Chapter 6
Pathway Analysis, Causal Mediation, and the Identification of Causal Mechanisms

Leonce Röth

Abstract This chapter presents the systematic analysis of causal mechanisms from the perspective of pathway analysis as an essential complement to conventional approaches to causation. It builds on the evidence that credible causal identification defies design-based strategies such as randomization or linear mediation analysis unless their research designs are supported by reliable mechanistic knowledge. The chapter reasons that the reliable causal identification of a mechanism requires the concept of 'natural indirect effect' and a double-nested counterfactual strategy. It discusses the empirical quantification of causal mechanisms and its underlying assumptions, offers empirical examples that clarify them, and reviews the conditions and limits of the strategy.

Learning Objectives
After studying this chapter, you will be able to:

- Understand the meaning of a mechanism from the pathway perspective.
- Learn how a counterfactual perspective on causality relates to mechanistic thinking.
- Learn how to identify and quantify causal mechanisms using non-parametric procedures.
- Understand why randomization alone does not suffice to identify causal mechanisms.
- Learn how to identify mechanisms when treatment and mediator interact.
- Understand the crucial assumptions under which indirect natural effect estimates equal identified causal mechanisms.

L. Röth (✉)
University of Cologne, Cologne, Germany
e-mail: Leonce.Roeth@uni-koeln.de

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_6

6.1 Introduction

An increasingly popular postulate of causal analysis maintains that good research includes some account of *how* one variable generates another to underpin a causal claim. Causal mechanisms are at the center of research in small-n analyses, often are a crucial part of the theoretical argument in large-n studies, and prove indispensable for scholars of systematic pathway analysis. In some accounts, a credible causal mechanism makes the difference between explanatory and non-explanatory propositions (Waldner, 2007, 146; Kiser & Hechter, 1991, 5; Mayntz, 2004, 14; Hedström, 2008).

Asking not just for a cause of an effect but also for the intermediate process in between is a deeper or second form of asking *why* (Pearl & Mackenzie, 2018, 299–300). The response to this deeper *why* always complements other types of evidence but remains crucial for qualifying the external and internal validity of causal relations. Indeed, mechanisms can raise our confidence in the established validity of a causal association – or undermine it (internal validity). Moreover, their knowledge can change the inference on evidence even from well-executed trials and improve the next experimental setup. This is because mechanisms convey information on the scope conditions of a causal association, which expose the limits of causal effects and their underlying processes (external validity). Besides, knowledge of mechanisms can reveal multiple pathways between cause and outcome, thus guiding us to more effective interventions.

A textbook illustration of these points comes from one of the earliest documented controlled experiments. In 1747, James Lind observed that eating citrus fruits prevents scurvy; understanding and validating the mechanism between citrus intake and scurvy prevention took another 183 years. In the meantime, the link from citrus to scurvy was discredited because the mechanism and its scope conditions remained unknown.[1]

The central intuition about the citrus treatment was that it involved vitamin C – a particular type of acid, later called 'ascorbic' in recognition of its scurvy preventive properties. We now know that vitamin C oxidizes when exposed to heat and light or put in contact with copper. In other words, the citrus treatment only works under specific scope conditions. Back then, however, the juice was heated for conservation, copper pipes were in widespread use, and exposure to light was regular. Thus, many attempts to produce lime juice for sea travels proved ineffective against scurvy.

Furthermore, mechanisms take time to unfold. Today we know that the intake of ascorbic acid activates the synthesis of the enzyme collagen IV. Collagen is a structural protein necessary for healthy blood vessels, muscle, skin, bone, cartilage, and other connective tissues. Ascorbic acid is required for various biosynthetic pathways; when these pathways decay, humans develop a series of symptoms

[1] The startling history of the cure for scurvy is well told in Lewis (1972). Pearl and Mackenzie (2018) recall it to illustrate mediation. This chapter's version enriches the history with some recent knowledge about the causal mechanism, and gives center stage to its scope conditions.

collectively assembled in the diagnosis of scurvy. Moreover, humans cannot synthesize collagen without ascorbic acid and have a low capacity to store it. As collagen IV synthesis stops 4–12 weeks after the last intake of ascorbic acid, symptoms of scurvy start to be visible after 4 weeks. The citrus intake also appeared ineffective for sea travels as the diffusion of steam navigation made many sea trips too short for the symptoms to show. However, Arctic expeditions remained long enough, and many seafarers suffered from scurvy in expeditions until the early twentieth century.[2]

For long, the wrong inference that citrus intake is ineffective for scurvy prevention survived due to the lack of knowledge of the mechanism of activation of collagen IV synthesis. Filling this gap proved crucial for restoring the causal association, as the mechanism disclosed many necessary scope conditions required for it to hold – namely, time, temperature, and exposure to light or copper. These conditions imply that the link between the effect of the treatment and the outcome can only be established in a study period of at least 4 weeks and if the ascorbic acid is kept intact. Moreover, they suggest that the link blurs whenever equivalent pathways are activated – for instance, if seafarers can eat raw meat or any fresh food containing sufficient ascorbic acid. Thus, perfect randomization of citrus intake may not reveal its preventive effect when its design does not take the relevant scope conditions of the mechanism into account.

In short, the knowledge of mechanisms improves three vital criteria of scientific inference – reliability and internal and external validity. But how to study mechanisms systematically?

In the following, I present the answer provided by the particular version of pathway analysis that merges graph theory with a counterfactual model of causality into a powerful framework for identifying mechanisms. This development is roughly 15 years old and still in full swing. It has taken computer science and biology by storm: biostatisticians now usually run millions of pathway models a minute to analyze gene expressions and understand the mechanisms linking a drug treatment and its effect. In comparison, social scientists still seem hesitant to embrace the many benefits that such a pathway perspective can bring. This chapter's first and foremost intention is to reduce hesitation.[3]

To this end, Sect. 6.2 locates the mechanistic why-question in the philosophy of science and discusses the assumptions under which a generic definition of a pathway or mediator[4] can be called 'a mechanism'. Then, Sect. 6.3 discusses how to distinguish between mechanistic associations and causal mechanisms. To this end, it dwells upon a remarkable strength of this method for pathway analysis – a

[2] Notably, the two expeditions of Robert Falcon Scott to Antarctica in 1903 and 1911 suffered greatly from scurvy.

[3] Excellent discussions of causal identification of mechanisms using graph theory are in Morgan and Winship (2015, Chap. 10); Pearl and Mackenzie (2018, Chap. 9); VanderWeele (2015, Part One). This chapter owes almost everything to these contributions. However, it takes a more specific angle on the causal identification of mechanisms in the social sciences.

[4] Note that, in some disciplines, the identification of mechanism is synonymous with causal mediation analysis. Here, instead, mediation is considered a special instance of pathway analysis.

graphical rendering of causal assumptions that helps to lay out the structural conditions under which pathways are causally identified or mistaken. Thus, it clarifies how the graph perspective improves on one of the most applied and cited methods in the history of the social sciences – the so-called Baron-Kenny approach to mediation analysis – and, in so doing, enhances our conditioning strategies.

Section 6.4 discusses the innovative core of pathways analysis – namely, the 'decomposition' and the quantification of the total, direct, and indirect effects on observational data. Indeed, Judea Pearl and others spearheaded a causal revolution when they defined the conditions of causally identified pathways and developed non-parametric formulae to decompose total effects into direct and indirect ones (Pearl, 2022). This quantification strategy of pathway effects took time to be accepted and faced some deep-rooted skepticism from the more conventional quarters of causal analysis (e.g., Rubin, 2004; Rubin, 2005). Nevertheless, social science scholars are slowly getting familiar with indirect effects and their underlying counterfactual theory of causation (see Imbens, 2020).

Section 6.5 replicates one influential model from development economics and sketches another from educational research. The first example demonstrates how strong supposedly mechanistic inference based on innovative cluster randomization in Kenya can be misleading. The second example shows how pathways analysis can draw important mechanistic lessons from a randomized controlled trial run in the United States to seemingly no effect. These examples prove mechanistic knowledge essential to validate and refine even causal evidence from compelling research designs.

The last section of this chapter intends to keep the promises of the pathway approach in check and dispel the illusion that causal identification is a simple technical exercise. As randomized controlled trials or instrumental variable applications show, the devil lies in the detail of the exclusion restrictions; in this respect, pathway causal identification is even more demanding than total effects via randomization or quasi-randomization. Pathway analysis reminds us that our models seldom ensure the perfect causal identification of a mechanism. Indeed, the complexity of the real world typically defies our attempts to draw exhaustive causal maps with analytic tools that require exclusion restrictions. Nonetheless, these restrictions ensure the transparent rigor that qualifies evidence as causal and distinct from mere association.

