mediation effects defined in the counterfactual framework in § 1.3, our first attempt was to define a "probability of direct causation" (PDC) and a "probability of indirect causation" (PIC) as

Definition 5.5.1 (Probabilities of direct causation)

$$PDC = P[Y^*(0, M(0)) = 0 | Y^*(1, M(0)) = 1]$$

and

Definition 5.5.2 (Probabilities of indirect causation)

$$PIC = P[Y^*(1, M(0)) = 0 | Y^*(1, M(1)) = 1].$$

Although Definition 5.5.1 and Definition 5.5.2 have useful applications, disentangling PC_A as a combination of PIC and PDC did not give any clear solution.

$$P[Y(0) = 0, Y(1) = 1] = P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 1] =$$

$$= P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0] + P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 1] =$$

$$= P[Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0] - P[Y^*(0, M(0)) = 1, Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0]$$

$$+ P[Y^*(0, M(0)) = 0, Y^*(1, M(0)) = 1] - P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 0, Y^*(1, M(0)) = 1] =$$

$$= P[Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0] + P[Y^*(0, M(0)) = 0, Y^*(1, M(0)) = 1] +$$

$$- \{P[Y^*(0, M(0)) = 1, Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0] + P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 0, Y^*(1, M(0)) = 1]\}$$

where we used the property
$P(A = a, B = b, C = c) = P(B = b, C = c) - P(A = 1 - a, B = b, C = c)$ for binary variable A,B and C taking values $\{0, 1\}$. For simplicity of notation we will refer to PC instead of PC_A.

We propose to disentangle the probability of causation as

$$\begin{aligned}
PC &= P[Y(0) = 0 | Y(1) = 1] = \frac{P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 1]}{P[Y^*(1, M(1)) = 1]} = \\
&= \frac{P[Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0]}{P[Y^*(1, M(1)) = 1]} + \frac{P[Y^*(0, M(0)) = 0, Y^*(1, M(0)) = 1]}{P[Y^*(1, M(1)) = 1]} \\
&\quad - \frac{1}{P[Y^*(1, M(1)) = 1]} \{P[Y^*(0, M(0)) = 1, Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0] \\
&\quad + P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 0, Y^*(1, M(0)) = 1]\} = \\
&= PIC + PDC \cdot \frac{P[Y^*(1, M(0)) = 1]}{P[Y^*(1, M(1)) = 1]} - \frac{1}{P[Y^*(1, M(1)) = 1]} \{P[Y^*(0, M(0)) = 1, Y^*(1, M(1)) = 1, Y^*(1, M(0)) = 0] + \\
&\quad + P[Y^*(0, M(0)) = 0, Y^*(1, M(1)) = 0, Y^*(1, M(0)) = 1]\}.
\end{aligned}$$

Unfortunately, this was not as clear as the case for EoC questions.

5.5.2 Linear programming

Our second attempt follows the reasoning of Tian and Pearl (2000) in [76]. Given the five variables X, $Y^*(0, M(0))$, $Y^*(0, M(1))$, $Y^*(1, M(0))$ and $Y^*(1, M(1))$ (with $i, j, k, r, s = \{0, 1\}$), we can specify 32 parameters, each one corresponding to one of the following joint probabilities

$$p_{ijkrs} = P(Y^*(1, M(1)) = i,\ Y^*(1, M(0)) = j,\ Y^*(0, M(1)) = k,\ Y^*(0, M(0)) = r,\ X = s)$$

which are constrained by the usual probability axioms

$$\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\sum_{r=0}^{1}\sum_{s=0}^{1} p_{ijkrs} = 1.$$

For example for $(i, j, k, r, s) = (1, 0, 1, 1, 1)$ and for the consistency condition on $Y^*(x, M(x))$ we have

$$\begin{aligned} p_{10111} &= \mathrm{P}(Y^*(1, M(1)) = 1,\ Y^*(1, M(0)) = 0,\ Y^*(0, M(1)) = 1,\ Y^*(0, M(0)) = 1,\ X = 1) = \\ &= \mathrm{P}(Y = 1,\ Y^*(1, M(0)) = 0,\ Y^*(0, M(1)) = 1,\ Y^*(0, M(0)) = 1,\ X = 1) \end{aligned}$$

Let's define further the jointly distribution of $Y(x, m)$ and the jointly distribution of $M(x)$, for $i, j, k, r = \{0, 1\}$

$$\begin{aligned} y_{ijkr} &= \mathrm{P}(Y^*(1, 1) = i,\ Y^*(1, 0) = j,\ Y^*(0, 1) = k,\ Y^*(0, 0) = r) \\ m_{ij} &= \mathrm{P}(M(1) = i,\ M(0) = j). \end{aligned}$$

If the consistency condition holds, it easy to prove the following rules

$$\begin{aligned} \mathrm{P}(Y = 1,\ X = 1) &= \mathrm{P}(Y(1) = 1,\ X = 1) = \sum_j\sum_k\sum_r p_{1jkr1} \\ \mathrm{P}(Y = 1,\ X = 0) &= \mathrm{P}(Y(0) = 1,\ X = 0) = \sum_i\sum_j\sum_k p_{ijk10} \\ \mathrm{P}(Y = 0,\ X = 1) &= \mathrm{P}(Y(1) = 0,\ X = 1) = \sum_j\sum_k\sum_r p_{0jkr1} \\ \mathrm{P}(Y = 0,\ X = 0) &= \mathrm{P}(Y(0) = 0,\ X = 0) = \sum_i\sum_j\sum_k p_{ijk00} \end{aligned}$$

and

$$\mathrm{P}(Y(1) = 1) = \mathrm{P}(Y^*(1, M(1)) = 1) = \sum_j\sum_k\sum_r\sum_s p_{1jkrs}$$

$$\mathrm{P}(Y(0)=1)=\mathrm{P}(Y^*(0,M(0))=1)=\sum_i\sum_j\sum_k\sum_s p_{ijk1s}$$
$$\mathrm{P}(Y^*(1,M(0))=1)=\sum_i\sum_k\sum_r\sum_s p_{i1krs}$$
$$\mathrm{P}(Y^*(0,M(1))=1)=\sum_i\sum_j\sum_r\sum_s p_{ij1rs}.$$

Suppose that we observed $X_A=1$ and $Y_A=1$ but not M_A. We can express the probability of causation for Ann's case as

$$\begin{aligned} PC_A &= \mathrm{P}(Y(0)=0|Y(1)=1,\,X=1)= \\ &= \frac{\mathrm{P}(Y(0)=0,\,Y(1)=1,\,X=1)}{\mathrm{P}(Y(1)=1,\,X=1)}=\frac{\sum_{j,k}p_{1jk01}}{\sum_{j,k,r}p_{1jkr1}}. \end{aligned} \tag{5.19}$$

The numerator above can be written as

$$\sum_{j,k}p_{1jk01}=\sum_{jkrs}p_{1jkrs}-\sum_{j,k}p_{1jk00}-\sum_{j,k}p_{1jk10}-\sum_{j,k}p_{1jk11}\le\sum_{jkrs}p_{1jkrs}=\mathrm{P}(Y^*(1,M(1))=1)$$

leading an upper bound for PC_A equals to

$$PC_A=\frac{\mathrm{P}(Y(0)=0,\,Y(1)=1,\,X=1)}{\mathrm{P}(Y(1)=1,\,X=1)}\le\frac{\mathrm{P}(Y^*(1,M(1))=1)}{\mathrm{P}(Y(1)=1,\,X=1)}. \tag{5.20}$$

Let us suppose that any of the $X-Y$, $X-M$ and $M-Y$ relation are confounded by C. Then from Equation (3.15), we can use the following decomposition

$$P[Y^*(x,M(\tilde{x}))=1]=\sum_m P(Y^*(x,m)=1|M(\tilde{x})=m)\,P(M(\tilde{x})=m)$$

that will produce an upper bound for PC_A in complete mediation analysis

$$PC_A\le\frac{\mathrm{P}(Y^*(1,M(1))=1)}{\mathrm{P}(Y(1)=1,\,X=1)}=\frac{\mathrm{P}(Y^*(1,1)=1)\,\mathrm{P}(M(1)=1)+\mathrm{P}(Y^*(1,0)=1)\,\mathrm{P}(M(1)=0)}{\mathrm{P}(Y(1)=1,\,X=1)}.$$

Under the assumption of no-confounding for every causal relation in the DAG in Figure 3.1, the above upper bound is estimable from nonexperimental data as

$$PC_A\le\frac{\mathrm{P}(Y=1|X=1,\,M=1)\,\mathrm{P}(M=1|X=1)+\mathrm{P}(Y=1|X=1,\,M=0)\,\mathrm{P}(M=0|X=1)}{\mathrm{P}(Y=1,\,X=1)}. \tag{5.21}$$

However, we do not obtain a different lower bound

$$\begin{aligned} PC_A &= \frac{\mathrm{P}(Y(0)=0,\, Y(1)=1)}{\mathrm{P}(Y=1|X=1)} = \frac{\sum p_{1jk0s}}{\mathrm{P}(Y=1|X=1)} = \\ &= \frac{\mathrm{P}(Y(1)=1) - \mathrm{P}(Y(0)=1) + \sum p_{0jk1s}}{\mathrm{P}(Y=1|X=1)} \geq \frac{\mathrm{P}(Y(1)=1) - \mathrm{P}(Y(0)=1)}{\mathrm{P}(Y=1|X=1)} \\ &= 1 - \frac{1}{RR} \end{aligned}$$

On the other hand the numerator in Equation (5.19) can be written also as

$$\begin{aligned} \sum_{j,k} p_{1jk01} &= \sum p_{ijk0s} - \sum p_{0jk00} - \sum p_{0jk01} - \sum p_{1jk00} \leq \sum p_{1jk0s} = \mathrm{P}(Y^*(0,M(0))=0) = \\ &= \mathrm{P}(Y^*(0,1)=0)\,\mathrm{P}(M(0)=1) + \mathrm{P}(Y^*(0,0)=0)\,\mathrm{P}(M(0)=0) = \\ &= \mathrm{P}(Y=0|X=0,M=1)\,\mathrm{P}(M=1|X=0) + \mathrm{P}(Y=0|X=0,M=0)\,\mathrm{P}(M=0|X=0). \end{aligned} \tag{5.22}$$

Considering both Equations (5.21) and (5.22), we obtained two possible upper bounds for PC_A not different from what Tian and Pearl (2000) found in [76]

$$PC_A \leq \frac{min\{\mathrm{P}(Y^*(0,M(0))=0), \mathrm{P}(Y^*(1,M(1))=1)\}}{\mathrm{P}(Y=1|X=1)}.$$

Thus, considering an additional variable in the problem but making no any special assumptions about the joint distribution, would not lead to more precise bounds.

5.5.3 Bound for PC in Mediation Analysis using Copulas

Let us consider the following decomposition of PC_A

$$\mathrm{PC}_A = \mathrm{P}(Y(0)=0 \mid X=1, Y(1)=1) = \frac{\mathrm{P}(Y(0)=0,\, Y(1)=1 \mid X=1)}{\mathrm{P}(Y(1)=1 \mid X=1)}. \tag{5.23}$$

Assuming only the consistency condition on $Y^*(x, M(x))$ we can obtain an interesting result, for the numerator above, that can be seen as a generalized mediation formula to bivariate distributions

Definition 5.5.3 (G-formula for bivariate distributions)

$$\begin{aligned} &P(Y(0)=0, Y(1)=1|X=1) = P(Y^*(0,M(0))=0, Y^*(1,M(1))=1|X=1) \\ &= \sum_{m_0=0}^{1}\sum_{m_1=0}^{1} P(Y^*(0,M(0))=0, Y^*(1,M(1))=1|M(0)=m_0, M(1)=m_1, X=1)P(M(0)=m_0, M(1)=m_1|X=1) \\ &= \sum_{m_0}\sum_{m_1} P(Y^*(0,m_0)=0, Y^*(1,m_1)=1|M(0)=m_0, M(1)=m_1, X=1)P(M(0)=m_0, M(1)=m_1|X=1). \end{aligned} \tag{5.24}$$

Equation (5.24) underlines a *bivariate dependence structure* between two hypothetical world: one where we interviewing setting each subject to be exposed and one where we interviewing setting each subject to be unexposed. Unfortunately, as we discussed in § 5, estimating these bivariate distributions from real data is not possible. When considering multivariate structures and dependencies between random variables, copula functions are among one the most exhaustive statistical tools [9, 44, 71, 83]. Given a set of random variables $X_1, \ldots, X_n$, the joint cumulative distribution function (CDF) $F_X(x_1, \ldots, x_n) = P[X_1 \leq x_1, \ldots, X_n \leq x_n]$ completely describe the dependencies among them. Given n CDF and given some information on the dependencies between them, we can deduce the multivariate CDF.

