Chapter 23: Causal Mediation

Contents

Throughout most of this book we have asked: "What is the **total** causal effect of treatment A on outcome Y?" This question collapses all causal pathways between A and Y into a single number. But scientific and policy questions often require a finer decomposition: *how much* of the effect of A on Y operates through a particular intermediate variable (a **mediator** M), and how much operates through other pathways?

Mediation analysis has a long history in the behavioral and social sciences, but the traditional approach — based on coefficient differences in linear regression — has significant conceptual and practical limitations. This chapter examines those limitations, defends a properly formulated mediation analysis, and introduces a modern interventionist framework that resolves many of the classical difficulties.

This chapter is based on Hernán and Robins (2020, chap. 23, pp. 323–330).

Key message: The classical approach to mediation (Baron and Kenny, path analysis) conflates total, direct, and indirect effects in ways that can produce misleading conclusions. The counterfactual formulation precisely defines each effect as a potential outcome, reveals the assumptions required for identification, and suggests estimation strategies that parallel those developed for total effects.

1 23.1 Mediation Analysis Under Attack (p. 323)

Classical mediation analysis, as popularized by Baron and Kenny (1986), decomposes the total effect of A on Y into:

- A **direct effect**: the effect of A on Y not through M.
- An **indirect effect** (mediated effect): the effect of A on Y through M.

In a linear structural equation model, this decomposition is:

$$Y = \beta_0 + \beta_1 A + \beta_2 M + \varepsilon_Y, \quad M = \alpha_0 + \alpha_1 A + \varepsilon_M.$$

The indirect effect is $\alpha_1 \cdot \beta_2$ (the "product of coefficients") and the direct effect is β_1. The total effect is $\beta_1 + \alpha_1 \cdot \beta_2$.

1.1 Criticisms of the Classical Approach

Definition 1.1 (Limitations of Classical Mediation Analysis)**.** The classical approach fails in several important ways:

1. **Non-linear models**: In logistic or survival regression, the "difference of coefficients" and "product of coefficients" methods do not estimate the same quantity and neither equals the natural indirect effect.
2. **Interaction**: If A and M interact in their effects on Y (i.e., the effect of M on Y differs by A), classical mediation analysis misses this interaction entirely.
3. **Confounding of the mediator**: If there are common causes of M and Y (confounders of the mediator-outcome relationship), the regression coefficients do not have causal interpretations.
4. **Unmeasured A–M interaction for non-linearities**: Even correcting for the above, the classical estimands do not correspond to well-defined causal quantities.

These criticisms motivate a formal counterfactual definition of direct and indirect effects, which makes the identifying assumptions explicit and yields estimators that work in non-linear models with interaction.

Historical note: The critique of classical mediation using the potential outcomes framework was developed in the 2000s and 2010s by VanderWeele, Robins, Pearl, and others. The key insight is that the classical "direct" and "indirect" effects are estimable only under strong — and usually unstated — assumptions about the absence of $A \times M$ interaction and the absence of unmeasured confounders of the $M \to Y$ relationship.

2 23.2 A Defense of Mediation Analysis (p. 325)

Despite these criticisms, mediation analysis remains scientifically valuable when conducted with appropriate care. Understanding *how* a treatment works — through what mechanisms — is essential for:

- **Surrogate endpoint validation**: If a treatment works only through biomarker M, then M may serve as a valid surrogate for Y in future trials.
- **Mechanism-based intervention design**: If the indirect pathway through M is dominant, interventions targeting M directly (without A) may be effective.
- **Effect decomposition for policy**: Policy makers may want to know how much of the health disparity between two groups is "explained" by a specific mediator.

The counterfactual framework provides precise definitions that make these scientific goals achievable without the ambiguities of the classical approach.