6.2 Can Pathways Be Mechanisms?

Sometimes, the concepts of mechanism, pathway, and mediation can be confusing. All three terms adhere to the general idea of increasing causal depth by diminishing the contiguity of time and space between cause and outcome. However, what exactly is considered a cause–effect framework and a mechanistic framework is subject to the relative status of a research field and is constantly in flux (see also Chap. 2, Sect. 2.3.1).

What appears to be a sufficiently deep causal mechanism in one particular research tradition and time can be perceived as a superficial association in another. Ideally, research fields increase causal depth over time and remain cautious about the trade-off between desirable specificity and useful parsimony (Craver & Kaplan, 2020). The balance of specificity and parsimony changes while research progresses, and what was considered a mechanism once might be addressed as separate cause–effect relations. Recall from the introduction that it took 183 years to detect the crucial acid for the mechanism between citrus intake and scurvy prevention. During the attempts to isolate ascorbic acid, the intake of vitamin C could have been appropriately described as the causal mechanism. In light of new knowledge, researchers today focus on way more specific biosynthesis pathways as distinct causal relationships. In short, researchers have approached the old mechanism to more causal depth. Philosophers of science call this kind of deepening process "bottoming-out" (see Fig. 6.1) or, in simpler terms, delivering on the demand for the explanation that can stop the infinite regress in causal analysis.

Aiming at fundamental explanations has had a strong appeal for a long time now in the social sciences (see Elster, 1989; Goldthorpe, 2001; Hedström et al., 1998; Hedström & Ylikoski, 2010; Knight & Winship, 2013). Nonetheless, causal mechanisms are also seen as the least understood kind of causal claim (Gerring, 2010; Hedström & Ylikoski, 2010; Waldner, 2012).

Some scholars use the term "mechanism" to refer to a series of events between the original cause and the outcome (Abell, 2004; Mahoney, 2012; Morgan & Winship, 2015; Pearl, 2009, Pearl & Mackenzie, 2018). The concept of "pathway", too, indicates a chain of mediators connecting a cause to an outcome. Thus, some have embraced the term "mechanism" for the analysis of pathways across cases (see Gerring, 2010; Imai et al. 2011; Weller & Barnes, 2014; Woodward, 2003, 350–58; Runhardt, 2015; Morgan & Winship, 2015, 325–352). Other scholars, however, try to exclusively use the term "causal mechanism" for process tracing within single cases (for example, Beach, 2017). These scholars adhere to the "process" or "physical" theories of causation that provide a substantive account of what causal processes are in light of what science tells us about the world (Dowe, 2000, 1–11 and Chap. 10).

Far from a terminological subtlety, these usages point to a fundamental divide over the concept of mechanism. The first group considers causality a matter of epistemology that can be addressed with probabilistic or counterfactual models. From this standpoint, establishing causation is an exercise in logic that many techniques

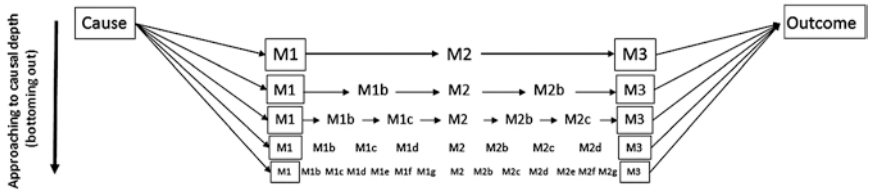

Fig. 6.1 Approaching to causal depth

can perform – provided that they afford comparisons ("type" causality; see Rohlfing & Zuber, 2021, 1634–35). In contrast, the holders of the process theory of causation maintain that causality is necessarily local – which means that it is manifest only in individual cases ("token" causality). Following the process view, within every unique case, causality exists in fine-grained sequences of entities' activities that have to satisfy the criterion of seamless productive continuity (Dowe, 2000). From the perspective of bottoming-out, the process viewpoint on mechanistic causation raises the highest possible demand on causal depth.

A pathway as a sequence of mediators (or interactions) cannot satisfy the ontological criteria established by the process view of mechanistic causation. First, seamless productive continuity can hardly be demonstrated by pathway analysis. Second, the very strength of pathway analysis lies in inferences from comparisons across cases or samples. In short, from the process view on causation, pathways do not deserve the term "mechanism". However, this reservation is a relative rarity in the social sciences. Most scholars are satisfied with an evidential view on mechanisms as a cause-to-effect pathway that at least includes one mediator. Even without satisfying the high demands from the process view, pathway analysts also approach causal depth as they want to know what connects a supposed cause and its outcome at the fundamental level, hence in a general form. As we will see in the next part, the biggest strength of pathway analysis in that ambition for deeper explanations is epistemological. Pathway analysis has developed clear and transparent criteria to distinguish causal mechanisms from mechanistic associations.

6.3 Identifying Causal Mechanisms with Graphs

Causal identification is a general problem independent of the commitment to a mechanistic theory (Pearl, 2009). Pearl's metaphor of a "ladder of causation" renders the solutions to the identification problem as a historical endeavor to more reliable causal knowledge (Pearl & Mackenzie, 2018, 23–52). In this line of thought, scientists moved from the regularity theory over probabilistic theory to the interventionist theory before reaching the top level of the counterfactual theory. As Pearl's argument goes, counterfactuals win the highest pitch as they synthesize and improve on previous solutions to causal identification problems.

From a regularity viewpoint, only the perfect sequence of the candidate cause and outcome constitutes evidence for causation. In our scurvy example, the regularity criterion requires that every citrus intake prevents scurvy without exceptions. The scope conditions of the mechanism demonstrated this bare inference mostly wrong. Under some circumstances, citrus can fail, or the causal effect might be observed without citrus. In Pearl's account, the limits of perfect regularity motivate the shift toward the probabilistic account of causality.

The probabilistic account admits that a causal relation unfolds or fails due to scope conditions and alternative mechanisms but maintains that many of them remain unknown. Hence, our best knowledge about citrus intake can focus on

whether it affects the probability of getting scurvy net of contextual vagaries – that is, on average. However, evidence that a factor affects the probability of an outcome does not constitute evidence for causation either. A limit of the probabilistic approach is that it cannot establish the direction of causation – a problem known as "asymmetry" or "endogeneity". In light of observed probability, for instance, it might also be that scurvy causes lemon intake.

The problem of asymmetry is solved when the candidate cause precedes the outcome. The best way of ensuring this order is to get some control over the candidate causal factor. So, if we prescribe citrus intake to healthy and compliant seafarers once on board, we can gather more convincing evidence of its contribution to the probability of getting scurvy. This approach is at the heart of the 'interventionist' school of causality.

With the asymmetry problem being solved, the thorniest issue of causal identification takes center stage. Even in an interventionist framework, confounders can bias the identification. Thus, we might mistake the sequence of two events as causal despite it being due to a third unobserved factor instead. Logically, the counterfactual theory of causation can discriminate between a confounded relationship and a causal one. The observed event is the real cause when it precedes the outcome, *and* its manipulation resonates with a change in the outcome that would not have occurred without the intervention. Thus, the counterfactual subsumes all preceding approaches to causal identification. Moreover, it embraces the 'would haves' and, on this basis, can offer a single theoretical solution to both asymmetry and confounding problems.

The counterfactual approach is deeply embedded in pathway analysis with graphs. Its notation responds to the problem of asymmetry by using directed arrows to clarify the direction of causality in contrast to the equal sign typical of the regression framework. Directed arrows connect "nodes" or variables in structures of dependency that recall family trees. Thus, the nodes in a path of directed arrows can be indicated as "grand-parent", "parent", "child", and "grand-child." These structures embody strong and weak causal assumptions. An arrow between two nodes indicates a weak causal assumption. It renders the direction of dependency – the fact that values of the child variable change in response to the values taken by the parent variable – but neither its sign[5] nor the size of the causal effect. The strongest causal assumption is the absence of an arrow between two nodes, as it signals that the corresponding variables take their values independently of one another. Furthermore, pathway analysts have introduced the so-called "*do*-operator" to mimic an intervention on an arrow and model the effect of its removal on observational data. This operator marks a relevant difference from conventional counterfactual studies based on non-intervention.

[5] However, some biologists introduced a distinction in the notation of the positive and the negative effects.

6.3.1 Closing the Backdoor

Graph theory offers a transparent strategy to tackle the two crucial problems of causal identification, namely, asymmetry and confounding. Figure 6.2 illustrates the task in its simplest form.