Let us consider again the random variables $X_1, \ldots, X_n$ with continuous marginal distributions $F_i(x) = \mathrm{P}(X_i \leq x)$. Applying the probability integral transformation to each component of the vector, we obtain

$$(U_1, \ldots, U_n) = (F_1(X_1), \ldots, F_n(X_n))$$

where each marginal has uniform distribution. The joint distribution of the n random variables is called *copula* of $X_1, \ldots, X_n$.

This can be written as

$$\begin{aligned} C(u_1, \ldots, u_n) &= C(F_1(x_1), \ldots, F_n(x_n)) = P[F_1(X_1) \leq F_1(x_1), \ldots, F_n(X_n) \leq F_n(x_n)] \\ &= F(x_1, \ldots, x_n) \end{aligned}$$

with $x_i = F_i^{-1}(u_i)$ for $i = 1, \ldots, n$.

In general we have the following definition and theorem

Definition 5.5.4 *A copula is the distribution function of a random variable in R^n with uniform-$(0, 1)$ marginals. Alternatively a copula is any function $C : [0, 1]^n \to [0, 1]$ with the following properties [73]:*

1. *$C(x_1, \ldots, x_n)$ is increasing in each component x_i;*
2. *$C(1, \ldots, 1, x_i, 1, \ldots, 1) = x_i$ for all $i \in \{1, \ldots, n\}$, $x_i \in [0, 1]$;*
3. *for all $(a_1, \ldots, a_n), (b_1, \ldots, b_n) \in [0, 1]^n$ with $a_i \leq b_i$ we have*

$$\sum_{i_1=1}^{2} \sum_{i_n=1}^{2} (-1)^{i_1 + \ldots + i_n} C(x_{1i_1}, \ldots, x_{ni_n}) \geq 0$$

where $x_{j1} = a_j$ and $x_{j2} = b_j$ for all $j \in \{1, \ldots, n\}$.

One of the most famous results, relative to the theory of copulas, is the following theorem due to Sklar (1959) in [71].

Theorem 5.5.1 (Sklar 1959) *Let $F(x_1, \ldots, x_n)$ be a joint cumulative distribution function with marginals $F_i(x_i)$. Then there exists a copula C such that, for all real values $(x_1, \ldots, x_n)$*

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)).$$

If the marginals $F_i(x_i)$ are all continuous, the copula is unique; otherwise it is uniquely determined on $range(F_1) \times \ldots \times range(F_n)$ which is the cartesian product of the ranges of the marginals CDF's. Conversely, if C is a copula and $F_i(x_i)$ are univariate CDF's then $F(x_1, \ldots, x_n)$ is a joint CDF with margins $F_i(x_i)$.

According to Theorem 5.5.1, once we know the marginal distributions $F_i(x_i)$ and the function C, we can completely reconstruct the joint distribution F.

A fundamental result, in copula's theory, is described by the following theorem which defines lower and upper bounds for copulas

Theorem 5.5.2 (Fréchet-Hoeffding bounds) *Let C be a multivariate copula and $X_1, \ldots, X_p$, p random variables. For every observed value $x_1, \ldots, x_p$*

$$\max\{\sum_{i=1}^{p} x_i + 1 - p, 0\} \leq C(x_1, \ldots, x_p) \leq \min\{x_1, \ldots, x_p\}$$

In terms of variables and CDF this theorem leads to the following bounds (see Avellana in [3])

$$\max\{\sum_{i=1}^{p} F(x_i) + 1 - p, 0\} \leq F(x_1, \ldots, x_p) \leq \min\{F(x_1), \ldots, F(x_p)\}. \tag{5.25}$$

For $p = 2$ the Inequality (5.25) leads directly to the bounds for PC_A in the simple analysis framework defined by the Inequality (5.4). For conditional probabilities, this theorem will be

Theorem 5.5.3 (Fréchet-Hoeffding for conditional probabilities) *If A, B and C are binary variable taking values $\{a, a'\}$, $\{b, b'\}$ and $\{c, c'\}$ then*

$$\max\{F(a|c) + F(b|c) - 1, 0\} \leq F(a, b|c) \leq \min\{F(a|c), F(b|c)\} \tag{5.26}$$

where $F(a|c) = F(A = a|C = c)$.

Proof 5.5.1 *For the upper bound*

$$F(a,b|c)+F(a,b'|c)=F(a|c) \;\Rightarrow\; F(a,b|c)=F(a|c)-F(a,b'|c) \;\Rightarrow\; F(a,b|c)\leq F(a|c)$$

given that $F(a,b'|c)\geq 0$.

For the lower bound

$$\begin{aligned}
&F(a,b|c)+F(a,b'|c)+F(a',b|c)+F(a',b'|c)=1\\
&\qquad\qquad \textit{then}\\
&F(a,b|c)=1-F(a,b'|c)-F(a',b|c)-F(a',b'|c)\\
&\qquad=1-F(b'|c)-F(a',b|c)\\
&\qquad=F(b|c)-F(a',b|c)+F(a',b'|c)-F(a',b'|c)\\
&\qquad=F(b|c)-F(a'|c)+F(a',b'|c)\\
&\qquad=F(b|c)+F(a|c)-1+F(a',b'|c).
\end{aligned}$$

In conclusion we will have $F(a,b|c)\geq F(b|c)+F(a|c)-1$ *given that* $F(a',b'|c)\geq 0$.

Here and after we will refer to the following assumptions (named $A\#$):

No-confounding Assumptions

1. $Y(x,m)\perp\!\!\!\perp M|X$ that is no $M-Y$ confounding **(A1)**;
2. $Y(x,m)\perp\!\!\!\perp X$ that is no $X-Y$ confounding **(A2)**;
3. $M(x)\perp\!\!\!\perp X$ that is no $X-M$ confounding **(A3)**.

We can directly apply Theorem 5.5.3 to equation (5.24)

$$\begin{aligned}
&A\leq \mathrm{P}(M(0)=m_0,M(1)=m_1|X=1)\leq B\\
&A=\max\{\mathrm{P}(M(0)=m_0|X=1)+\mathrm{P}(M(1)=m_1|X=1)-1,0\}\\
&B=\min\{\mathrm{P}(M(0)=m_0|X=1),\mathrm{P}(M(1)=m_1|X=1)\}.
\end{aligned}$$

Under the assumption **(A3)**, we were able to measure the above bounds from nonexperimental data

$$A=\max\{\mathrm{P}(M=m_0|X=0)+\mathrm{P}(M=m_1|X=1)-1,0\} \tag{5.27}$$
$$B=\min\{\mathrm{P}(M=m_0|X=0),\mathrm{P}(M=m_1|X=1)\}. \tag{5.28}$$

On the other hand, using Theorem 5.5.3, we obtained

$$C \leq \mathrm{P}(Y^*(0,m_0)=0,\ Y^*(1,m_1)=1,\ M(0)=m_0,\ M(1)=m_1|X=1) \leq D \tag{5.29}$$
$$C = \max\{\mathrm{P}(Y^*(0,m_0)=0,\ Y^*(1,m_1)=1|X=1) + \mathrm{P}(M(0)=m_0,\ M(1)=m_1|X=1) - 1, 0\}$$
$$D = \min\{\mathrm{P}(Y^*(0,m_0)=0,\ Y^*(1,m_1)=1|X=1),\ \mathrm{P}(M(0)=m_0,\ M(1)=m_1|X=1)\}.$$

Dividing the Inequality (5.29) by $\mathrm{P}(M(0),M(1)|X)$, we obtained bounds for the first term in Equation (5.24)

$$C' \leq \mathrm{P}(Y^*(0,m_0)=0, Y^*(1,m_1)=1|M(0)=m_0, M(1)=m_1, X=1) \leq D' \tag{5.30}$$
$$C' = \max\{\frac{\mathrm{P}(Y^*(0,m_0)=0, Y^*(1,m_1)=1|X=1)-1}{\mathrm{P}(M(0)=m_0, M(1)=m_1|X=1)} + 1, 0\}$$
$$D' = \min\{\frac{\mathrm{P}(Y^*(0,m_0)=0, Y^*(1,m_1)=1|X=1)}{\mathrm{P}(M(0)=m_0, M(1)=m_1|X=1)}, 1\}.$$

The inequality (5.30) can be further decomposed considering again Theorem 5.5.3 for

$$\begin{aligned} &E \leq \mathrm{P}(Y^*(0,m_0)=0, Y^*(1,m_1)=1|X=1) \leq F \\ &E = \max\{\mathrm{P}(Y^*(0,m_0)=0|X=1) + \mathrm{P}(Y^*(1,m_1)=1|X=1) - 1, 0\} \\ &F = \min\{\mathrm{P}(Y^*(0,m_0)=0|X=1), \mathrm{P}(Y^*(1,m_1)=1|X=1)\}. \end{aligned} \tag{5.31}$$

In conclusion, according to (5.31) for the numerator and (5.28) for the denominator, the term C' in (5.30) can be minimized as

$$C' \geq \max\{\frac{max\{\mathrm{P}(Y^*(0,m_0)=0|X=1) + \mathrm{P}(Y^*(1,m_1)=1|X=1) - 1, 0\} - 1}{min\{\mathrm{P}(M(0)=m_0|X=0), \mathrm{P}(M(1)=m_1|X=1)\}} + 1, 0\}. \tag{5.32}$$

In the same way, the term D' in the Inequality (5.30) can be maximized as

$$D' \leq \min\{\frac{min\{\mathrm{P}(Y^*(0,m_0)=0|X=1), \mathrm{P}(Y^*(1,m_1)=1|X=1)\}}{max\{\mathrm{P}(M(0)=m_0|X=1) + \mathrm{P}(M(1)=m_1|X=1) - 1, 0\}}, 1\}. \tag{5.33}$$

Given the assumptions **A1**, **A2** and **A1** and assuming consistency for every counterfactual variables considered, the inequalities 5.32 and 5.33 can be maximized and minimized by estimable quantities as

$$C' \geq \max\{\frac{max\{\mathrm{P}(Y=0|X=0,M=m_0)+\mathrm{P}(Y=1|X=1,M=m_1)-1,0\}-1}{min\{\mathrm{P}(M=m_0|X=0),\mathrm{P}(M=m_1|X=1)\}}+1,0\} \tag{5.34}$$

and

$$D' \leq \min\{\frac{min\{\mathrm{P}(Y=0|X=0,M=m_0),\mathrm{P}(Y=1|X=1,M=m_1)\}}{max\{\mathrm{P}(M=m_0|X=0)+\mathrm{P}(M=m_1|X=1)-1,0\}},1\}, \tag{5.35}$$

considering the following identifiability rules

$$\begin{aligned}\mathrm{P}(Y^*(x,m)=0|X=\tilde{x}) &= \mathrm{P}(Y^*(x,m)=0|X=x)\ \textbf{(A2)}\\ &= \mathrm{P}(Y^*(x,m)=0|X=x,M=m)\ \textbf{(A1)}\\ &= \mathrm{P}(Y=0|X=x,M=m)\ \textbf{Consistency of } Y^*(x,m)\end{aligned}$$

and

$$\begin{aligned}\mathrm{P}(M(x)=m_x|X=\tilde{x}) &= \mathrm{P}(M(x)=m_x|X=x)\ \textbf{(A3)}\\ &= \mathrm{P}(M=m_x|X=x)\ \textbf{Consistency of M(x)}.\end{aligned}$$

Thus, considering the estimable bounds (5.34) and (5.35) for the first term in Equation (5.24) and the bounds (5.27) and (5.28) for the second term, we get new estimable bounds for PC_A in partial mediation analysis.

Unfortunately, over 10000 simulations of the values included in the bounds (5.34), (5.35), (5.27) and (5.28), none of them lead to tighter bounds than the bounds in the inequality (5.4) for the simple case of X on Y. It seems that, bounding the probability of causation in mediation analysis, requires stronger conditions than the simply no-confounding assumptions and consistency to get better bounds for PC_A.

In the next sections we will test this hypothesis.