Definition 2.1 (Natural Direct Effect (NDE))**.** The **natural direct effect** of A (comparing $a = 1$ to $a = 0$) is

$$\text{NDE} = \text{E}\left[Y^{a=1,M^{a=0}}\right] - \text{E}\left[Y^{a=0,M^{a=0}}\right],$$

where $Y^{a,m}$ is the potential outcome under treatment a and mediator value m, and $M^{a=0}$ is the potential value of the mediator under no treatment. The NDE asks: what is the effect of A on Y if the mediator were "held" at the value it would have taken under no treatment?

Definition 2.2 (Natural Indirect Effect (NIE))**.** The **natural indirect effect** of A (comparing $a = 1$ to $a = 0$) is

$$\text{NIE} = \text{E}\left[Y^{a=1,M^{a=1}}\right] - \text{E}\left[Y^{a=1,M^{a=0}}\right],$$

which captures the effect of changing the mediator from $M^{a=0}$ to $M^{a=1}$ while holding treatment at $a = 1$. The NIE asks: how much of the treatment effect is attributable to the change in M?

2.1 The Decomposition

The **total effect** decomposes as:

$$\text{E}[Y^{a=1}] - \text{E}[Y^{a=0}] = \text{NDE} + \text{NIE}.$$

This decomposition always holds on the additive (difference) scale. On other scales (ratio, odds ratio), decompositions exist but are more complex.

Nested counterfactuals: The quantity $Y^{a=1,M^{a=0}}$ is a **cross-world** counterfactual: it asks about the outcome when treatment is set to 1 but the mediator takes the value it would have taken if treatment had been set to 0. This requires simultaneously imagining two different treatment worlds, which cannot be directly observed.

Identification: The NDE and NIE are identified under four conditions: (1) no unmeasured A–Y confounders, (2) no unmeasured M–Y confounders, (3) no unmeasured A–M confounders, and (4) no M–Y confounder that is itself caused by A. Condition (4) is particularly restrictive: if treatment affects a common cause of M and Y, the cross-world counterfactuals are not identified from observed data.

3 23.3 Empirically Verifiable Mediation (p. 327)

The natural direct and indirect effects require identifying assumptions that cannot be empirically verified — in particular, the cross-world consistency condition assumes that the potential outcome $Y^{a=1,M^{a=0}}$ is well-defined and corresponds to a realizable intervention. This is conceptually problematic when the mediator M cannot in principle be manipulated independently of A.

An alternative is to focus on **controlled direct effects** and related estimands that do not require cross-world counterfactuals.

Definition 3.1 (Controlled Direct Effect (CDE))**.** The **controlled direct effect** of A at mediator level m is

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

where both A and M are set simultaneously to specified values. The CDE asks: what is the effect of A on Y when M is held at level m by external intervention?

The CDE is identified under the same conditions as the average causal effect of A, *plus* no unmeasured M–Y confounders. Crucially, the CDE does not require the cross-world consistency assumption and does not depend on condition (4) above (no A-affected M–Y confounder).

3.1 Empirical Content

The CDE corresponds to a **physical intervention**: setting $M = m$ for everyone in the population and then comparing treatment assignments. This is the kind of intervention that could, in principle, be conducted (e.g., in a sequential randomized trial that randomizes both A and M).

The limitation of the CDE is that it does not provide a single measure of the "indirect effect"; rather, the CDE at m captures the direct effect when the mediator is fixed at m, and one obtains a different CDE for each value of m.

Relationship between CDE and NDE:

When there is no $A \times M$ interaction on the additive scale (i.e., $\mathrm{E}[Y^{1,m}] - \mathrm{E}[Y^{0,m}]$ does not depend on m), the CDE equals the NDE for any value of m. In this case, both estimands agree and the choice is immaterial. When interaction is present, they differ, and the analyst must decide which is of interest.

Four-way decomposition: VanderWeele (2014) proposed a four-way decomposition of the total effect into components due to: (1) the controlled direct effect (CDE at $m = 0$), (2) a pure indirect effect (no interaction), (3) an interaction component attributable to the mediator, (4) an interaction component attributable to treatment. This decomposition is especially useful when $A \times M$ interaction is present and of scientific interest.