On the left-hand side of Fig. 6.2, we see the identification for the total effect framework, as in a typical correlation or regression analysis. To declare the association between X and Y causal, we first need to demonstrate that X precedes Y and not the other way around. This assumption is embodied in the direction of the arrows. The second task is to check that the association between X and Y is not confounded by third factors such as C. Path X ← C → Y is a so-called "open back-door path" and can be seen as a pipe where non-causal variance is flowing that confounds the true relationship between X and Y. Back-door paths can be closed in two ways. First, by conditioning on C. If we can hold C constant, the back-door paths between X and Y are closed, and the association between X and Y is not confounded anymore. To hold confounders constant is a common identification strategy – for example, in multivariate regressions where we regress Y on X and condition on C (Pearl & Mackenzie, 2018, 157). A second widespread approach is the randomization of X. If we assign the treatment condition of X randomly, all associations running into X are broken, and, therefore, all back-door paths are closed (compare middle part of Fig. 6.2). Experimental designs build on the randomization of the treatment. In quasi-experimental designs – such as regression discontinuity or instrumental variables – randomness in the assignment to treatment arises indirectly from natural factors or events independently of the causal channel of interest (see Chap. 3). If we can rule out both reversed causality and confounding, the associations between X and Y imply causation by necessity. The power of the back-door criterion is that it reveals under which conditions associations are causal even based on observational data.

In a mechanistic framework, the two conditions for a causal interpretation of associations are the same: X needs to precede Y, and all back-door paths between X and Y need to be closed, as on the right-hand side of Fig. 6.2. However, these conditions allow the causal interpretation of the total effect between X and Y, not the causal interpretation of the other quantities of interest to a mechanistic framework – namely, the effect of X on M (X → M, M being the mediator), and the effect of M on Y (M → Y; Y being the outcome). More conditions must be fulfilled to allow for a causal interpretation of the associations b and c on the right-hand side of Fig. 6.2.

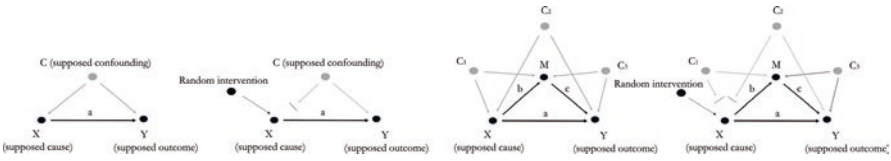

Fig. 6.2 Causal identification with and without a mechanism

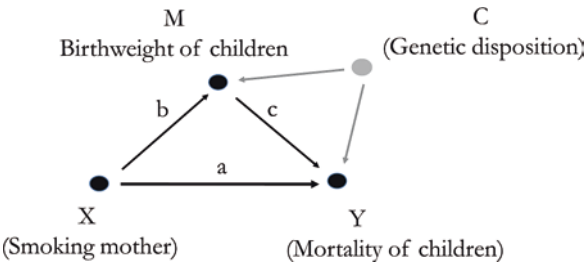

Fig. 6.3 Collider bias in mediation analysis

X has to precede M, and M has to precede Y. Furthermore, all three associations (a, b, and c) have to be un-confounded to reveal the 'true' causal effect from X → M, from M → Y, and the remaining effect of X → Y. In that framework, the total effect equals the sum of the effect from X over M to Y (the *indirect* effect) and the remaining effect of X on Y (the *direct* effect).

If we randomize the treatment X of a mediation model, the randomized treatment blocks all arrows running into X. In the example on the right-hand side of Fig. 6.2, the randomization means ruling out the confounding of C1 and C2 so that the total effect of X on Y still is the true causal effect. However, even with a randomized treatment, we are still unable to quantify the indirect effect. The reason is that C3 is left unconditioned and confounds the relationship between M and Y (path c). Randomization of the treatment does close all back-door paths running into X but does not suffice to identify mechanisms. Unfortunately, the problem of potential confounding between M and Y runs even deeper.

Figure 6.3 represents a famous causal model of the effect of smoking on child mortality. It represents precisely the constellation described on the right-hand side of Fig. 6.2 and represents a fundamental problem of mechanistic identification, the collider bias. The collider bias has troubled statisticians for centuries and led to uncountable false inferences, the birth-weight paradox just being a prominent example.[6]

Let us consider the example in Fig. 6.3. In the mid-1960s, Jacob Yerushalmy pointed out that smoking during pregnancy seemed to benefit the health of children if the baby happened to be born underweight – the so-called "birth-weight paradox" (see Yerushalmy, 1971).[7] Until 2006, this paradox remained unexplained.

In an extensive data set, Yerushalmy found unexpected relationships. Babies of smokers were lighter than babies of non-smokers. However, within the group of low-birth-weight babies, the babies of smoking mothers had a better survival rate than those of non-smokers. It was as if the mother's smoking had a protective effect within the group of babies being born underweight. The inference was that "there is no causal path from smoking to mortality" (Yerushalmy, 1971). How come?

Yerushalmy's findings are the consequence of a problematic conditioning strategy. He was unaware of the importance of genetic disposition and operated under

[6] It likely was Barbara Burks who first modeled the problem using causal graphs in 1926.

[7] An excellent discussion of the birthweight paradox can be found in Wilcox (2006).

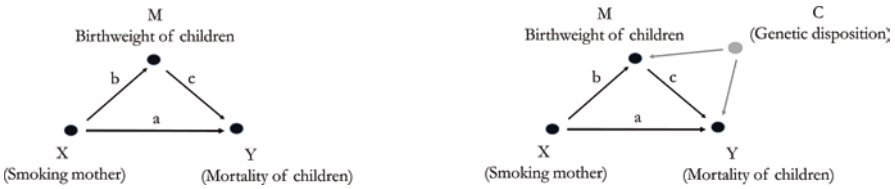

Fig. 6.4 Collider bias in mediation analysis

the assumption of the left model in Fig. 6.4. However, even within that model, it does not make sense to condition on birthweight. Birthweight is not a confounder, but a mediator. Conditioning on the mediator means correcting for the variance that runs through it. In the example, it means controlling for the *indirect* effect of birthweight. The remaining effect of X on Y is typically seen as the *direct* effect.

Conditioning on a mediator is justified to separate the indirect effect (X → M → Y) from the direct one (X → Y). As such, it lies at the heart of the conventional mediation analysis. Indeed, conventional mediation analysis compares effect estimates of the cause based on two separate regressions. The crucial difference runs between the estimate of the coefficient of X on Y in a model without a mediator and in one conditioned on the mediator. As an illustration, if 100% of the variance of the effect from cause X runs through mediator M, conditioning on M leads to a null coefficient of the cause. Baron and Kenny (1986) define three necessary, but not sufficient, conditions for detecting mediation along these lines[8]:

- X has to be significantly related to M.
- M has to be significantly related to Y.
- The total association between X and Y has to decrease when M is kept in the model.

This reasoning allows inferring four types of mediations based on how the effect between X on Y changes when we condition on M (see Fig. 6.5).

Conventional mediation analysis speaks of 'full mediation' when the total variance is associated with the path from X via M to Y (indirect effect), and the direct effect of X on Y leaves nothing unexplained. "Partial mediation" is inferred from a reduced direct effect of X on Y after conditioning on the mediator. "No evidence for mediation" is inferred when the conditioning on the mediator does not affect the direct effect from X on Y. Finally, "inconsistent mediation" is inferred when the adjustment on the mediator reverses the direction of the effect of X on Y.

The birth weight paradox is an instructive example of inconsistent mediation. The reason is that the most prominent factor for low birth weight is a specific genetic disposition that sorts an even higher impact on mortality than smoking. Genetic dispositions confound the path M → Y, as illustrated on the right-hand side of

[8] Note that this paper is one of the most cited papers in scientific history.

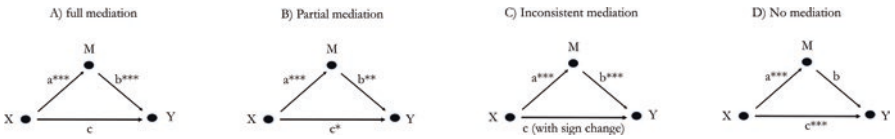

Fig. 6.5 Types of mediation. (**Note:** *** refers to the level of significance)

Fig. 6.4. It is easy to see that Yerushalmy overlooked an important confounder; what is not so easy to see is that Yerushalmy conditioned on a *collider*.

A collider is given when the same outcome depends on two different causes or, in graphical terms, when at least two arrows point to the same node. In Fig. 6.4, birthweight is a mediator (X → M → Y) and a collider (X → M ← C). Adjusting for the collider means opening a closed back-door path from X over C to Y. In other words, conditioning on birthweight creates a spurious positive association between the smoking of mothers and children's survival because genetic dispositions confound the relationship between birth weight and child mortality.

In short, Yerushalmy's surprising findings follow from this troublesome conditioning strategy. Conditioning on birth weight leads to an entirely new comparison within the stratum of children with low weight at birth. Within this new stratum, smoking mothers seem to affect babies' survival positively. However, this association is spurious. Genetic disposition has an even stronger effect on birth weight than smoking, and unless controlled for, it biases the association between birth weight and child mortality.