5.5.4 Bounds for PC_A assuming bivariate conditions

Let us consider the following bivariate assumptions for counterfactual outcome and mediator (named $B\#$):

Bivariate Assumptions

1. $(Y^*(0,m_0),Y^*(1,m_1)) \perp\!\!\!\perp (M(0),M(1))|X$ **(B1)**

2. $(Y^*(0,m_0), Y^*(1,m_1)) \perp\!\!\!\perp X$ **(B2)**

3. $(M(0), M(1)) \perp\!\!\!\perp X$ **(B3)**

Then we can decompose the numerator of PC_A

$$\mathrm{PC}_A = \mathrm{P}(Y(0)=0|Y(1)=1,X=1) = \frac{\mathrm{P}(Y(0)=0,Y(1)=1|X=1)}{\mathrm{P}(Y(1)=1|X=1)}$$

using the generalized mediation formula in Definition 5.5.3 and assuming conditions $B1$, $B2$ and $B3$ as

$$\begin{aligned}
&\mathrm{P}(Y(0)=0,Y(1)=1|X=1) = \mathrm{P}(Y^*(0,M(0))=0,Y^*(1,M(1))=1|X=1)\\
&= \sum_{m_0=0}^{1}\sum_{m_1=0}^{1} \mathrm{P}(Y^*(0,M(0))=0,Y^*(1,M(1))=1|M(0)=m_0,M(1)=m_1,X=1)\mathrm{P}(M(0)=m_0,M(1)=m_1|X=1)\\
&= \sum_{m_0}\sum_{m_1} \mathrm{P}(Y^*(0,m_0)=0,Y^*(1,m_1)=1|M(0)=m_0,M(1)=m_1,X=1)\mathrm{P}(M(0)=m_0,M(1)=m_1|X=1)\\
&= \sum_{m_0}\sum_{m_1} \mathrm{P}(Y^*(0,m_0)=0,Y^*(1,m_1)=1|X=1)\mathrm{P}(M(0)=m_0,M(1)=m_1) \ \textbf{B1 and B3}\\
&= \sum_{m_0}\sum_{m_1} \mathrm{P}(Y^*(0,m_0)=0,Y^*(1,m_1)=1)\mathrm{P}(M(0)=m_0,M(1)=m_1) \ \textbf{B2}
\end{aligned} \tag{5.36}$$

Upper Bound

Let us consider the Theorem 5.5.2 for $p=2$. The joint probability of two events is always smaller than the single probabilities, i.e. $\mathrm{P}(a,b) \leq min\{\mathrm{P}(a),\mathrm{P}(b)\}$, then

$$\mathrm{P}(Y(0)=0,Y(1)=1|X=1) \leq \sum_{m_0=0}^{1}\sum_{m_1=0}^{1} min\{\mathrm{P}(Y^*(0,m_0)=0),\mathrm{P}(Y^*(1,m_1)=1)\}\min\{\mathrm{P}(M(0)=m_0),\mathrm{P}(M(1)=m_1)\}. \tag{5.37}$$

This will lead to the following new upper bound for PC_A in partial mediation analysis

$$\begin{aligned}
&\mathrm{P}(Y(0)=0,\,Y(1)=1|X=1) \leq min\{\mathrm{P}(Y^*(0,0)=0),\mathrm{P}(Y^*(1,0)=1)\}\cdot min\{\mathrm{P}(M(0)=0),\mathrm{P}(M(1)=0)\}\\
&+ min\{\mathrm{P}(Y^*(0,0)=0),\mathrm{P}(Y^*(1,1)=1)\}\cdot min\{\mathrm{P}(M(0)=0),\mathrm{P}(M(1)=1)\}\\
&+ min\{\mathrm{P}(Y^*(0,1)=0),\mathrm{P}(Y^*(1,0)=1)\}\cdot min\{\mathrm{P}(M(0)=1),\mathrm{P}(M(1)=0)\}\\
&+ min\{\mathrm{P}(Y^*(0,1)=0),\mathrm{P}(Y^*(1,1)=1)\}\cdot min\{\mathrm{P}(M(0)=1),\mathrm{P}(M(1)=1)\}
\end{aligned} \tag{5.38}$$

Equation (5.38) underlines 64 different combinations of $\mathrm{P}(Y^*(x,m)=y)$ and $\mathrm{P}(M(x)=m)$. These can lead to 64 different upper bounds for PC_A in mediation analysis. The smaller probability, out of this 64 combinations, will be the designed value to bound from above the probability of causation for Ann's case.

Let us consider the bound for PC_A found in § 5.1 for the simple analysis of X on Y

$$PC = \mathrm{P}(Y(0) = 0 | Y(1) = 1,\ X = 1) \leq \frac{\min\{\mathrm{P}(Y(0) = 0), \mathrm{P}(Y(1) = 1)\}}{\mathrm{P}(Y(1) = 1)} \quad (5.39)$$

where

$$\mathrm{P}(Y(0) = 0) = \mathrm{P}(Y^*(0,0) = 0)\mathrm{P}(M(0) = 0) + \mathrm{P}(Y^*(0,1) = 0)\mathrm{P}(M(0) = 1) \quad (5.40)$$
$$\mathrm{P}(Y(1) = 1) = \mathrm{P}(Y^*(1,0) = 1)\mathrm{P}(M(1) = 0) + \mathrm{P}(Y^*(1,1) = 1)\mathrm{P}(M(1) = 1) \quad (5.41)$$

Comparing the above equations with (5.38), we can notice that they are a precise combination of (5.38). Then, bounding the probability of causation in the presence of a mediator considering only 5.40 and 5.41, will not describe every possible combination of the 64 probabilities. Thus bounding the probability of causation in the presence of a mediator using only the simple framework of X on Y might lead to wider bounds.

Identifiability Assumptions
In addition to the usual consistency assumptions on every counterfactual variable considered, to estimate the upper bound in (5.38) we need further assumptions on no-confounding between any variable in the mediation mechanism

1. $Y^*(x,m) \perp\!\!\!\perp M|X$ that is no $M - Y$ confounding **(A1)**;
2. $Y^*(x,m) \perp\!\!\!\perp X$ that is no $X - Y$ confounding **(A2)**;
3. $M(x) \perp\!\!\!\perp X$ that is no $X - M$ confounding **(A3)**;

such that

$$\begin{aligned}
\mathrm{P}(Y^*(x,m) = y) &= \mathrm{P}(Y^*(x,m) = y | X = x) \ \textbf{(A2)} \\
&= \mathrm{P}(Y^*(x,m) = y | X = x, M = m) \ \textbf{(A1)} \\
&= \mathrm{P}(Y = y | X = x, M = m) \ \textbf{Consistency on } Y^*(x,m)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{P}(M(x) = m) &= \mathrm{P}(M(x) = m | X = x) \ \textbf{(A3)} \\
&= \mathrm{P}(M = m | X = x). \ \textbf{Consistency on } M(x)
\end{aligned}$$

Lower Bound

According to Theorem 5.5.2 for $p = 2$, we can obtain a lower bound for PC_A in (5.36) as

$$\begin{aligned}
\mathrm{PC}_A &= \mathrm{P}(Y(0)=0|Y(1)=1,\ X=1) = \frac{\mathrm{P}(Y(0)=0,\ Y(1)=1|X=1)}{\mathrm{P}(Y(1)=1|X=1)} \\
&= \frac{\sum_{m_0}\sum_{m_1}\mathrm{P}(Y^*(0,m_0)=0,\ Y^*(1,m_1)=1)\mathrm{P}(M(0)=m_0,\ M(1)=m_1)}{\mathrm{P}(Y(1)=1|X=1)} \\
&\geq \frac{\sum_{m_0,m_1} max\{0,\mathrm{P}(Y^*(0,m_0)=0)+\mathrm{P}(Y^*(1,m_1)=1)-1\}\cdot max\{0,\mathrm{P}(M(0)=m_0)+\mathrm{P}(M(1)=m_1)-1\}}{\mathrm{P}(Y(1)=1|X=1)}.
\end{aligned}$$

If $max\{0,\mathrm{P}(Y^*(0,m_0)=0)+\mathrm{P}(Y^*(1,m_1)=1)-1\}$ is zero or $max\{0,\mathrm{P}(M(0)=m_0)+\mathrm{P}(M(1)=m_1)-1\}$ is zero, the lower bound will be zero. Otherwise we have

$$\begin{aligned}
\mathrm{PC}_A &\geq \frac{\sum_{m_0,m_1}[\mathrm{P}(Y^*(0,m_0)=0)+\mathrm{P}(Y^*(1,m_1)=1)-1]\cdot[\mathrm{P}(M(0)=m_0)+\mathrm{P}(M(1)=m_1)-1]}{\mathrm{P}(Y(1)=1|X=1)} \\
&= \frac{\sum_{m_0,m_1}\{\mathrm{P}(Y^*(1,m_1)=1)\mathrm{P}(M(1)=m_1)+[\mathrm{P}(Y^*(0,m_0)=0)-1]\mathrm{P}(M(0)=m_0)\}}{\mathrm{P}(Y(1)=1|X=1)}+ \\
&+ \frac{\sum_{m_0,m_1}\{[\mathrm{P}(Y^*(0,m_0)=0)-1][\mathrm{P}(M(1)=m_1)-1]+\mathrm{P}(Y^*(1,m_1)=1)[\mathrm{P}(M(0)=m_0)-1]\}}{\mathrm{P}(Y(1)=1|X=1)} \\
&= \frac{\sum_{m_0,m_1}\mathrm{P}(Y^*(1,m_1)=1)\mathrm{P}(M(1)=m_1)}{\mathrm{P}(Y(1)=1|X=1)} - \frac{\sum_{m_0,m_1}\mathrm{P}(Y^*(0,m_0)=1)\mathrm{P}(M(0)=m_0)}{\mathrm{P}(Y(1)=1|X=1)}+ \\
&- \frac{\sum_{m_0,m_1}\mathrm{P}(Y^*(0,m_0)=1)[\mathrm{P}(M(1)=m_1)-1]}{\mathrm{P}(Y(1)=1|X=1)} + \frac{\sum_{m_0,m_1}\mathrm{P}(Y^*(1,m_1)=1)[\mathrm{P}(M(0)=m_0)-1]}{\mathrm{P}(Y(1)=1|X=1)} \\
&= \frac{\sum_{m_1}\mathrm{P}(Y^*(1,m_1)=1)\mathrm{P}(M(1)=m_1)}{\mathrm{P}(Y(1)=1|X=1)} - \frac{\sum_{m_0}\mathrm{P}(Y^*(0,m_0)=1)\mathrm{P}(M(0)=m_0)}{\mathrm{P}(Y(1)=1|X=1)}+ \\
&- \frac{\sum_{m_0}\mathrm{P}(Y^*(0,m_0)=1)\sum_{m_1}[\mathrm{P}(M(1)=m_1)-1]}{\mathrm{P}(Y(1)=1|X=1)} + \frac{\sum_{m_1}\mathrm{P}(Y^*(1,m_1)=1)\sum_{m_0}[\mathrm{P}(M(0)=m_0)-1]}{\mathrm{P}(Y(1)=1|X=1)} \\
&= \frac{\sum_{m_1}\mathrm{P}(Y^*(1,m_1)=1)\mathrm{P}(M(1)=m_1)}{\mathrm{P}(Y(1)=1|X=1)} - \frac{\sum_{m_0}\mathrm{P}(Y^*(0,m_0)=1)\mathrm{P}(M(0)=m_0)}{\mathrm{P}(Y(1)=1|X=1)} \qquad (5.42)
\end{aligned}$$

where

$$\begin{aligned}
\frac{\sum_{m_0}\mathrm{P}(Y^*(0,m_0)=1)\sum_{m_1}[\mathrm{P}(M(1)=m_1)-1]}{\sum_{m_1}\mathrm{P}(Y^*(1,m_1)=1)\mathrm{P}(M(1)=m_1)} &= 0 \\
\frac{\sum_{m_1}\mathrm{P}(Y^*(1,m_1)=1)\sum_{m_0}[\mathrm{P}(M(0)=m_0)-1]}{\sum_{m_1}\mathrm{P}(Y^*(1,m_1)=1)\mathrm{P}(M(1)=m_1)} &= 0.
\end{aligned}$$

Identifiability Assumptions

Under the assumptions of consistency and no-confounding **A1**, **A2** and **A3** we can estimate the lower bound in (5.42) as

$$\mathrm{PC}_A \geq= \frac{\sum_{m_1}\mathrm{P}(Y^*(1,m_1)=1)\mathrm{P}(M(1)=m_1)}{\mathrm{P}(Y(1)=1|X=1)} - \frac{\sum_{m_0}\mathrm{P}(Y^*(0,m_0)=1)\mathrm{P}(M(0)=m_0)}{\mathrm{P}(Y(1)=1|X=1)}$$

$$
\begin{aligned}
&= 1 - \frac{\mathrm{P}(Y(0)=1)}{\mathrm{P}(Y(1)=1)} \\
&= 1 - \frac{\mathrm{P}(Y=1|X=0)}{\mathrm{P}(Y=1|X=1)} \quad \textbf{(A2)} \\
&= 1 - \frac{1}{\mathrm{RR}}. \qquad (5.43)
\end{aligned}
$$

Then, again, we do not obtain a different lower bound for PC_A even considering a partial mediation mechanism. This is perhaps consistent with what we found in § 5.4.

5.5.5 Bounds for PC_A assuming bivariate and univariate conditions

The conditions assumed in § 5.5.4 underline a dependence structure between the two hypothetical worlds arising after setting all patients to be exposed and all patients to be unexposed in the same time. For example, **(B3)** $(M(0), M(1)) \perp\!\!\!\perp X$ underlines a connection between what would have happened at the mediator if every subject would be exposed and what would have happened at the mediator if every subject would be unexposed relative to X. If we believe in the causal mechanism of X on M, the bivariate distribution $(M(0), M(1))$ is all is needed to infer the causation of X on M.
In fact, they imply the existence of a multivariate structure between counterfactuals that can be considered much more realistic than the simplistic assumption of independence.

Unfortunately, we can not evaluate any of this multivariate structure from the data. In this section we will weaken the bivariate assumptions **(B1)** to **(B3)** to obtain the same results found in § 5.5.4.

Let us consider the following assumptions regarding both univariate and bivariate potential distributions:

Bivariate and Univariate Assumptions

1. $Y^*(x, m) \perp\!\!\!\perp (M(0), M(1))|X$ **(C1)**;
2. $Y^*(x, m) \perp\!\!\!\perp X$ that is no $X - Y$ confounding **(A2)**;
3. $M(x) \perp\!\!\!\perp X$ that is no $X - M$ confounding **(A3)**.

Note that assumption **(C1)** $Y^*(x, m) \perp\!\!\!\perp (M(0), M(1))|X$ implies both $Y^*(x, m) \perp\!\!\!\perp M(0)|X$ and $Y^*(x, m) \perp\!\!\!\perp M(1)|X$, that is no $M - Y$ confounding. It involves a certain independence, given X, between the counterfactual values $Y^*(x, m)$,

arose setting exposure and mediator to a particular value, to both counterfactual outcomes $M(0)$ and $M(1)$. It can be seen as a generalization of the univariate hypothesis $Y^*(x,m) \perp\!\!\!\perp M(x)|X$.

Upper Bound

Using a combination of assumptions **(C1)**, **(A2)** and **(A3)** and of Theorem 5.5.3, we can obtain an upper bound for PC_A as

$$
\begin{aligned}
&\text{P}(Y(0)=0, Y(1)=1|X=1) = \text{P}(Y^*(0,M(0))=0,\ Y^*(1,M(1))=1|X=1)\\
&= \sum_{m_0=0}^{1}\sum_{m_1=0}^{1} \text{P}(Y^*(0,M(0))=0,\ Y^*(1,M(1))=1|M(0)=m_0,\ M(1)=m_1,\ X=1)\text{P}(M(0)=m_0,\ M(1)=m_1|X=1)\\
&= \sum_{m_0}\sum_{m_1} \text{P}(Y^*(0,m_0)=0,\ Y^*(1,m_1)=1|M(0)=m_0,\ M(1)=m_1,\ X=1)\text{P}(M(0)=m_0,\ M(1)=m_1|X=1)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textbf{Consistency on } Y^*(x,M(x))\\
&\leq \sum_{m_0}\sum_{m_1} min\{\text{P}(Y^*(0,m_0)=0|M(0)=m_0,\ M(1)=m_1,\ X=1), \text{P}(Y^*(1,m_1)=1|M(0)=m_0,\ M(1)=m_1,\ X=1)\}\\
&\cdot min\{\text{P}(M(0)=m_0|X=1), \text{P}(M(1)=m_1|X=1)\} \qquad \textbf{Theorem 5.5.3 on } Y^*(x,M(x)) \textbf{ and on } M(x)\\
&= \sum_{m_0}\sum_{m_1} min\{\text{P}(Y^*(0,m_0)=0|X=1), \text{P}(Y^*(1,m_1)=1|X=1)\} \cdot min\{\text{P}(M(0)=m_0|X=1), \text{P}(M(1)=m_1|X=1)\}.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textbf{(C1)} \quad (5.44)
\end{aligned}
$$

Assumptions **C1**, **A2** and **A3** will be enough to estimate the above quantities from the data, for example for $\text{P}(Y^*(0,m_0)=0|X=1)$ and $\text{P}(Y^*(1,m_1)=1|X=1)$ we have:

$$
\begin{aligned}
\text{P}(Y^*(x,m)=y|X=\tilde{x}) &= \text{P}(Y^*(x,m)=y|X=x)\ \textbf{(A2)}\\
&= \text{P}(Y^*(x,m)=y|X=x, M(x)=m)\ \textbf{(C1)}\\
&= \text{P}(Y=y|X=x, M=m)\ \textbf{Consistency of } Y^*(x,M(x))
\end{aligned}
$$

For $\text{P}(M(0)=m_0|X=1)$ and $\text{P}(M(1)=m_1|X=1)$:

$$
\begin{aligned}
\text{P}(M(x)=m|X=\tilde{x}) &= \text{P}(M(x)=m|X=x)\ \textbf{(A3)}\\
&= \text{P}(M=m|X=x)\ \textbf{Consistency of } M(x)
\end{aligned}
$$

The upper bound in (5.44), computed assuming only one bivariate condition, is the same as the one found in (5.38) assuming three different bivariate conditions

$$
\begin{aligned}
&\text{P}(Y(0)=0,\ Y(1)=1|X=1) \leq min\{\text{P}(Y^*(0,0)=0), \text{P}(Y^*(1,0)=1)\} \cdot min\{\text{P}(M(0)=0), \text{P}(M(1)=0)\} && (5.45)\\
&+ min\{\text{P}(Y^*(0,0)=0), \text{P}(Y^*(1,1)=1)\} \cdot min\{\text{P}(M(0)=0), \text{P}(M(1)=1)\} && (5.46)\\
&+ min\{\text{P}(Y^*(0,1)=0), \text{P}(Y^*(1,0)=1)\} \cdot min\{\text{P}(M(0)=1), \text{P}(M(1)=0)\} && (5.47)\\
&+ min\{\text{P}(Y^*(0,1)=0), \text{P}(Y^*(1,1)=1)\} \cdot min\{\text{P}(M(0)=1), \text{P}(M(1)=1)\} && (5.48)
\end{aligned}
$$

Lower Bound

For the lower bound we obtained

$$
\begin{aligned}
&\mathrm{P}(Y(0)=0, Y(1)=1|X=1) = \mathrm{P}(Y^*(0,M(0))=0,\ Y^*(1,M(1))=1|X=1)\\
&= \sum_{m_0=0}^{1}\sum_{m_1=0}^{1} \mathrm{P}(Y^*(0,M(0))=0,\ Y^*(1,M(1))=1|M(0)=m_0,\ M(1)=m_1,\ X=1)\mathrm{P}(M(0)=m_0,\ M(1)=m_1|X=1)\\
&= \sum_{m_0=0}^{1}\sum_{m_1=0}^{1} \mathrm{P}(Y^*(0,m_0)=0,\ Y^*(1,m_1)=1|M(0)=m_0,\ M(1)=m_1,\ X=1)\mathrm{P}(M(0)=m_0,\ M(1)=m_1|X=1)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textbf{Consistency on } Y^*(x,M(x))\\
&\geq max\{0, \mathrm{P}(Y^*(0,m_0)=0|M(0)=m_0,\ M(1)=m_1,\ X=1) + \mathrm{P}(Y^*(1,m_1)=1|M(0)=m_0,\ M(1)=m_1,\ X=1) - 1\}\\
&\qquad\qquad\qquad\qquad \cdot max\{0, \mathrm{P}(M(0)=m_0|X=1) + \mathrm{P}(M(1)=m_1|X=1) - 1\} \qquad \textbf{Theorem 5.5.3}\\
&= max\{0, \mathrm{P}(Y^*(0,m_0)=0|X=1) + \mathrm{P}(Y^*(1,m_1)=1|X=1) - 1\}\\
&\qquad\qquad\qquad\qquad \cdot max\{0, \mathrm{P}(M(0)=m_0|X=1) + \mathrm{P}(M(1)=m_1|X=1) - 1\} \qquad \textbf{C1} \qquad (5.49)
\end{aligned}
$$

Assumptions **C1**, **A2** and **A3** will be enough to estimate the above quantities from the data. The lower bound above, computed assuming only one bivariate condition, is the same as the one found in Equation (5.43) assuming three different bivariate conditions.

On assumption C1

Dawid (1979) in [14] and Constantinou and Dawid (2015) in [10] provide useful rules for conditional independence such as

$$X \perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp W|Y,Z \Leftrightarrow X \perp\!\!\!\perp (W,Y)|Z. \qquad (5.50)$$

In terms of potential mediator and outcome this will become

$$Y(x,m) \perp\!\!\!\perp M(1)|X \text{ and } Y(x,m) \perp\!\!\!\perp M(0)|X, M(1) \Leftrightarrow Y(x,m) \perp\!\!\!\perp (M(0), M(1))|X. \qquad (5.51)$$

That is

$$\textbf{B1} \text{ and } Y(x,m) \perp\!\!\!\perp M(0)|X, M(1) \Leftrightarrow C1. \qquad (5.52)$$

Condition $Y(x,m) \perp\!\!\!\perp M(0)|X, M(1)$, that we will call **D1**, is not new in the literature, is also assumed by Daniels *et al.* (2012) in [13]. However, this is not weaker than **D1**. It still assumes a dependence between cross world counterfactuals.

5.6 Comparisons

In this section we will compare the bounds found in the simple analysis framework § 5.1 with those obtained considering a complete mediation mechanism in § 5.4

and those obtained considering a partial mediation mechanism assuming both univariate and bivariate conditions as discussed in § 5.5.5. Here we will focus on comparing these bounds to obtain the best information from the data.

The numerator of the upper bound of PC_A in the simple analysis framework (5.4), which ignores the mediator, may be written as

$$\begin{aligned}&\min\{\mathrm{P}(Y^*(0,0)=0)\mathrm{P}(M(0)=0)+\mathrm{P}(Y^*(0,1)=0)\mathrm{P}(M(0)=1),\mathrm{P}(Y^*(1,0)=1)\mathrm{P}(M(1)=0)+\mathrm{P}(Y^*(1,1)=1)\mathrm{P}(M(1)=1)\}\\&=\min\{\alpha+\beta,\gamma+\delta\}.\end{aligned}\tag{5.53}$$

We can see that both (5.45) and (5.46) are smaller or at least equal than α, while both (5.47) and (5.48) are smaller or at least equal than β. So $2(\alpha+\beta)$, and similarly $2(\gamma+\delta)$, are bigger than the sum of (5.45), (5.46), (5.47) and (5.48). Thus the upper bound accounting for the mediator can not exceed twice that when ignoring it. However, as we will see in the next section subsection 5.6.1, it could be larger or smaller than that bound. On the other hand, we did not obtain a different lower bound.

Let us consider again the generalized g-formula

$$\begin{aligned}&\mathrm{P}(Y(0)=0,Y(1)=1|X=1)\\&=\sum_{m_0}\sum_{m_1}\mathrm{P}(Y^*(0,m_0)=0,Y^*(1,m_1)=1|M(0)=m_0,M(1)=m_1,X=1)\mathrm{P}(M(0)=m_0,M(1)=m_1|X=1).\end{aligned}$$

In the special case of complete mediation, the terms with $m_0 = m_1$, in the sum above, must be 0 leading to the following upper bound

$$\begin{aligned}&\mathrm{P}(Y(0)=0,Y(1)=1|X=1)\leq\\&+min\{\mathrm{P}(Y^*(0,0)=0),\mathrm{P}(Y^*(1,1)=1)\}\cdot min\{\mathrm{P}(M(0)=0),\mathrm{P}(M(1)=1)\}\\&+min\{\mathrm{P}(Y^*(0,1)=0),\mathrm{P}(Y^*(1,0)=1)\}\cdot min\{\mathrm{P}(M(0)=1),\mathrm{P}(M(1)=0)\}.\end{aligned}$$

Under the hypothesis of no direct effect, we can always equate $Y^*(x,m) = Y^*(m)$ for every x such that $Y^*(m) = Y^*(0,m) = Y^*(1,m)$. Then, assuming a complete mediation mechanism between X and Y, the numerator of PC_A can be written as

$$\begin{aligned}&\mathrm{P}(Y(0)=0,Y(1)=1|X=1)\leq\\&min\{\mathrm{P}(Y^*(0)=0),\mathrm{P}(Y^*(1)=1)\}\cdot min\{\mathrm{P}(M(0)=0),\mathrm{P}(M(1)=1)\}+\\&+min\{\mathrm{P}(Y^*(1)=0),\mathrm{P}(Y^*(0)=1)\}\cdot min\{\mathrm{P}(M(0)=1),\mathrm{P}(M(1)=0)\}\\&=min\{y^*_{0+},y^*_{+1}\}\cdot min\{m_{0+},m_{+1}\}+min\{y^*_{+0},y^*_{1+}\}\cdot min\{m_{1+},m_{+0}\}\end{aligned}$$

which collects all the four different possibilities in Table 5.3. Thus, starting from the bounds for PC_A in the more general case of partial mediation, assuming no

direct effect from X to Y, we obtained the bound for PC_A in the case of complete mediation. The bounds found in § 5.5.5 for partial mediation, are then consistent with the bounds found in § 5.4 for complete mediation.