The graph-theoretical solution of the birth weight paradox offers at least two important lessons. First, while conditioning on confounders closes back-door paths and yields unbiased associations, conditioning on mediators and/or collider variables leads to biased associations. Second, and more important for the causal identification of mechanisms, standard mediation analysis proves unreliable. Conditioning on a collider has caused uncountable "mediation fallacies" (Pearl & Mackenzie, 2018, 315). Despite the increased awareness, the pervasiveness of the problem can still be underestimated. Indeed, mediation fallacies are not limited to the cases of inconsistent mediation. Instead, they may affect all types of conventional mediation with significant consequences. If a collider cannot be ruled out, regression-based mediation analysis cannot be trusted to produce reliable effect estimates as we cannot quantify the bias introduced by conditioning on the mediator.

Figure 6.6 illustrates a more complex causal system where we might be interested in the relative importance of pathway X → M1 → M2 → Y versus pathway X → M3 → Y. This identification task clearly falls beyond the possibilities of the regression framework and demands the more powerful approach to pathway analysis that graphs afford instead.

The overall model entails 11 variables and consists of 16 paths. The back-door criteria guide us to an effective conditioning strategy. There is no confounding between X and Y and the total effect represents the true causal effect, as we declare the causal system exhaustive. However, estimating the indirect effect of the two

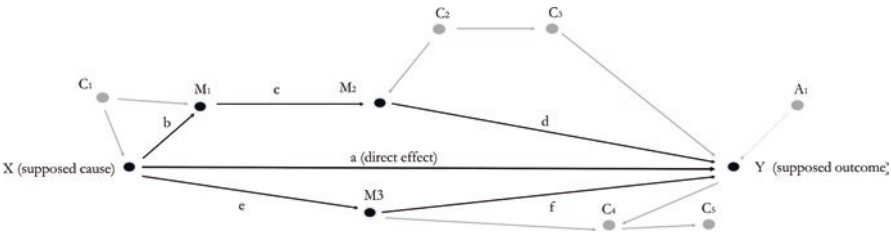

Fig. 6.6 More complex pathways

pathways of interest requires conditioning. The effect of path b is biased unless we condition on C1. The effect of path d is biased unless we condition on C2, C3, or C2 and C3 – conditioning on any of these confounders blocks the back-door path M2 ← C2 → C3 → Y effectively. A1 could be considered an alternative explanation for Y on which it is unnecessary to condition because it does not affect the quantities of interest. C4 and C5 should not be conditioned on: C4 is a collider and would open the non-active backdoor path M3 → C4 → C5 → Y; similarly, C5 should not be conditioned because of the extended collider rule that even 'descendants' of colliders, too, activate back-door paths.

The overall goal of the conditioning strategy guided by the back-door criterion is to block all the paths that generate non-causal associations between the cause and the outcome without inadvertently blocking any of the paths that generate the causal effect itself (Morgan & Winship, 2015, 109). Conditioning on C in Fig. 6.2 is a viable option whereas conditioning on M in Fig. 6.3 opens an otherwise closed back-door path. Eventually, with Morgan and Winship (2015, 109), the back-door criterion can be defined as follows:

> If one or more back-door paths connect the causal variable to the outcome variable, the causal effect is identified by conditioning on a set of variables Z if
>
> ***Condition 1*:** All back-door paths between the causal variable and the outcome variable are blocked after conditioning on Z, which will always be the case if each back-door path
>
> (a) Contains a chain of mediation, where the middle variable is in Z or
> (b) Contains a fork of mutual dependence, where the middle variable is in Z or
> (c) Contains an inverted fork of mutual causation, where the middle variable and all of its descendants are not in Z
>
> and
>
> ***Condition 2*:** No variables in Z are descendants of the causal variable that lie on any of the directed paths that begin at the causal variable and reach the outcome variable.

However, closing the back-doors is only one of two possible identification strategies.

6.3.2 Closing the Front Door

The front-door criterion provides another interesting identification strategy derived from causal graph theory in cases where essential confounders remain unobserved. For example, let us turn to the prize-winning paper on skills and the labor market by

Glynn and Kashin (2018). Glynn and Kashin applied the front-door criterion to a well-known dataset on the effect of the Job Training Partnership Act (JTPA). The Act institutes a job training program to equip participants with different skills. The dataset contains data on the people who applied for the program, whether they showed up, and their earnings over 18 months. The study includes a randomized control trial (RCT) and an observational component. Figure 6.7 provides the causal graphs of the general problem (left), the example (middle), and the front-door approach (right).

The variable *signed up* records whether a person did enroll to the job training, the variable *showed up* whether the enrollee did use the services. The program can only affect the earnings if users showed up, so the absence of a direct arrow between *signed up* to *earnings* can be easily justified. In other words, the entire effect is mediated. Let us say cause, outcome, and mediator are all affected by the general motivation of an applicant, but unfortunately, we have not measured motivation. In a causal graph, an unmeasured variable is typically depicted by a hollow node.

The logic of the front door is to block all paths running into M – in other words, to shield the mediator. In the example of Fig. 6.7, we might randomly call applicants off and compare the randomly canceled applicants with those given real training. With all front-door paths being closed, the estimates of paths b and c can be calculated and are unbiased by definition. In that example, absent a direct effect, the indirect effect equals the total effect, and the estimate using the front-door equals the estimate based on the randomization of X. Glynn and Kashin compared the front-door predictions with those from a randomized controlled experiment, and found the results very similar (Glynn & Kashin, 2018).

The front-door approach could remove almost all of the bias introduced by the omission of the confounder of motivation. In contrast, a simultaneous estimation using the back-door without the possibility of conditioning on motivation showed substantial differences to both the experimental results and the front-door approach (Glynn & Kashin, 2017, 2018).

With Morgan and Winship (2015, 333–334), the front-door criterion can be defined as follows:

> If one or more unblocked back-door paths connect a causal variable to an outcome variable, the causal effect is identified by conditioning on a set of observed variables, M, that make up an identifying mechanism if

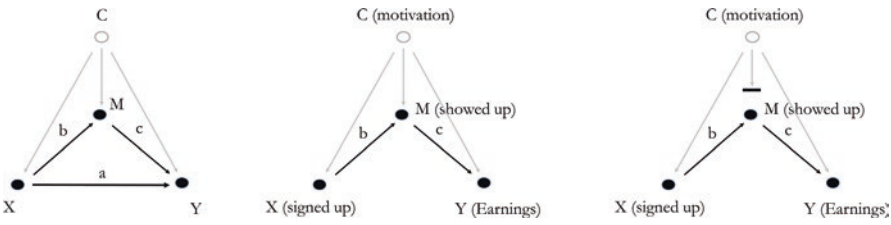

Fig. 6.7 How to shield a mediator

Condition 1 (*exhaustiveness*): The variable in the set M intercepts all directed paths from the causal variable to the outcome variable.

and

Condition 2 (*isolation*): No unblocked back-door paths connect the causal variable to the variables in the set M, and all back-door paths from the variables in the set M to the outcome variable can be blocked by conditioning on the causal variable.

At this point, we have learned two different ways to identify causal mechanisms. By definition, closing all back-door paths or closing all front-door paths leads to causal estimates even with observational data. The logic of back-door paths explains why the identification of indirect effect is neither ensured by the randomization of the cause nor by conditioning on the mediator as applied by conventional regression-based mediation analysis. The next section discusses how indirect and direct effects can nonetheless be identified.

6.4 Identifying Indirect Effects

For a long time, mediation analysts defined:

$$\text{Total Effect} = \text{Direct Effect} + \text{Indirect Effect}$$

This formula understands the indirect effect as a residual category. The Baron-Kenny approach (1986) is entirely built upon this logical pillar. As a straightforward consequence, the conventional approach advised conditioning on the mediator to arrive at the direct effect and, in force of the composition assumption, calculating the indirect effect of mediation as the total minus the direct effect.

The first problem, as already seen, is that the composition stands if M and Y are not confounded or, in other words, if a collider bias can be ruled out. The second problem is that the estimate of the residual is only credible in strictly linear systems. Once we relax the linearity assumption, the composition rule fails (Pearl & Mackenzie, 2018, 322–336).[9]

6.4.1 Indirect Effect in Non-linear Systems

The language of indirect, direct, and total effects evolved in the 1970s, but only recently was the indirect effect defined in causal terms. This shift entailed embracing counterfactual thinking.

[9] The problem of conventional mediation analysis is very fundamental. Mediation analysis based on the difference methods (Baron & Kenny, 1986; Judd and Kenny, 1981) and linear regression models suffer from problems in the presence of interactions, non-linearities, binary outcomes, unobserved confounders, and other modeling complications (see Shpitser, 2013).