We can even discuss how the upper bound for PC_A in a partial mediation mechanism can be considered as the sum of an indirect measure given by the sum of (5.46) and (5.47) and a direct measure given by the sum of (5.45) and (5.48). In fact, equations (5.46) and (5.47), are all that is needed to construct an upper bound for PC_A in complete mediation where no direct effect is present.

Moreover, equations (5.46) and (5.47) together, do not exceed $\alpha + \beta$. So assuming a complete mediation mechanism between X and Y, we never do worse than the simple analysis of X on Y.

5.6.1 Examples

In the case of partial mediation, taking account of information of the mediator can, but need not, yield a tighter upper bound. Thus suppose that a good experimental study tested the same drug taken by Ann, and produced the data reported in the tables below. In the first case, considering the mediator yields better bounds, but in the second we do better by ignoring it.

	Die	Live	Total
Exposed	69	31	100
Unexposed	24	76	100

Table 5.4: Deaths in individuals exposed and unexposed to the same drug taken by Ann

Introducing M, suppose we identify the following probabilities, consistent with the table above

$$\begin{aligned} &\mathrm{P}(Y^*(0,0)=0)=0.98 && \mathrm{P}(Y^*(0,1)=0)=0.165 \\ &\mathrm{P}(Y^*(1,0)=0)=0.315 && \mathrm{P}(Y^*(1,1)=0)=0.143 \\ &\mathrm{P}(M(0)=0)=0.73 && \mathrm{P}(M(1)=0)=0.981 \end{aligned}$$

We obtain: $0.65 \leq \mathrm{PC}_A \leq 0.81$ when accounting for the mediator, and $0.65 \leq \mathrm{PC}_A \leq 1$ when ignoring it.

On the other hand, let us consider an experimental study the data reported in the table below

Suppose we have the following probabilities, consistent with the table above

	Die	Live	Total
Exposed	78	22	100
Unexposed	68	32	100

Table 5.5: Deaths in individuals exposed and unexposed to the same drug taken by Ann

$$\begin{aligned} &\mathrm{P}(Y^*(0,0)=0)=0.98 && \mathrm{P}(Y^*(0,1)=0)=0.67 \\ &\mathrm{P}(Y^*(1,0)=0)=0.09 && \mathrm{P}(Y^*(1,1)=0)=0.27 \\ &\mathrm{P}(M(0)=0)=0.04 && \mathrm{P}(M(1)=0)=0.26 \end{aligned}$$

We obtain: $0.59 \leq \mathrm{PC}_A \leq 0.94$ when accounting for the mediator, but $0.59 \leq \mathrm{PC}_A \leq 0.87$ when ignoring it.

Chapter 6

Conclusions and further aims

In the EoC framework, Mediation analysis in the presence of unmeasured confounding is complex. We faced this problem in a real case when high parity is strongly associated with wheezing and likely to be mediated by birth weight. Several researchers designed methods capable of dealing with such problem. VanderWeele proposed a method not affected by collider bias that consists in conditioning on the estimated risk of being LBW instead of conditioning on M itself. In the dataset analyzed, in both situations of a rare and a regular outcome, this approach led to the same conclusions made on stratified measures. However, this method is highly affected by the poor predictive strength of the model designed to predict the mediator. In fact it is not simple to find, in a real case, such strong mediator's predictors able to avoid this issue. Furthermore, it precludes more real situations of an unmeasured intermediate confounding that might be responsible of the apparent paradox. Other problems may be caused may be the low prevalence of the mediator. On the other hand, mediation effects pointed out similar conclusions: the exposure appeared to act as a risk factor only in the normal birth weight intervention group, what is meant to be the least at risk, while they were not statistically associated in the low birth weight intervention group. Furthermore, in both cases, there were not evidence of an indirect effect of the exposure on the outcome trough the mediator. In the case of a rare outcome, there were not such evidence neither when accounting for the mediated interactive effect in the natural direct effect rather than in the indirect effect. In this situation we have to focus on sensitivity analysis. For the case of a rare outcomes, the sensitivity analysis we suggested, involved a graphical explanation of the hypothetical relational assumptions between U, M and Y that might mask the indirect effect. In the case of only linear relationship, we obtained a protective corrected indirect effects. However, by this method, we were not able to assess the magnitude of the unmeasured U affecting both mediator and outcome. We could at least state if it were protective or harmful. This can be done considering the second sensitivity analysis techniques. Setting the parameter γ, the hypothetical effect of U on Y, and the hypothetical prevalence of U in every strata of exposure, mediator and confounders, we were able to answer the

following research question: if we could control for U, what degree of confounding is capable of showing a $CDE(1)$ bigger (or at least equal) than $CDE(0)$? We saw that, only values of $\gamma \geq 4$ could reverse the relationship between $CDEs$, that is, only considerable unmeasured confounding U would reverse this relation. The corrected indirect effect associated to this sensitivity parameters, in line with the previous graphical result, should then be protective. In the case of a regular outcome, similar conclusions can be drawn from stratified OR, mediation effects and from conditioning on the risk of an intermediate. In particular we examined, via sensitivity analysis, two possible scenarios: un unmeasured variable U that affects the mediator-outcome relation and an additional unmeasured intermediate variable that is affected by the exposure. We saw that the differences between biased $CDEs$, in the case of one mediator, increased with the degree of confounding. Furthermore, when accounting for U, the corrected gap between $CDEs$ should show a completely different situation (only negative values) than the estimated results not accounting for U. However, the difference found in the NINFEA dataset can only be explained by a huge degree of confounding, that is with S bigger than 9 when each or both γ_u and β_u are bigger than 3. According to this result, for low prevalence of LBW, we can not longer conclude the absence of an indirect effect via birth weight in the case of one mediator. For the case of two mediators (one mediator of interest and one intermediate confounder): the sensitivity analysis shown similar conclusions for $CDEs$, except the results in 4.13e which are not consistent with the result found in the Ninfea dataset. On the other hand, 4.13f pointed out a biased indirect effect always bigger and closer to one than the indirect effect through LBW alone. In conclusion, to reverse the relationship between $CDEs$ and to produce the difference found in the NINFEA sample, the bias affecting this dataset should be considerable and perhaps unlikely. On the other hand, this severe degree of confounding is consistent with the simulations produced by Basso *et al.* in [6, 5]. The graphical sensitivity analysis that we proposed here, are a simple and useful tools capable of investigating different paradoxical scenarios. Further extensions of this work will include a statistical package for the sensitivity analysis proposed by VanderWeele and graphically illustrated by our figures.

From a CoE perspective, bounding the probability of causation in Mediation Analysis turned out to be challenging and interesting. The important implications of PC_A in real cases, encourage the researcher to focus on studying methods capable of producing more precise bounds. In this thesis we saw that, considering a complete mediation mechanism, we never do worse than the simple analysis of X on Y. However, when a partial mediation mechanism disentangle the relation between exposure and mediator, usual assumptions of no unmeasured confounding were not enough to obtain smaller bounds. Considering an additional bivariate assumption between potential variables, we obtained new bounds for PC_A in the case of partial mediation. In particular, we saw that bounding the probability of causation not accounting for the mediator (as in the simple framework of X

on Y) will not be exhaustive of all possible scenarios and hence, might lead to wider bounds. This bivariate hypothesis is not new in the literature and might be connected to similar assumptions. In this particular case it requires independence between some potential variables.

This work has several possible extensions. It could be very interesting to apply these theoretical formulas to real data combining also information on both covariates and mediators. Another development could be to include Ann's mediator value in the formulation of PC_A and define bounds for the "controlled probability of causation". Another extension that seems to be promising is to implement different copula functions in order to obtain an exact estimate for the probability of causation which must lie in the bounds found in this thesis.

Bibliography

[1] Italian web-based birth cohort ninfea. `https://www.progettoninfea.it/`. Accessed: 2016-01-04.

[2] C. V. Ananth and T. J. VanderWeele. Placental abruption and perinatal mortality with preterm delivery as a mediator: disentangling direct and indirect effects. *American journal of epidemiology*, page kwr045, 2011.

[3] C. M. C. Avellana. The importance of being the upper bound in the bivariate family. *SORT: statistics and operations research transactions*, 30(1):55–84, 2006.

[4] 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, 1986.

[5] O. Basso and A. J. Wilcox. Intersecting birth weight-specific mortality curves: solving the riddle. *American journal of epidemiology*, 169(7):787–797, 2009.

[6] O. Basso, A. J. Wilcox, and C. R. Weinberg. Birth weight and mortality: causality or confounding? *American Journal of Epidemiology*, 164(4):303–311, 2006.

[7] K. A. Bollen. *Structural equations with latent variables*. John Wiley & Sons, 2014.

[8] A.-M. Brooks, R. S. Byrd, M. Weitzman, P. Auinger, and J. T. McBride. Impact of low birth weight on early childhood asthma in the united states. *Archives of pediatrics & adolescent medicine*, 155(3):401–406, 2001.

[9] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula methods in finance*. John Wiley & Sons, 2004.

[10] P. Constantinou and A. P. Dawid. Extended conditional independence and applications in causal inference. *arXiv preprint arXiv:1512.00245*, 2015.

[11] R. M. Daniel, B. L. De Stavola, S. Cousens, and S. Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

[12] R. M. Daniel, B. L. De Stavola, and S. N. Cousens. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata Journal*, 11(4):479, 2011.

[13] M. J. Daniels, J. A. Roy, C. Kim, J. W. Hogan, and M. G. Perri. Bayesian inference for the causal effect of mediation. *Biometrics*, 68(4):1028–1036, 2012.

[14] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

[15] A. P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000.

[16] A. P. Dawid. The role of scientific and statistical evidence in assessing causality. *Perspectives on Causation*, pages 133–147, 2011.

[17] A. P. Dawid. Statistical causality from a decision-theoretic perspective. *arXiv preprint arXiv:1405.2292*, 2014.

[18] A. P. Dawid, D. L. Faigman, and S. E. Fienberg. Fitting science into legal contexts assessing effects of causes or causes of effects? *Sociological Methods & Research*, page 0049124113515188, 2013.

[19] A. P. Dawid, R. Murtas, and M. Musio. Bounding the probability of causation in mediation analysis. *Selected papers of the 47th Scientific meeting of the Italian Statistical Society, Springer Book*, Editors: Tonio Di Battista, Elas Moreno, Walter Racugno. In press.

[20] A. P. Dawid, M. Musio, and F. S. E. From statistical evidence to evidence of causality. *Bayesian Analysis*, 2015, Online first.

[21] B. L. De Stavola, R. M. Daniel, G. B. Ploubidis, and N. Micali. Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. *American journal of epidemiology*, 181(1):64–80, 2015.

[22] V. Didelez, A. P. Dawid, and S. Geneletti. Direct and indirect effects of sequential treatments. *arXiv preprint arXiv:1206.6840*, 2012.

[23] S. Geneletti. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):199–215, 2007.

[24] A. F. Hayes and K. J. Preacher. Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behavioral Research*, 45(4):627–660, 2010.

[25] M. Heaman, D. Kingston, B. Chalmers, R. Sauve, L. Lee, and D. Young. Risk factors for preterm birth and small-for-gestational-age births among canadian women. *Paediatric and perinatal epidemiology*, 27(1):54–61, 2013.

[26] M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health*, 60(7):578–586, 2006.

[27] M. A. Hernan and J. M. Robins. *Causal inference*. CRC, 2010.

[28] S. Hernández-Díaz, E. F. Schisterman, and M. A. Hernán. The birth weight paradox uncovered? *American journal of epidemiology*, 164(11):1115–1120, 2006.