Let us start with the direct effect using the *do*-calculus. In the simple graph of treatment (X), mediator (M), and outcome (Y), we get the direct effect of X on Y when we intervene on X without allowing M to change. We *do*(M = 0) and randomly assign units to *do*(X = 1) or *do*(X = 0). We call this the 'controlled direct effect' or CDE.

CDE(0) raises when we force the mediator to take on the value of zero and can be computed as

$$CDE(0) = Pr\left(\mathrm{Y} = 1 \mid do\left(\mathrm{X} = 1\right), do\left(\mathrm{M} = 0\right)\right) - Pr\left(\mathrm{Y} = 1 \mid, do\left(\mathrm{X} = 0\right) \mid, do\left(\mathrm{M} = 0\right)\right)$$

Had we forced the mediator to be 1, we would have denoted the resulting controlled direct effect as CDE(1). In practice, however, this alternative strategy could prove unwise as it forces M on instances of X that are potentially implausible to observe. Moreover, inferring the direct effect from the difference between CDE(1) and CDE(0) is to infer from an over-controlled experiment.

The so-called 'natural direct effect' or NDE offers an alternative perspective. We randomize X, but let M take the value it would naturally do. The 'would' indicates that a counterfactual is required and can be calculated as follows:

$$NDE = Pr\left(\mathrm{Y}_{\mathrm{M=M0}} = 1 \mid do\left(\mathrm{X} = 1\right)\right) - Pr\left(\mathrm{Y}_{\mathrm{M=M0}} = 1 \mid do\left(\mathrm{X} = 0\right)\right).$$

The NDE subtracts the probability of having a positive outcome without the treatment (X = 0) under M equal to zero from the probability of having a positive outcome with the treatment (X = 1) again under null M. In short, the NDE holds the mediator constant while the treatment is forced toward specific values. Indirect effects, unlike direct effects, have no controlled version because there is no way to disable the direct path by holding some variable constant.

Indirect effects have a natural version, too, which again requires thinking in counterfactual terms. The natural indirect effect (NIE) is when we would abstain from the treatment, but allow the mediator to be present. Understanding the causal properties of the indirect effect requires a double-nested counterfactual. In formal terms, we can define the natural indirect effect as follows:

$$NIE = Pr\left(\mathrm{Y}_{\mathrm{M=M1}} = 1 \mid do\left(\mathrm{X} = 0\right)\right) - Pr\left(\mathrm{Y}_{\mathrm{M=M0}} = 1 \mid do\left(\mathrm{X} = 0\right)\right)$$

The first term indicates the probability of a positive outcome under absent treatment and present mediator. From this quantity, we subtract the probability of the positive outcome under the 'natural' situation where both the treatment and mediator are given.

The counterfactual M1 must be computed for each observation on a case-by-case basis. This requirement places the natural indirect effect out of the experimenters' reach as they may not know the value of the mediator M1 for any particular

treatment X at the level of the individual unit. However, assuming there is no confounding between X and M as well as M and Y (i.e., ruling out the confounding and the collider bias), the NIE can still be computed on observational data. The natural indirect effect entails denying the treatment to anyone, and letting the mediator take the value it would have in the presence of the counterfactual treatment for each individual. The difference yields Pearl and Mackenzie (2018, 333) mediation formula as follows:

$$NIE = \sum_{m} \left[Pr\left(\mathrm{X} = 1\right) - Pr\left(\mathrm{X} = 0\right) \right] \cdot Pr\left(\mathrm{Y} = 1 \mid, \mathrm{X} = 0 \mid, \mathrm{M} = m\right)$$

The expression stands for the effect of X on M in the subset of the units where the mediator takes the value *m* (in square brackets) times the probability that Y = 1 when X = 0 and the mediator takes the value *m*. So formulated, the NIE exposes the source of the product-of-coefficients idea and casts the product of two non-linear effects. Moreover, this formula allows calculating what is *explained by* mediation and the percentage *owed to mediation*.

6.4.2 Indirect Effect When the Cause and the Mediator Interact

The identification of indirect effects becomes more complex when the mediator and the supposed cause (or "exposure") interact. A unified perspective on the decomposition of the total effect in a case where the independent variable of interest interacts with the mediator has been provided by VanderWeele (2014).

So far, effect decomposition has meant to split a total effect into an indirect and direct one. In the presence of exposure-mediator interaction, two components need to be added: the one due to interaction only; the other due to mediation and interaction (see VanderWeele, 2014, 751). The counterfactual assumptions to identify the effect quantities are similar to those required to analyze causal mediation without interaction. As in the case of causal mediation, indirect effects including interactions require double-nested counterfactuals, whereas the direct effect requires weaker assumptions. The attribution of the interaction quantities to either the indirect or direct effect, instead, remains an empirical question. Figure 6.8 illustrates two possible response strategies based on VanderWeele (2014, 757).

The fourfold decomposition depicted in Fig. 6.8 encompasses both decompositions for mediation and interaction.

For interaction, the reference interaction (INT*ref*) and the mediated interaction (INT*med*) combine to the portion attributable to interaction (PAI). The portion attributable to interaction (PAI) combines with the controlled direct effect (CDE) and the pure indirect effect (PIE) to give the total effect (TE).

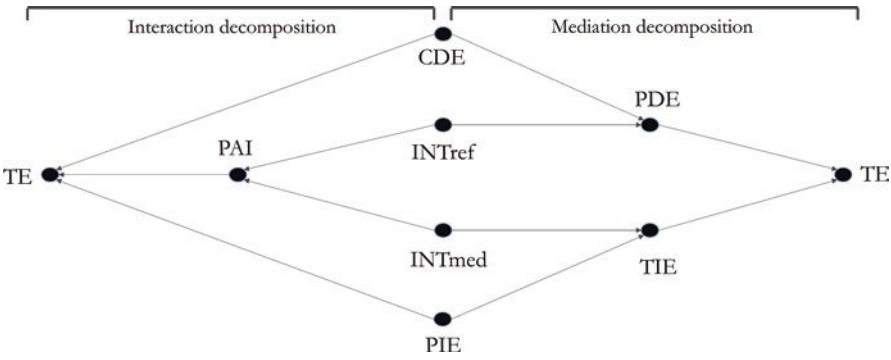

Fig. 6.8 Fourfold decomposition

For mediation, the controlled direct effect and the reference interaction (INT*ref*) combine to give the pure direct effect (PDE); the pure indirect effect (PIE) combines with the mediated interaction (INT*med*) to give the total indirect effect (TIE), and the pure direct effect (PDE) combines with total indirect effect (TIE) to give the total effect (TE).

6.4.3 Wrapping Up

The graph theory reveals that the identification of causal mechanisms requires counterfactuals. The natural indirect effect is when we abstain from the treatment, but the mediator is present. Contrasted with the state where both the treatment and the mediator are present, we can quantify how much of the effect of X on Y is captured by the mediator M, and how much of Y is owed to the mediator M alone. Such a natural indirect effect gauges a causal mechanism once the back-door criterion is satisfied, e.g., all back-door paths are closed.

The consequences of this definition are far-reaching. The identification of causal mechanisms appears as out of reach to the conventional mediation analysis than to randomization. What appears as bad news can also be a good insight, as the natural indirect effect yields a mediation formula stripped of any parametric assumptions. Under some assumptions, this formula allows quantifying the causal mechanism based on observational data. Section 6.5 demonstrates this claim with the example of a renowned identification debate.

6.5 Applications

6.5.1 A Mechanistic View on the Worm Wars

In this application case, I add a causal mediation view to the "worm wars" – a famous debate over the interpretation of influential cluster randomization in Kenya that, besides other studies, brought one of its authors, Michael Kremer, the Nobel Memorial Prize in Economic Sciences in 2019.

The study originates from the evidence that nearly two billion people worldwide – mostly children – are infected by intestinal worms. These species inhabit the human digestive tract; they spread by expelling their eggs via the body waste of infected people. Without good sanitation, these microscopic eggs can find their way, unnoticed, onto the skin or food of another person. Once someone ingests an egg, the reinfection cycle continues. Poor sanitation facilities and hygiene practices allow infections to spread locally.

In 2004, a landmark study showed that an inexpensive medication to treat parasitic worms could improve health and school attendance for millions of children in many developing countries (Miguel & Kremer, 2004). Eleven years later, a headline in *The Guardian* reported that the deworming treatment had been debunked. In 2021, a carefully exercised replication study restated the original findings (see Ozier, 2021). Why so?