[29] G. Hesslow. Two notes on the probabilistic approach to causality. *Philosophy of science*, pages 290–292, 1976.

[30] P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[31] 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.

[32] B. Hu, J. Shao, and M. Palta. Pseudo-r^ 2 in logistic regression model. *Statistica Sinica*, 16(3):847, 2006.

[33] C. M. Judd and D. A. Kenny. Process analysis estimating mediation in treatment evaluations. *Evaluation review*, 5(5):602–619, 1981.

[34] P. S. Laplace. Mémoire sur la probabilité des causes par les évènemens. *Mémories de Mathématique et de Physique, presentés à l'Académie Royale des Sciences*, pages 621–656, 1774.

[35] P. S. Laplace. Memoir on the probability of the causes of events. *Statistical Science*, pages 364–378, 1986.

[36] D. Lewis. *Counterfactuals*. Harvard University Press, 1973.

[37] D. P. MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2008.

[38] D. P. MacKinnon and A. J. Fairchild. Current directions in mediation analysis. *Current Directions in Psychological Science*, 18(1):16–20, 2009.

[39] D. P. MacKinnon, G. Warsi, and J. H. Dwyer. A simulation study of mediated effect measures. *Multivariate behavioral research*, 30(1):41–62, 1995.

[40] T. F. Mebrahtu, R. G. Feltbower, and R. C. Parslow. Effects of birth weight and growth on childhood wheezing disorders: findings from the born in bradford cohort. *BMJ open*, 5(11):e009553, 2015.

[41] R. Murtas, A. P. Dawid, and M. Musio. Probability of causation: bounds and identification for partial and complete mediation analysis. *Manuscript*, 2016.

[42] R. Murtas, B. L. De Stavola, D. Zugna, and L. Richiardi. Birth order and wheezing: an example of sensitivity analysis for the mediating effect of birth weight. *Manuscript submitted for publication*, 2016.

[43] B. Muthén. Applications of causally defined direct and indirect effects in mediation analysis using sem in mplus. *Manuscript submitted for publication*, 2011.

[44] R. B. Nelsen. *An introduction to copulas*, volume 139. Springer Science & Business Media, 2013.

[45] W. H. Oddy, J. K. Peat, and N. H. de Klerk. Maternal asthma, infant feeding, and the risk of asthma in childhood. *Journal of allergy and clinical immunology*, 110(1):65–67, 2002.

[46] N. Pearce, R. Beasley, C. Burgess, and J. Crane. *Asthma epidemiology: principles and methods*. Oxford University Press New York, 1998.

[47] J. Pearl. *Bayesian networks: A model of self-activated memory for evidential reasoning*. University of California (Los Angeles). Computer Science Department, 1985.

[48] J. Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121(1-2):93–149, 1999.

[49] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[50] J. Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001.

[51] J. Pearl. *Causality*. Cambridge university press, 2009.

[52] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[53] J. Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[54] M. L. Petersen, S. E. Sinisi, and M. J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.

[55] C. Pizzi, B. L. De Stavola, F. Merletti, R. Bellocco, I. dos Santos Silva, N. Pearce, and L. Richiardi. Sample selection and validity of exposure–disease association estimates in cohort studies. *Journal of epidemiology and community health*, 65(5):407–411, 2011.

[56] C. Pizzi, B. L. De Stavola, N. Pearce, F. Lazzarato, P. Ghiotti, F. Merletti, and L. Richiardi. Selection bias and patterns of confounding in cohort studies: the case of the ninfea web-based birth cohort. *Journal of epidemiology and community health*, 66(11):976–981, 2012.

[57] T. S. Richardson and J. M. Robins. Single world intervention graphs: a primer. In *Second UAI Workshop on Causal Structure Learning, Bellevue, Washington*, 2013.

[58] L. Richiardi, I. Baussano, L. Vizzini, J. Douwes, N. Pearce, and F. Merletti. Feasibility of recruiting a birth cohort through the internet: the experience of the ninfea cohort. *European journal of epidemiology*, 22(12):831–837, 2007.

[59] L. Richiardi, I. Baussano, L. Vizzini, J. Douwes, N. Pearce, and F. Merletti. Feasibility of recruiting a birth cohort through the internet: the experience of the ninfea cohort. *European journal of epidemiology*, 22(12):831–837, 2007.

[60] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

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

[62] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[63] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

[64] J. Roy, J. W. Hogan, and B. H. Marcus. Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*, 9(2):277–289, 2008.

[65] D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[66] D. Seidman, A. Laor, R. Gale, D. Stevenson, and Y. Danon. Is low birth weight a risk factor for asthma during adolescence? *Archives of disease in childhood*, 66(5):584–587, 1991.

[67] R. Shaw, K. Woodman, J. Crane, C. Moyes, J. Kennedy, and N. Pearce. Risk factors for asthma symptoms in kawerau children. *The New Zealand medical journal*, 107(987):387–391, 1994.

[68] I. Shpitser. *Graph-Based Criteria of Identifiability of Causal Questions*. Wiley Online Library, 2012.

[69] I. Shpitser and J. Pearl. Complete identification methods for the causal hierarchy. *The Journal of Machine Learning Research*, 9:1941–1979, 2008.

[70] J. H. Silber, P. R. Rosenbaum, D. Polsky, R. N. Ross, O. Even-Shoshan, J. S. Schwartz, K. A. Armstrong, and T. C. Randall. Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? *Journal of Clinical Oncology*, 25(10):1169–1175, 2007.

[71] A. Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

[72] J. Splawa-Neyman, D. Dabrowska, T. Speed, et al. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

[73] S. E. Starkstein and M. Merello. *Psychiatric and cognitive disorders in Parkinson's disease*. Cambridge University Press, 2002.

[74] M. Szumilas. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3):227, 2010.

[75] E. J. Tchetgen Tchetgen and T. J. VanderWeele. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *BMJ open*, 25(2):282–91, March 2014.

[76] J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

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

[78] T. J. VanderWeele. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.)*, 21(4):540, 2010.

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

[80] T. J. VanderWeele and Y. Chiba. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, biostatistics, and public health*, 11(2), 2014.

[81] T. J. VanderWeele, S. L. Mumford, and E. F. Schisterman. Conditioning on intermediates in perinatal epidemiology. *Epidemiology (Cambridge, Mass.)*, 23(1):1, 2012.

[82] S. Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

[83] A. Zilio. Il metodo delle copule in finanza: inferenza, calcolo del var ed implementazione in r. 2005.

Appendix A

Software development

Here we report a selection of methods presented in § 4.2 and implemented in Stata 12.

```
*********************************************************************
*****CHAPTER:4, SECTION: 2, SUBSECTION: 3. RARE OUTCOME RESULTS******
*********************************************************************

use dataset.dta,clear

**************************Stratified OR*******************************
gen xm=x*m
logit y i.x i.xm i.m anno_child mat_age gest_age smoking mat_asthma
lincom _b[1.x]+_b[1.xm]
di exp(r(estimate))

******************Partitioning the causal effects********************
set seed 12345
capture program drop direct_effect_adjusted
program direct_effect_adjusted
qui logit  m x mat_age gest_age smoking
scalar beta0=_b[_cons]
scalar beta_x=_b[x]
scalar beta_1a=_b[mat_age]
scalar beta_2b=_b[gest_age]
scalar beta_2c=_b[smoking]

qui logit y x m xm anno_child gest_age smoking mat_asthma
scalar theta0=_b[_cons]
scalar theta_x=_b[x]
scalar theta_m=_b[m]
```

```
scalar theta_xm=_b[xm]
scalar theta_2b=_b[gest_age]
scalar theta_2c=_b[smoking]
scalar theta_3=_b[anno_child]
scalar theta_4=_b[mat_asthma]

* mean children: born in 2009 with 39 weeks of gestation form a
* non-smoking mother of 33 years corresponding in the model to
* anno_child==0 (because already centered at 2009)
* mat_age==0 (because already centered at 33)
* gest_age==2 (because centered at 37)
* and smoking=0

*Mediation effects for a mean individual
scalar cde0=exp(theta_x)
scalar cde1=exp(theta_x+theta_xm)
scalar pnde=exp(theta_x)*((1+exp(theta_m+theta_xm+beta_0 ///
    +beta_2b*2))/(1+exp(theta_m+beta_0+beta_2b*2)))
scalar tnie=((1+exp(beta_0+beta_2b*2))*(1+exp(theta_m+theta_xm ///
    +beta_0+beta_x+beta_2b*2)))/((1+exp(beta_0+beta_x+beta_2b*2))*(1 ///
    +exp(theta_m+theta_xm+beta_0+beta_2b*2)))
scalar tnde=(exp(theta_x)*(1+exp(theta_m+theta_xm+beta_0+beta_x///
    +beta_2b*2)))/(1+exp(theta_m+beta_0+beta_x+beta_2b*2))
scalar pnie=((1+exp(beta_0+beta_2b*2))*(1+exp(theta_m+beta_0+beta_x ///
    +beta_2b*2)))/((1+exp(beta_0+beta_x+beta_2b*2))*(1 ///
    +exp(theta_m+beta_0+beta_2b*2)))
scalar tce=pnde*tnie
scalar K=tnie/pnie
scalar list
end

bootstrap direct_effect_adjusted cde0_c=cde0 cde1_c=cde1 pnde_c=pnde ///
    tnie_c=tnie tnde_c=tnde pnie_c=pnie tce_c=tce1 ///
    K_c=K, reps(10000) nodots

***************VanderWeeles approaches to the paradox***************
use dataset_rare.dta,clear
logit m gest_age mat_age smoking dets

predict risk, pr
centile risk,centile(95.0) ///.1370409
gen h=0
replace h=1 if risk>=.1370409
```

```
replace h=. if risk==.

gen xh=x*h
logit y x h xh anno_child gest_age smoking
lincom _b[x]+_b[xh]
di exp(r(estimate))

***************Sensitivity Analysis: Rare Outcome*******************
*simulations

clear
set obs 10000

gen pi_00=runiform()
gen pi_10=runiform()
gen pi_11=runiform()
gen pi_01=runiform()

*the prevalence of U among unexposed was set to be bigger than the
*prevalence of U among exposed
gen rule_0=(pi_10<pi_00)
gen rule_1=(pi_11<pi_01)

foreach gamma of numlist 2 3 4 5 7{

gen bias_cde_0=((1+(‘gamma’-1)*pi_10)/(1+(‘gamma’-1)*pi_00))
gen bias_cde_1=((1+(‘gamma’-1)*pi_11)/(1+(‘gamma’-1)*pi_01))
gen cde_unb_0=3.32/bias_cde_0
gen cde_unb_1=1.02/bias_cde_1
gen diff=cde_unb_0-cde_unb_1

twoway (scatter pi_10 pi_00 if diff<=0 & rule_0==1 & rule_1==1) ///
    (function x, range(pi_00) n(2)), xtitle(pi_00) ///
    ytitle(pi_10) title(g=‘gamma’)

twoway (scatter pi_11 pi_01 if diff<=0 & rule_0==1 & rule_1==1)///
    (function x, range(pi_01) n(2)), xtitle(pi_01) ytitle(pi_11) ///
    title(g=‘gamma’)
drop bias_* cde_* diff
}

*plotted graph (4.7 and 4.8)
```

```
gen bias_cde_0=((1+(4-1)*pi_10)/(1+(4-1)*pi_00))
gen bias_cde_1=((1+(4-1)*pi_11)/(1+(4-1)*pi_01))

gen cde_unb_0=3.32/bias_cde_0
gen cde_unb_1=1.02/bias_cde_1
gen diff=cde_unb_0-cde_unb_1

keep if  rule_0==1 &  rule_1==1 & diff<=0
*only 4 combinations selected

gen n=_n

twoway (scatter pi_10 pi_00 if n==1, msymbol(square)) ///
    (scatter pi_10 pi_00 if n==2, msymbol(triangle)) ///
    (scatter pi_10 pi_00 if n==3, msymbol(diamond)) ///
    (scatter pi_10 pi_00 if n==4, msymbol(plus)) ///
    (function x, range(pi_00) n(2)), xtitle(pi_00) ///
    ytitle(pi_10) title(g=4) name(graph_4_1, replace)

twoway (scatter pi_11 pi_01 if n==1, msymbol(square)) ///
    (scatter pi_11 pi_01 if n==2, msymbol(triangle)) ///
    (scatter pi_11 pi_01 if n==3, msymbol(diamond)) ///
    (scatter pi_11 pi_01 if n==4, msymbol(plus)) ///
    (function x, range(pi_00) n(2)), xtitle(pi_01)///
    ytitle(pi_11) title(g=4) name(graph_4_2, replace)

*bias affecting PNDE and TNIE
use "dataset_rare.dta", clear
logit m x age gest mcp_msmoke
gen p_m1=exp(_b[_cons]+_b[x]*0+_b[gest]*2)/(1+exp(_b[_cons]+_b[x]*0+_b[gest]*2))
sum p_m1 //0.05
gen p_m0=1/(1+exp(_b[_cons]+_b[x]*0+_b[gest]*2))
sum p_m0 //0.95
gen num=exp(_b[_cons]+_b[x]*0+_b[gest]*2)
sum num //0.05

di ((1+3*0.38)+(1+3*0.05)*0.05)/((1+3*0.4)+(1+3*0.93)*0.05) //0.92
di 3.11/0.91 //3.42
di 1.01*0.91 //0.92
```