Miguel and Kremer convincingly argued that, due to the infectiousness of the worms, individual treatments are unlikely to be effective because children will quickly re-infect themselves with other children. Consequently, they run an encompassing field experiment in Kenya using cluster randomization at the school level. The experiment compared more than 25,000 treated children across three waves to a control group for each wave with similar attributes except for the suppressed treatment. They found a remarkable effect of the treatment on school attendance not only in the treatment area (up to 3 km) but also in the surrounding areas (3–6 km from the treatment).

Replication analyses have mainly confirmed the direct effect in the treatment areas. However, the spillover effects became subject to debate and turned insignificant in some specifications (for example, Aiken et al., 2014). The debate about the replication involved many influential scholars, was covered by several blogs, and eventually came to be known as the "worm wars". A systematic review of the debate seemed to restore the trust in the key findings of the original study. Ozier (2021) concluded that, if anything, years of debates and replication have reinforced his belief in the main effect. In short, it appeared as if the treatment of Miguel and Kremer had indeed sorted a substantial positive impact on children's school attendance.

However, there is a second line of skepticism, less concerned with the significance levels of the total effects but with the plausibility of the indirect effect. The indirect effect, as we have learned, considers the probability of a positive outcome (school attendance) given that we do not have a treatment (no de-worming drug

intake), but we set the mediator (being, in fact, de-wormed) to the values as if we would have had treatment (de-worming drug intake). We contrast this with the probability of a positive outcome (school attendance) under natural conditions where the treatment is given (de-worming drug intake) and the mediator too (being de-wormed). Based on Pearl's mediation formulae, we can compute the natural indirect effect using observational data. The results can be given a causal interpretation if we can exclude confounding between the mediator (being de-wormed) and the outcome (school attendance).

This mechanistic perspective on the study is of great interest for at least two reasons. First, experts in deworming cast considerable doubt on the findings. Epidemiologists refused to include the paper in a meta-study for methodological reasons (no blinded treatment was performed) and referred instead to existing epidemiological studies that, if at all, showed very modest effects of deworming on school attendance. In other words, the authors of a Cochrane review were unconvinced that de-worming could have had such a substantial effect as reported in Miguel and Kremer (Taylor-Robinson et al., 2015). Second, the authors of the original experiment framed their study and their results as if they had strong evidence for the entire mechanism. In the words of the authors' abstract, "[*d*]*eworming substantially improved health and school participation among untreated children in both treatment schools and neighboring schools, and these externalities are large enough to justify fully subsidizing treatment*." (Miguel & Kremer, 2004, 159). In short, the authors' inference is that their evidence point to a clear recommendation for subsidizing de-worming treatments because de-wormed students have a higher likelihood of attending school. Is it the de-worming via the drug intake that causes students to attend school more often?

Based on the original data, the mediation formulae can be used to put the mechanistic claim under scrutiny. Table 6.1 includes all probabilities required to compute the natural indirect, natural direct, and the total effect based on the replication data of Miguel and Kremer (2014), Miguel et al. (2014).[10] By relating indirect and direct effect quantities to the total effect, we can draw valuable conclusions. The natural indirect effect supports the suspicion of the epidemiologists. Only 1.8% of the total effect would be achieved by worm-free students alone. In contrast, 94.2% of the total effect is related to the natural direct effect of the treatment other than

[10] For the replication, I use a very simple model based on the drug treatment in the first period of the field experiment. The experiment had three waves, but the comparison groups changed during the waves and because the effect on school attendance is predominantly a result of the first wave, I focus on the first wave only. For the mediator, I use the reversed indicator of any moderate or heavy worm infection based on the WHO standard in 1999. I see the mechanism present when a treated student is indeed free of worms. For the outcome, I use a dummy of students being present in school at times of the surprise visit. The current documentation of the data is exemplary (see Miguel and Kremer, 2014; Miguel et al. 2014; Hicks and Nekesa, 2014).

Table 6.1 Probabilities of the treatment, the mechanism, the outcome and the natural direct (NDI), indirect (NIE), and total effect (NTE)

Treatment condition, mediator condition, and outcome probabilities						
Treatment	Dewormed	Present in school (in %)			Treatment	Dewormed (in %)
Yes	Yes	0.90			No	0.55
Yes	No	0.86			Yes	0.59
No	Yes	0.86				
No	No	0.85				
				Inference		
NIE	0.05	NIE/TE	1.8	1.8% of the school attendance effect would be achieved by worm-free students alone		
NDE	2.7	NDE/TE	94.2	94.2% of the attendance effect is related to the treatment other than deworming students		
TE	2.9	1-NDE/TE	5.8	5.8% of attendance effect is owed to the capacity of the treatment to deworm students		

Note: Compare equations for NIE, NDE, and TE above.

deworming students. Finally, 5.8% of the effect on attendance is owed to the capacity of the treatment to deworm students.[11]

How do we make sense of these numbers?

Humphreys (2015) documented and commented on the worm wars in close detail, driven by concerns for the mechanistic element of the study. He points to several important aspects that can be learned from the documentation of the experiment. Based on background information and the skeptical comments of epidemiologists, we might add several pathways between treatment and outcome (see Fig. 6.9). The causal graph reveals that the estimate above of the natural indirect effect is not identified. There is nothing identified in this system of pathways because too many nodes are unobserved. Let us briefly describe the pathways in Fig. 6.9.

One element of the treatment is the drug intake that seems to effectively de-worm students. The effect of de-worming alone is relatively weak, as the path analysis in Table 6.1 confirms. The drug intake has as least two more effects on attendance that cannot be isolated given the existing data. De-wormed students create spillovers, and spillovers might feedback to the treated. This feedback is problematic because it undermines the assumption of the independence of the treatment group and the control group – the problem that compelled resorting to cluster randomization in the first place.

Beyond spillovers, the drug intake can create placebo effects. Students feel better because of the drug, irrespective of being de-wormed, which might increase school

[11] An alternative way of modeling these numbers would be to use readymade packages in software such as R or Stata. In Stata, you would use the model builder and simple graph the mediation model. After the estimation of all path-coefficients, the effects can be decomposed into total, direct, and indirect effects using the teffects command (see Bollen, 1989; Sobel, 1987). Note that this command still assumes linearity and leads to biased estimates in this case.

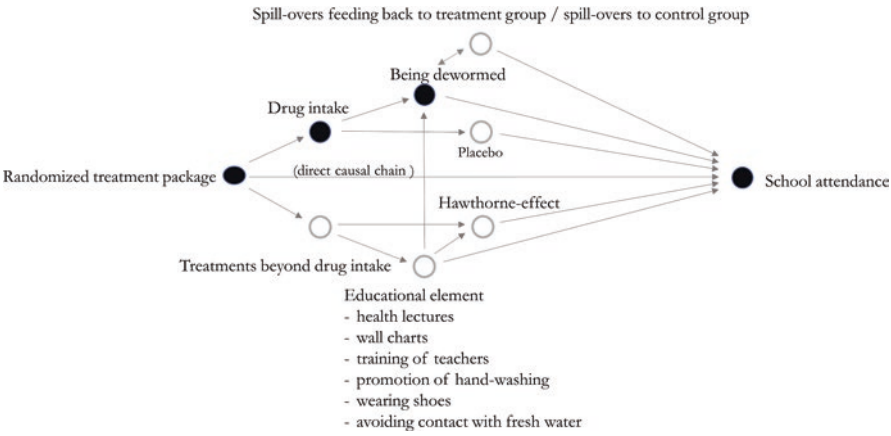

Fig. 6.9 Mechanisms in the worm wars

attendance. Since the control group was not treated with a placebo, we cannot estimate the placebo effect. More worrisome is how the research group treated the treatment group beyond the drug intake. The documentation files list health lectures, wall charts in the schools, training of teachers in the treatment schools, encouragements of the treated students for handwashing, wearing shoes, and avoiding freshwater (see Hicks & Nekesa, 2014, 7).[12] This extensive treatment had obvious health effects – including a contribution to de-worming – which suggests that the treated students likely became well aware of being subject to an encompassing treatment package. Thus, at least three more paths follow from that treatment beyond drug intake.