Here we report a selection of methods presented in § 4.3 and implemented in Stata 12.

```
**********************************************************************
```

```
****CHAPTER:4, SECTION: 3, SUBSECTION: 2. REGULAR OUTCOME RESULTS****
**********************************************************************

use dataset_regular.dta,clear

******************Partitioning the causal effects********************

*gformula version 1.13

gformula y x m xm anno_child mat_age smoking gest_age, mediation ///
    outcome(y) exposure(x) mediator(m) base_confs(anno_c age ///
    mcp_msmoke gest) com(y:logit, m:logit) eq(y: x m xm anno_child ///
    smoking gest_age, m: x mat_age smoking gest_age) derived(xm)///
    derrules(xm:x*m) control(m:1) minsim obe moreMC samples(1000) ///
    simulations(10000) seed(79) logOR

gformula y x m xm anno_child mat_age smoking gest_age, mediation ///
    outcome(y) exposure(x) mediator(m) base_confs(anno_c age ///
    mcp_msmoke gest) com(y:logit, m:logit) eq(y: x m xm anno_child ///
    mat_age smoking gest_age, m: x mat_age smoking gest_age) derived(xm)///
    derrules(xm:x*m) control(m:0) minsim obe moreMC samples(1000) ///
    simulations(10000) seed(79) logOR

**********************************************************************
****CHAPTER:4, SECTION: 3, SUBSECTION: 3. REG. OUT. Sensitivity******
**********************************************************************

******************g-formula by hand, one mediator********************

clear
set seed 100

program drop simulations
program define simulations, rclass
version 12.0
clear
syntax [, s(integer 10) obs(integer 10000) p_x(real 0.25) ///
    gamma0(real -2.9) gammax(real -0.44) gammau(real 0) ///
    p_u(real 0) beta0(real -1.85) betax(real 2.16) betam(real 1.09)///
    betaxm(real 1) betau(real 1)]

set obs `obs'
qui gen id=_n
```

```
qui gen u=runiform()<‘p_u’
qui gen x=runiform()<‘p_x’
qui gen p_m=(exp(‘gamma0’+‘gammax’*x+‘gammau’*u)/(1+exp(‘gamma0’ ///
    +‘gammax’*x+‘gammau’*u)))

qui gen m=runiform()<p_m

qui gen xm=x*m
qui gen p_y=(exp(‘beta0’+‘betax’*x+‘betam’*m+‘betaxm’*xm+‘betau’*u)/(1 ///
    +exp(‘beta0’+‘betax’*x+‘betam’*m+‘betaxm’*xm+‘betau’*u)))
qui gen y=runiform()<p_y

return scalar prev_m=‘prev_m’
return scalar prev_y=‘prev_y’

preserve

expand ‘s’

sort id

qui by id:gen original=_n==1

*GFORMULA BY HAND with U: unbiased results
qui logit m x u if original==1
qui gen M_0=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[u]*u)))) //M(0)
qui gen M_1=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u)))) //M(1)

qui logit y x m xm u if original==1
*P(Y(0,0)=1)
qui gen p_Y_00=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*0+_b[xm]*0*0+_b[u]*u))))
*P(Y(0,1)=1)
qui gen p_Y_01=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*1+_b[xm]*0*1+_b[u]*u))))
*P(Y(1,0)=1)
qui gen p_Y_10=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*0+_b[xm]*1*0+_b[u]*u))))
*P(Y(1,1)=1)
qui gen p_Y_11=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*1+_b[xm]*1*1+_b[u]*u))))

qui egen m_p_Y_00=mean(p_Y_00) //E(Y(0,0))
qui egen m_p_Y_01=mean(p_Y_01) //E(Y(0,1))
qui egen m_p_Y_10=mean(p_Y_10) //E(Y(1,0))
qui egen m_p_Y_11=mean(p_Y_11) //E(Y(1,1))
```

```
*P(Y(0,M(0))=1)
qui gen p_Y_0_M_0=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*M_0 ///
    +_b[xm]*0*M_0+_b[u]*u))))
*P(Y(1,M(0))=1)
qui gen p_Y_1_M_0=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*M_0 ///
    +_b[xm]*1*M_0+_b[u]*u))))
*P(Y(1,M(1))=1)
qui gen p_Y_1_M_1=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*M_1 ///
    +_b[xm]*1*M_1+_b[u]*u))))

qui egen m_p_Y_0_M_0=mean(p_Y_0_M_0) //E(Y(0,M(0)))
qui egen m_p_Y_1_M_0=mean(p_Y_1_M_0) //E(Y(1,M(0)))
qui egen m_p_Y_1_M_1=mean(p_Y_1_M_1) //E(Y(1,M(1)))

*log(TCE)=log(E(Y(1,M(1)))/1-E(Y(1,M(1))))-log(E(Y(0,M(0)))/1-E(Y(0,M(0))))
local lTCE_unb=log(m_p_Y_1_M_1/(1-m_p_Y_1_M_1)) ///
    -log(m_p_Y_0_M_0/(1-m_p_Y_0_M_0))
*log(NDE)=log(E(Y(1,M(0)))/1-E(Y(1,M(0))))-log(E(Y(0,M(0)))/1-E(Y(0,M(0))))
local lNDE_unb=log(m_p_Y_1_M_0/(1-m_p_Y_1_M_0)) ///
    -log(m_p_Y_0_M_0/(1-m_p_Y_0_M_0))
*log(NIE)=log(TCE)-log(NDE)
local lNIE_unb=`lTCE_unb'-`lNDE_unb'


*log(CDE(0))=log(E(Y(1,0))/1-E(Y(1,0)))-log(E(Y(0,0))/1-E(Y(0,0)))
local lCDE_0_unb=log(m_p_Y_10/(1-m_p_Y_10))-log(m_p_Y_00/(1-m_p_Y_00))
*log(CDE(1))
local lCDE_1_unb=log(m_p_Y_11/(1-m_p_Y_11))-log(m_p_Y_01/(1-m_p_Y_01))

drop p_Y* m_p_Y* M_*

return scalar lcde_0_unb=`lCDE_0_unb'
return scalar lcde_1_unb=`lCDE_1_unb'
return scalar lnde_unb=`lNDE_unb'
return scalar ltce_unb=`lTCE_unb'
return scalar lnie_unb=`lNIE_unb'

*GFORMULA BY HAND without U: biased results
qui logit m x if original==1
qui gen M_0=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*0)))) //M(0)
qui gen M_1=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*1)))) //M(1)
```

```
qui logit y x m xm if original==1
*P(Y(0,0)=1)
qui gen p_Y_00=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*0+_b[xm]*0*0))))
*P(Y(0,1)=1)
qui gen p_Y_01=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*1+_b[xm]*0*1))))
*P(Y(1,0)=1)
qui gen p_Y_10=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*0+_b[xm]*1*0))))
*P(Y(1,1)=1)
qui gen p_Y_11=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*1+_b[xm]*1*1))))

qui egen m_p_Y_00=mean(p_Y_00) //E(Y(0,0))
qui egen m_p_Y_01=mean(p_Y_01) //E(Y(0,1))
qui egen m_p_Y_10=mean(p_Y_10) //E(Y(1,0))
qui egen m_p_Y_11=mean(p_Y_11) //E(Y(1,1))

*P(Y(0,M(0))=1)
qui gen p_Y_0_M_0=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*M_0+_b[xm]*0*M_0))))
*P(Y(1,M(0))=1)
qui gen p_Y_1_M_0=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*M_0+_b[xm]*1*M_0))))
*P(Y(1,M(1))=1)
qui gen p_Y_1_M_1=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*M_1+_b[xm]*1*M_1))))

qui egen m_p_Y_0_M_0=mean(p_Y_0_M_0) //E(Y(0,M(0)))
qui egen m_p_Y_1_M_0=mean(p_Y_1_M_0) //E(Y(1,M(0)))
qui egen m_p_Y_1_M_1=mean(p_Y_1_M_1) //E(Y(1,M(1)))

*log(TCE)=log(E(Y(1,M(1)))/1-E(Y(1,M(1))))-log(E(Y(0,M(0)))/1-E(Y(0,M(0))))
local lTCE_b=log(m_p_Y_1_M_1/(1-m_p_Y_1_M_1))-log(m_p_Y_0_M_0/(1-m_p_Y_0_M_0))
*log(NDE)=log(E(Y(1,M(0)))/1-E(Y(1,M(0))))-log(E(Y(0,M(0)))/1-E(Y(0,M(0))))
local lNDE_b=log(m_p_Y_1_M_0/(1-m_p_Y_1_M_0))-log(m_p_Y_0_M_0/(1-m_p_Y_0_M_0))
*log(NIE)=log(TCE)-log(NDE)
local lNIE_b=‘lTCE_b’-‘lNDE_b’


*log(CDE(0))=log(E(Y(1,0))/1-E(Y(1,0)))-log(E(Y(0,0))/1-E(Y(0,0)))
local lCDE_0_b=log(m_p_Y_10/(1-m_p_Y_10))-log(m_p_Y_00/(1-m_p_Y_00))
*log(CDE(1))
local lCDE_1_b=log(m_p_Y_11/(1-m_p_Y_11))-log(m_p_Y_01/(1-m_p_Y_01))

drop p_Y* m_p_Y* M_*

return scalar lcde_0_b=‘lCDE_0_b’
```

```
return scalar lcde_1_b=‘lCDE_1_b’
return scalar lnde_b=‘lNDE_b’
return scalar ltce_b=‘lTCE_b’
return scalar lnie_b=‘lNIE_b’
end

local i=0
local drop i
foreach bu of numlist -0.7 1 2 3{
foreach gu of numlist -0.7 1 2 3{
foreach bm of numlist -0.7 1{
foreach g0 of numlist -2 -1.4 0.4 2.2{
local i=‘i’+1
local bu10=‘bu’*10
local gu10=‘gu’*10
local bm10=‘bm’*10
local g010=‘g0’*10
di
di "_____________________________________________________"
di "betam is ‘bm’; betau is ‘bu’; gammau is ‘gu’; gamma0 is ‘g0’"
di "you are at the interaction ‘i’ out of 128 "
di
simulate prev_m=r(prev_m) prev_y=r(prev_y) ///
    lcde_0_unb=r(lcde_0_unb) lcde_0_b=r(lcde_0_b) ///
    lcde_1_unb=r(lcde_1_unb) lcde_1_b=r(lcde_1_b) ///
lnde_unb=r(lnde_unb) lnde_b=r(lnde_b) lnie_unb=r(lnie_unb) ///
    lnie_b=r(lnie_b) ltce_unb=r(ltce_unb) ltce_b=r(ltce_b), ///
reps(1000) saving(path\results_‘bu10’_‘gu10’_‘bm10’_‘g010’_‘i’,replace): ///
    simulations, obs(10000) s(10) p_x(0.25) p_u(0.5) gamma0(‘g0’) ///
    gammax(-.91) gammau(‘gu’) beta0(-1.4) betax(0.8) betam(‘bm’) ///
    betaxm(0) betau(‘bu’)
}
}
}
}

**********g-formula by hand, two mediators**********

set seed 100
program drop simulations2
program define simulations2, rclass
version 12.0
clear
```

```
syntax [, s(integer 10) obs(integer 10000) p_x(real 0.25) ///
    alpha0(real 0.4) alphax(real 0) gamma0(real -2.9) ///
    gammax(real -0.44) gammau(real 0) gammaux(real 0) ///
    beta0(real -1.85) betax(real 2.16) betam(real 1.09) ///