First, the educational elements on health issues might have affected the well-being of students besides de-worming, which raises their probability to be present in school. Second, being so obviously treated might activate the Hawthorne effect, the rising willingness of participants to make the experiment a success in light of the efforts experimenters provided for the treated. For example, teachers might just encourage students in the treatment group to show up because they know that school attendance is an important measure (although it has to be noted that the measurement of school attendance was achieved by surprise visits). Third, health education

[12] The educational treatments at the school level were part of a separate intervention of the same NGO and could in principle be controlled based on the data (see Hicks & Nekesa, 2014, 5). In fact Miguel and Kremer condition on those interventions. They write "None of these programs involved health treatments for pupils, and given the cross-cutting design, are unlikely to complicate the identification of average treatment effects across PSDP program and comparison schools." Nonetheless, in many specifications Miguel and Kremer (2004) control for assignment to assistance through these other programs'. Only a page later, they write without considering any potential bias "[t]he educational component of the intervention focused on teaching children about avoiding the disease. Health educators explained the transmission vectors for different types of helminths [one of the relevant worm types] and also promoted hand-washing, wearing shoes, and avoiding contact with fresh water" (2014, 7).

affects the likelihood of being de-wormed besides de-worming drug intake and school attendance. Accordingly, the effect of being de-wormed on school attendance, including the spillover effects, is confounded. Knowing about the direction of the influence of health education (increasing de-worming and school attendance), the already weak indirect effect of de-worming via drug-intake on school attendance is most likely biased upwards. This perspective reveals that the authors make strong mechanistic inference without ever quantifying the importance of their hypothesized mechanism and without noticing that the indirect effect cannot be precisely identified, given the observable data at hand.

Such a mechanistic perspective also reveals the standing of the main criticism of the epidemiologists. The Cochrane reviewers classified the study as very weak in terms of evidence, predominantly because of the lack of placebo treatment of the control group. Indeed, except for the spillover path, all alternative paths between treatment and outcome could have been closed by placebo treatment. The consideration also applies to the educational health elements.

Thus, the mechanistic view qualifies the inference of this landmark study substantially. First, there is a confirmation of a significant indirect effect running from the treatment over being de-wormed to higher school attendance. However, this indirect effect explains a very marginal part of the increased school attendance. Way more important are the indirect effects triggered by the entire treatment package beyond the ability to de-worm students. The rise in school attendance is predominantly a composite of different pathways from the Hawthorne pathway over the health education pathway to a potential placebo pathway, combined around 54 times more powerful for school attendance than the de-worming effect. The overall inference to recommend the distribution of cheap drugs might be replaced by the recommendation to offer supposedly more expensive health education.

To be very clear about it, the study of Miguel and Kremer is comparatively well-executed and deserves to be praised for the logic of cluster randomization alone. Nonetheless, the mechanistic view on this experiment demonstrates that randomization does not allow for mechanistic inference. While the total effect of the treatment package might still be perfectly identified, the mechanistic view helps identify which elements of the treatment have created more or less powerful pathways to the outcome. It is extremely interesting to know how much Hawthorne, placebo, or health education contributed to the substantial rise in school attendance, as such effect decomposition can help to improve similar experiments in the future. Like in the lemon-scurvy example, experimenters need to disable these alternative pathways (exclusion restriction) for getting to the correct inference.

A mechanistic view may help to understand supposedly strong effects in well-executed experiments. Moreover, it can reveal causal mechanisms where experiments seem to yield nothing.

6.5.2 A Mechanistic View on a Chicago School Reform

In 1998, US secretary of education, William Bennet, called Chicago's public school the worst of the nation. However, several reforms in the late 1990s moved them from the worst to 'innovators of the nation'.[13] One of the core reforms involved a program called 'Algebra for All', compulsory prep courses for ninth graders in high school. At first sight, the program seemed a success as math scores rose significantly. However, the qualification of incoming ninth-grade students was already improving due to changes in the K-8 curriculums (an important confounder). Once controlled for this confounder, the reform turned out to be insignificantly related to the math performance of ninth graders. Here, the story would have typically found its end.

Luckily, Professor Guanghei Hong remained curious because she knew that when Algebra for All was introduced, more than the curriculum changed. The lower-achieving students found themselves in classrooms with higher-achieving students and could not keep up. Detrimental effects for students and teachers caused by mixed classes compared to remedial classes are well-known. In short, Mrs. Hong was suspicious of the unanticipated side effects of the treatment package. Testing the classroom environment as a mediator between reform and outcome clearly showed that this pathway had negative consequences. Once taken into consideration, the direct effect turned positive. The lesson seemed clear: removing the mixed classes and keeping the prep courses was the logical consequence and created a success story of the modified Algebra for All program.

Students in Chicago significantly benefited from a mechanistic view on an education program that has, at first sight, falsely been considered a failure. We learn from this example that different mechanisms can cancel each other out ("opposing mediation" as in Kenny [1998]), which demonstrates that even a null finding based on a randomized treatment can be worth considering with closer scrutiny on the level of mechanisms. The Algebra for All example is similar to the discredited causal link between lemons and scurvy prevention, although its revitalization took place in a substantially shorter period.

6.6 Thou Shall Not Raise Causal Illusions

Scholars of pathways have revolutionized our view on causal identification. The counterfactual perspective on pathways reveals that fundamental problems of causality – asymmetry and confounding – can logically be solved by closing either the back- or the front-door. This perspective embraces conventional counterfactual causal inference such as randomization or quasi-experiments. Causal graphs help to make its logic and assumptions very transparent. Applying the logic of the

[13] One of its inventors, Arne Duncan, became secretary of education under Barack Obama.

back-door to generally defined causal mechanisms reveals two things. First, conventional approaches are ill-suited for identifying causal mechanisms as they can mistake their structure. Pathway analysis solved that issue by focussing on indirect effects. This perspective reveals that causal mechanisms can be quantified by non-parametric comparisons of observable with counterfactual probabilities. To lend these numbers a causal meaning depends on a simple assumption: path estimates in a system of pathways must be unconfounded.

This unconfoundedness can unfortunately not be fully ensured by randomization – although the randomization of the treatment helps a lot to block all paths running into the candidate cause. Moreover, causal mechanisms can only be identified if a theoretically exhaustive causal system is given and all confounders are observed and conditioned on. Based on a theoretically defined causal system, effective strategies of de-confounding can be determined. The complexity of the task becomes apparent when we remind ourselves of the problem of the collider bias. The collider bias is an instance of a single confounded path in a system of pathways, leading in the worst of events to completely misleading estimates of the indirect and direct effects – such as when smoking mothers are understood to increase the survival rate of their children. Besides, complex pathways with sequences of many mediators can complicate the identification task and the chances for false inference multiply.

The pathways perspective on the identification of causal mechanisms is logically simple. However, mechanisms can only be identified given a theoretically exhaustive causal system where all the variables required to close the back-doors are measured, free of error, and conditioned. Empirically, these assumptions are hard to meet. Thus, research relying on pathways or causal mechanisms should avoid creating the causal illusion that the back-door criterion will easily tackle identification tasks.

The greater strength of the pathway approach is not to deliver a readymade tool for causal inference but a perspective that can boost the transparency over what is needed to identify a mechanism causally. It complements standard approaches of causal inference that typically seek to identify total effects. Analyses of mechanisms searching for indirect effects ask a deeper form of why. Preliminary answers to these deeper questions can at times be very generic, such as a single mediator connecting cause and outcome, and at times can also span to very complex systems of pathways. However, even the most generic mechanism can reveal a great deal. Thinking of lemons' ability to prevent scurvy, smoking mothers to decrease the survival rate of their children, the capacity of de-worming to increase school attendance or preparation courses to improve school performance. In all examples of this chapter, evidence on a single mediator considerably qualified the inference of a cause–effect relationship.

Despite the capacity of a mechanistic view to qualify the inference of even well-executed experiments, the added values are complementary. Randomized treatments facilitate the identification of causal mechanisms because important sources of confounding are erased by design. Mechanisms, in turn, improve the exercise and

inference on well-executed experiments too. The more we know about the mechanisms, the better we can identify total effects.

Suggested Readings

There are three books of great help to understand causal mediation. The most encompassing work on causal mediation analysis, including moderated mediation, is most likely VanderWeeles' book *Explanation in causal inference: methods for mediation and interaction*, published in 2015 by Oxford University Press. Although probably the most encompassing, it addresses the issue from the perspective of biostatistics. Easier access to causal mediation can prove Chapter 9 on *Mediation: The search for a mechanism* in Pearl and Mackenzie (2018), published by Basic Books. The entire textbook can be highly recommended to cast light on recent developments in causal identification against the background of the history of statistics. Finally, Chapter 10 on *Mechanisms and causal explanation* in Morgan and Winship (2015) lies somehow in between VanderWeeles' equation-based insights and Pearl and Mackenzie's captivating narrative. Their entire book on *Counterfactuals and causal inference* can be recommended, as it covers virtually all causal identification tasks from the perspective of the social sciences while preserving a deep commitment to graph theory and counterfactual thinking.

Helpful Websites

Beyond books, there are two highly informative websites on causal mediation. The one by David Kenny provides regular updates on mediation analysis and also covered issues in causal mediation (http://davidakenny.net/cm/mediate.htm). Alternatively, Columbia University provides information on causal mediation, including a recorded lecture of VanderWeele based on the Harvard Seminar Series in Biostatistics (https://www.publichealth.columbia.edu/research/population-health-methods/causal-mediation#websites).