    betau(real 1) betaxm(real 1) betaux(real 0) ///
    betaum(real 0)]

set obs `obs'
qui gen id=_n

qui gen x=runiform()<`p_x'
qui gen p_u=(exp(`alpha0'+`alphax'*x)/(1+exp(`alpha0'+`alphax'*x)))
qui gen u=runiform()<p_u
qui gen ux=u*x
qui gen p_m=(exp(`gamma0'+`gammax'*x+`gammau'*u+`gammaux'*ux)/(1 ///
    +exp(`gamma0'+`gammax'*x+`gammau'*u+`gammaux'*ux)))
qui gen m=runiform()<p_m
return scalar prev_m=`prev_m'

qui gen xm=x*m
qui gen um=u*m
qui gen ux2=u*x
qui gen p_y=(exp(`beta0'+`betax'*x+`betam'*m+`betau'*u+`betaxm'*xm ///
    +`betaux'*ux2+`betaum'*um)/(1+exp(`beta0'+`betax'*x+`betam'*m ///
    +`betau'*u+`betaxm'*xm+`betaux'*ux2+`betaum'*um)))
qui gen y=runiform()<p_y
return scalar prev_y=`prev_y'

preserve
expand `s'
sort id
qui by id:gen original=_n==1

*GFORMULA BY HAND with U: unbiased results

qui logit u x if original==1
qui gen u_0=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*0)))) //U(0)
qui gen u_1=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*1)))) //U(1)

qui logit m x u ux if original==1
*M(X=0,U(X=0))
qui gen p_m_00=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[u]*u_0+_b[ux]*0*u_0))))
*M(X=1,U(X=1))
```

```
qui gen p_m_11=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_1+_b[ux]*1*u_1))))
*M(X=0,U(X=1))
qui gen p_m_01=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[u]*u_1+_b[ux]*0*u_1))))
*M(X=1,U(X=0))
qui gen p_m_10=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_0+_b[ux]*1*u_0))))

qui egen m_p_m_00=mean(p_m_00)
qui egen m_p_m_11=mean(p_m_11)
qui egen m_p_m_01=mean(p_m_01)
qui egen m_p_m_10=mean(p_m_10)

qui logit y x u m xm ux um if original==1
*P(Y(1,U(1),M(1,U(0)) )=1)
qui gen p_Y_1110=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_1+_b[m]*m_p_m_10 ///
    +_b[xm]*1*m_p_m_10+_b[ux]*1*u_1+_b[um]*u_1*m_p_m_10))))
*P(Y(1,U(1),M(0,U(0)) )=1)
qui gen p_Y_1100=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_1+_b[m]*m_p_m_00 ///
    +_b[xm]*1*m_p_m_00+_b[ux]*1*u_1+_b[um]*u_1*m_p_m_00))))
*P(Y(1,U(1),M(1,U(1)) )=1)
qui gen p_Y_1111=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_1+_b[m]*m_p_m_11 ///
    +_b[xm]*1*m_p_m_11+_b[ux]*1*u_1+_b[um]*u_1*m_p_m_11))))
*P(Y(1,U(1),M(1,U(1)) )=1)
qui gen p_Y_1000=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_0+_b[m]*m_p_m_00 ///
    +_b[xm]*1*m_p_m_00+_b[ux]*1*u_0+_b[um]*u_0*m_p_m_00))))
*P(Y(1,U(1),M(1,U(1)) )=1)
qui gen p_Y_0000=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[u]*u_0+_b[m]*m_p_m_00 ///
+_b[xm]*0*m_p_m_00+_b[ux]*0*u_0+_b[um]*u_0*m_p_m_00))))

qui egen m_p_Y_1110=mean(p_Y_1110) //E[Y(1,U(1),M(1,U(0)) )]
qui egen m_p_Y_1100=mean(p_Y_1100) //E[Y(1,U(1),M(0,U(0)) )]
qui egen m_p_Y_1111=mean(p_Y_1111) //E[Y(1,U(1),M(1,U(1)) )]
qui egen m_p_Y_1000=mean(p_Y_1000) //E[Y(1,U(0),M(0,U(0)) )]
qui egen m_p_Y_0000=mean(p_Y_0000) //E[Y(0,U(0),M(0,U(0)) )]

*P(Y(1,U(1),0)=1)
qui gen p_c_Y_110=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_1+_b[m]*0 ///
    +_b[xm]*1*0+_b[ux]*1*u_1+_b[um]*u_1*0))))
*P(Y(0,U(0),0)=1)
qui gen p_c_Y_000=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[u]*u_0+_b[m]*0 ///
    +_b[xm]*0*0+_b[ux]*0*u_0+_b[um]*u_0*0))))
*P(Y(1,U(1),1)=1)
qui gen p_c_Y_111=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[u]*u_1+_b[m]*1 ///
    +_b[xm]*1*1+_b[ux]*1*u_1+_b[um]*u_1*1))))
```

```
*P(Y(0,U(0),1)=1)
qui gen p_c_Y_001=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[u]*u_0+_b[m]*1 ///
    +_b[xm]*0*1+_b[ux]*0*u_0+_b[um]*u_0*1))))

qui egen m_p_c_Y_110=mean(p_c_Y_110) //E[Y(1,U(1),0)]
qui egen m_p_c_Y_000=mean(p_c_Y_000) //E[Y(0,U(0),0)]
qui egen m_p_c_Y_111=mean(p_c_Y_111) //E[Y(1,U(1),1)]
qui egen m_p_c_Y_001=mean(p_c_Y_001) //E[Y(0,U(0),1)]

local lNIE_unb_M_110=log(m_p_Y_1110/(1-m_p_Y_1110)) ///
    -log(m_p_Y_1100/(1-m_p_Y_1100))
local lNIE_unb_U_100=log(m_p_Y_1100/(1-m_p_Y_1100)) ///
    -log(m_p_Y_1000/(1-m_p_Y_1000))
local lNIE_unb_UM_111=log(m_p_Y_1111/(1-m_p_Y_1111)) ///
    -log(m_p_Y_1110/(1-m_p_Y_1110))
local lNDE_unb_000=log(m_p_Y_1000/(1-m_p_Y_1000)) ///
    -log(m_p_Y_0000/(1-m_p_Y_0000))
local lTCE_unb=‘lNIE_unb_M_110’+‘lNIE_unb_U_100’ ///
    +‘lNIE_unb_UM_111’+‘lNDE_unb_000’

local lCDE_M_0_unb=log(m_p_c_Y_110/(1-m_p_c_Y_110)) ///
    -log(m_p_c_Y_000/(1-m_p_c_Y_000))
local lCDE_M_1_unb=log(m_p_c_Y_111/(1-m_p_c_Y_111)) ///
    -log(m_p_c_Y_001/(1-m_p_c_Y_001))

drop u_* m_p_m_* p_m_* m_p_Y_*
drop p_Y_* m_p_c_Y_* p_c_Y_*

return scalar lCDE_M_0_unb=‘lCDE_M_0_unb’
return scalar lCDE_M_1_unb=‘lCDE_M_1_unb’
return scalar lNIE_unb_M_110=‘lNIE_unb_M_110’
return scalar lNIE_unb_U_100=‘lNIE_unb_U_100’
return scalar lNIE_unb_UM_111=‘lNIE_unb_UM_111’
return scalar lNDE_unb_000=‘lNDE_unb_000’
return scalar lTCE_unb=‘lTCE_unb’

*GFORMULA BY HAND without U: biased results
qui logit m x if original==1
qui gen M_0=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*0)))) //M(0)
qui gen M_1=runiform()<(1/(1+exp(-(_b[_cons]+_b[x]*1)))) //M(1)

qui logit y x m xm if original==1
*P(Y(0,0)=1)
```

```
qui gen p_Y_00=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*0+_b[xm]*0*0))))
*P(Y(0,1)=1)
qui gen p_Y_01=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*1+_b[xm]*0*1))))
*P(Y(1,0)=1)
qui gen p_Y_10=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*0+_b[xm]*1*0))))
*P(Y(1,1)=1)
qui gen p_Y_11=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*1+_b[xm]*1*1))))

qui egen m_p_Y_00=mean(p_Y_00) //E(Y(0,0))
qui egen m_p_Y_01=mean(p_Y_01) //E(Y(0,1))
qui egen m_p_Y_10=mean(p_Y_10) //E(Y(1,0))
qui egen m_p_Y_11=mean(p_Y_11) //E(Y(1,1))

*P(Y(0,M(0))=1)
qui gen p_Y_0_M_0=(1/(1+exp(-(_b[_cons]+_b[x]*0+_b[m]*M_0+_b[xm]*0*M_0))))
*P(Y(1,M(0))=1)
qui gen p_Y_1_M_0=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*M_0+_b[xm]*1*M_0))))
*P(Y(1,M(1))=1)
qui gen p_Y_1_M_1=(1/(1+exp(-(_b[_cons]+_b[x]*1+_b[m]*M_1+_b[xm]*1*M_1))))

qui egen m_p_Y_0_M_0=mean(p_Y_0_M_0) //E(Y(0,M(0)))
qui egen m_p_Y_1_M_0=mean(p_Y_1_M_0) //E(Y(1,M(0)))
qui egen m_p_Y_1_M_1=mean(p_Y_1_M_1) //E(Y(1,M(1)))

*log(TCE)=log(E(Y(1,M(1)))/1-E(Y(1,M(1))))-log(E(Y(0,M(0)))/1-E(Y(0,M(0))))
local lTCE_b=log(m_p_Y_1_M_1/(1-m_p_Y_1_M_1)) ///
    -log(m_p_Y_0_M_0/(1-m_p_Y_0_M_0))
*log(NDE)=log(E(Y(1,M(0)))/1-E(Y(1,M(0))))-log(E(Y(0,M(0)))/1-E(Y(0,M(0))))
local lNDE_b=log(m_p_Y_1_M_0/(1-m_p_Y_1_M_0))///
    -log(m_p_Y_0_M_0/(1-m_p_Y_0_M_0))
*log(NIE)=log(TCE)-log(NDE)
local lNIE_b=`lTCE_b'-`lNDE_b'

*log(CDE(0))=log(E(Y(1,0))/1-E(Y(1,0)))-log(E(Y(0,0))/1-E(Y(0,0)))
local lCDE_0_b=log(m_p_Y_10/(1-m_p_Y_10))-log(m_p_Y_00/(1-m_p_Y_00))
*log(CDE(1))
local lCDE_1_b=log(m_p_Y_11/(1-m_p_Y_11))-log(m_p_Y_01/(1-m_p_Y_01))

drop p_Y* m_p_Y* M_*

return scalar lcde_0_b=`lCDE_0_b'
return scalar lcde_1_b=`lCDE_1_b'
return scalar lnde_b=`lNDE_b'
```

```
return scalar ltce_b=‘lTCE_b’
return scalar lnie_b=‘lNIE_b’

end

local i=0
local drop i
foreach ax of numlist -0.7 1{
foreach bu of numlist -0.7 1 2 3{
foreach gu of numlist -0.7 1 2 3{
foreach bm of numlist -0.7 1{
foreach g0 of numlist -2 -1.4 0.4 2.2{
local i=‘i’+1
local bu10=‘bu’*10
local gu10=‘gu’*10
local g010=‘g0’*10
local bm10=‘bm’*10
local ax10=‘ax’*10
di
di "_____________________________________________________________"
di "betam is ‘bm’; betau is ‘bu’; gamma0 is ‘g0’; gammau is ‘gu’; alphax is ‘ax’"
di "you are at the interaction ‘i’ out of 256 "
di
simulate prev_m=r(prev_m) prev_y=r(prev_y) ///
    lCDE_M_0_unb=r(lCDE_M_0_unb) lCDE_M_1_unb=r(lCDE_M_1_unb) ///
    lNIE_unb_M_110=r(lNIE_unb_M_110) lNIE_unb_U_100=r(lNIE_unb_U_100) ///
    lNIE_unb_UM_111=r(lNIE_unb_UM_111) ///
    lTCE_unb=r(lTCE_unb) lNDE_unb_000=r(lNDE_unb_000) ///
    lcde_1_b=r(lcde_1_b) lcde_0_b=r(lcde_0_b) lnie_b=r(lnie_b) ///
    lnde_b=r(lnde_b) ltce_b=r(ltce_b), reps(100) ///
    saving(path\results_‘bu10’_‘gu10’_‘g010’_‘bm10’_‘ax10’_‘i’,replace): ///
    simulations2, obs(1000) s(10) p_x(0.25) alpha0(0.1) alphax(‘ax’) ///
    gamma0(‘g0’) gammax(-.91) gammau(‘gu’) gammaux(0) ///
    beta0(-1.4) betax(0.8) betam(‘bm’) betau(‘bu’) betaxm(0) ///
    betaux(0) betaum(0)
}
}
}
}
}
```