Software Recommendations

Causal mediation, the identification of mechanisms, or causal pathway analysis are relatively new and characterized by rapid development. Formulas, methods, and software applications change accordingly. Nonetheless, several software packages have proven extremely useful.

1. R *mediation* package (Tingley et al. 2014):
   - the *mediate()* function estimates the natural direct and indirect effects based on Pearl's mediation formula,
   - X-M interaction may be conducted by the function test *TMint()* (significant finding implies that the no X-M interaction assumption does not hold).
   - the sensitivity analysis function *medsens()* allows for investigators to examine, through simulations, the robustness of their findings to potential unmeasured M-Y confounders.

   Results for all analyses are displayed using the *summary()* and *plot()* functions

2. SAS macro:
   - The SAS macro is a regression-based approach to estimating controlled direct and natural direct and indirect effects.
   - The macro can handle virtually every distributional and link assumption (compare Valeri et al., 2013).
3. Stata:
   - *paramed* package (no sensitivity analysis) (Emsley et al., 2013).
   - *ldecomp* (no sensitivity analysis) (Buis, 2010).
   - *medeff* (sensitivity analysis) (Hicks and Tingley, 2011).
   - *gformula* (helpful in case of post-treatment and time-varying confounding) (Daniel et al., 2011).

Review Questions

1. Under which conditions can mechanisms be causally identified?
2. What is a natural indirect effect in comparison to a controlled indirect effect?
3. Why randomization might identify cause-effect relationships but not neccessarily indirect effects?
4. Why might conventional mediation analysis be misleading for the causal identification of the mechanism?
5. How does mechanistic evidence help to improve the implementation of experiments?
6. What are the consequences of treatment-mediator interactions for the identification of mechanisms?
7. What are the limits of mechanistic causal identification?

References

Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? *Annual Review of Sociology, 30*, 287–310. https://doi.org/10.1146/annurev.soc.29.010202.100113

Aiken, A, Davey, C, Hayes, R, & Hargreaves, J. (2014). Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: A pure replication. *3ie replication paper* 3, part 1. Washington, DC: International initiative for impact evaluation (3ie). https://doi.org/10.1093/ije/dyv127.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Beach, D. (2017). What are we actually tracing? Process tracing and the benefits of conceptualizing causal mechanisms as systems. *Qualitative & Multi-Method Research, 14*(1/2), 15–22. https://doi.org/10.5281/zenodo.823306

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

Buis, M. (2010). Direct and indirect effects in a logit model. *The Stata Journal, 10*(1):11–29.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Dowe, P. (2000). *Physical causation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511570650

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

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812255

Emsley, R. & Liu, H. (2013). "PARAMED: Stata module to perform causal mediation analysis using parametric regression models," Statistical Software Components S457581, Boston College Department of Economics.

Gerring, J. (2010). Causal mechanisms: Yes, but…. *Comparative Political Studies, 43*(11), 1499–1526. https://doi.org/10.1177/0010414010376911

Glynn, A. N., & Kashin, K. (2017). Front-door difference-in-differences estimators. *American Journal of Political Science, 61*(4), 989–1002. https://doi.org/10.1111/ajps.12311

Glynn, A. N., & Kashin, K. (2018). Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association, 113*(523), 1040–1049. https://doi.org/10.1080/01621459.2017.1398657

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociological Review, 17*(1), 1–20. https://www.jstor.org/stable/522622

Hedström, P. (2008). Studying mechanisms to strengthen causal inferences in quantitative research. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 319–335). https://doi.org/10.1093/oxfordhb/9780199286546.003.0013

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*(1), 49–67. https://doi.org/10.1146/annurev.soc.012809.102632

Hedström, P., Swedberg, R., Hernes, G., & (Eds.). (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511663901

Hicks, J., & Nekesa, C. (2014). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Codebooks. Available at Havard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038

Hicks, R., & Tingley, D. (2011). mediation: STATA package for causal mediation analysis.

Humphreys, M. (2015). *What has been learned from the deworming replications: A nonpartisan view*. http://www.columbia.edu/~mh2245/w/worms.html [retrieved 01.11.2021].

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature, 58*(4), 1129–1179. https://doi.org/10.1257/jel.20191597

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765–789.

Judd, C. & Kenny, D. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*(5), 602–619.

Kiser, E., & Hechter, M. (1991). The role of general theory in comparative-historical sociology. *American Journal of Sociology, 97*(1), 1–30. https://doi.org/10.1086/229738

Knight, C., & Winship, C. (2013). The causal implications of mechanistic thinking: Identification using directed acyclic graphs (DAGs). In L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 275–299). Springer. https://doi.org/10.1007/978-94-007-6094-3_14

Lewis, H. E. (1972). Medical aspects of polar exploration: Sixtieth anniversary of Scotts last expedition: State of knowledge about scurvy in 1911. *Proceedings of the Royal Society of Medicine, 65*(1), 39–42. https://doi.org/10.1177/003591577206500116

Mahoney, J. (2012). The logic of process tracing tests in the social sciences. *Sociological Methods & Research, 41*(4), 570–597. https://doi.org/10.1177/0049124112437709

Mayntz, R. (2004). Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences, 34*(2), 237–259. https://doi.org/10.1177/0048393103262552

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica, 72*(1), 159–217. https://doi.org/10.1111/j.1468-0262.2004.00481.x

Miguel, E. & Kremer, M. (2014). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Guide to Replication of Miguel and Kremer (2004). Available at Havard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038

Miguel, E., Kremer, M., Hicks, J. & Nekesa, C. (2014). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Data User's Guide. Available at Havard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.. https://doi.org/10.1017/CBO9781107587991

Ozier, O. (2021). Replication Redux: The reproducibility crisis and the case of deworming. *The World Bank Research Observer, 36*(1), 101–130. https://doi.org/10.1093/wbro/lkaa005

Pearl, J. (2009). *Causality*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J. (2022). Direct and indirect effects. In Geffner, H., Dechter, R., & Halpern, J. Y. (Eds.). Probabilistic and Causal Inference: The Works of Judea Pearl (pp. 373–392).

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Rohlfing, I., & Zuber, C. I. (2021). Check your truth conditions! Clarifying the relationship between theories of causation and social science methods for causal inference. *Sociological Methods & Research, 50*(4), 1623–1659. https://doi.org/10.1177/0049124119826156

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31*(2), 161–170. https://doi.org/10.1111/j.1467-9469.2004.02-123.x

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331. https://doi.org/10.1111/j.1467-9469.2004.02-123.x

Runhardt, R. W. (2015). Evidence for causal mechanisms in social science: Recommendations from Woodward's manipulability theory of causation. *Philosophy of Science, 82*(5), 1296–1307. https://doi.org/10.1086/683679

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science, 37*(6), 1011–1035. https://doi.org/10.1111/cogs.12058

Sobel, M. E. (1987). Direct and indirect effects in linear structural equation models. *Sociological Methods and Research, 16*, 155–176. https://doi.org/10.1177/0049124187016001006

Taylor-Robinson, D. C., Maayan, N., Soares-Weiser, K., Donegan, D., & Garner, P. (2015). Deworming drugs for soil-transmitted intestinal Worms in children: Effects on nutritional indicators, haemoglobin, and school performance (review). *Cochrane Database of Systematic Reviews, 7*. https://doi.org/10.1002/14651858.CD000371.pub6

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis

VanderWeele, T. J. (2014). A unification of mediation and interaction: A four-way decomposition. *Epidemiology, 25*(5), 749–761. https://doi.org/10.1097/EDE.0000000000000121

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods, 18*(2), 137.

Waldner, D. (2007). Transforming inferences into explanations: Lessons from the study of mass extinctions. In R. N. Lebow & M. I. Lichbach (Eds.), *Theory and evidence in com-*

parative politics and international relations (pp. 145–175). Palgrave Macmillan. https://doi.org/10.1057/9780230607507_6

Waldner, D. (2012). Process tracing and causal mechanisms. In H. Kincaid (Ed.), *The Oxford handbook of philosophy of social science* (pp. 65–84). Oxford University Press.

Weller, N., & Barnes, J. (2014). *Finding pathways: Mixed-method research for studying causal mechanisms*. Cambridge University Press. https://doi.org/10.1017/CBO9781139644501

Wilcox, A. J. (2006). Invited commentary. The perils of birth weight—A lesson from directed acyclic graphs. *American Journal of Epidemiology, 164*(11), 1121–1123. https://doi.org/10.1093/aje/kwj276

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yerushalmy, J. (1971). The relationship of parents' cigarette smoking to outcome of pregnancy – Implications as to the problem of inferring causation from observed associations. *American Journal of Epidemiology, 93*(6), 443–456. https://doi.org/10.1093/oxfordjournals.aje.a121278