4 23.4 An Interventionist Theory of Mediation (p. 329)

The deepest conceptual challenge in mediation analysis arises from the cross-world nature of the NDE and NIE: $Y^{a=1,M^{a=0}}$ requires simultaneously setting $A = 1$ (at one level) and letting M be as it would be if $A = 0$ (a different level). This is not the result of any single intervention on (A, M).

The **interventionist** approach resolves this by restricting attention to estimands that correspond to interventions that *could actually be performed*.

> **Definition 4.1** (Interventionist (Stochastic) Mediation Estimands)**.** Let Q be a distribution over mediator values. Define the **randomized interventional analogue** of the indirect effect as
>
> $$\mathrm{rIIE} = \mathrm{E}\big[Y^{a=1,G_1}\big] - \mathrm{E}\big[Y^{a=1,G_0}\big],$$
>
> where $G_a \sim Q(M^a)$ denotes assigning the mediator at random from the distribution it would have under treatment a.

In this formulation, both components of the nested counterfactual correspond to actual interventions: set A to some value and draw M from a distribution. There is no cross-world problem because we are only asking "what would happen if A were set to 1 and M were drawn from the distribution it would have under $A = 0$?" — which is an intervention that could be implemented.

4.1 Identification of Interventionist Effects

The randomized interventional analogue of the indirect effect is identified under conditions that do not require assumption (4) from Section 23.2:

$$\mathrm{rIIE} = \mathrm{E}\left[\int_m \mathrm{E}[Y \mid A = 1, M = m, C]\, dF_{M|A=0,C}(m)\right] - \mathrm{E}[\mathrm{E}[Y \mid A = 1, M, C]],$$

where C are baseline covariates sufficient to control confounding of both the $A \to Y$ and $M \to Y$ pathways. This expression can be estimated from observed data using outcome regression, IP weighting, or doubly robust methods.

Why the interventionist approach matters:

The standard NDE requires cross-world consistency, which is often unjustifiable when the mediator cannot be independently manipulated. The interventionist approach by Vansteelandt and Daniel (2017) and the stochastic mediation approach by Didelez, Dawid, and Geneletti (2006) replace cross-world counterfactuals with stochastic interventions on the mediator, which are well-defined physically and identifiable under weaker assumptions.

Example: Suppose A is a job training program, M is employment status, and Y is income. The NDE requires holding employment at the value it *would have been* without training — but if training causes employment, what does it mean to "hold" employment at the no-training value?

The interventionist approach instead asks: "what is the effect on income if we randomly assign employment from the distribution it would have in the no-training group?" This is a well-defined intervention (randomly assign individuals to jobs or not) and is identifiable.

4.2 The Do-Operator and Mediation

In the structural causal model (SCM) framework of Pearl (2000), mediation is analyzed using the **do-operator** $\text{do}(M = m)$, which represents an intervention that sets M to a fixed value m by removing all arrows into M in the causal DAG.

Definition 4.2 (Direct Effect via the Do-Operator)**.** The **controlled direct effect** in Pearl's framework is:

$$\text{CDE}(m) = \text{E}[Y \mid \text{do}(A = 1), \text{do}(M = m)] - \text{E}[Y \mid \text{do}(A = 0), \text{do}(M = m)].$$

This is equivalent to the potential outcome expression $\text{E}\big[Y^{a=1,m}\big] - \text{E}\big[Y^{a=0,m}\big]$ under the consistency assumption.

The natural direct and indirect effects in SCMs require an additional assumption — that the system obeys a **composition rule** — which corresponds to the cross-world consistency condition above. The interventionist approach avoids this by replacing $\text{do}(M = M^{a=0})$ with $\text{do}(M \sim F_{M^{a=0}})$, a stochastic do-intervention.

i Fine Point 23.1: Mediation with Time-Varying Mediators

When the mediator M is itself time-varying — for example, when A is a baseline treatment and M_k is a biomarker measured at each follow-up visit — mediation analysis becomes a special case of the time-varying treatment framework from Chapters 19–21.
The sequence $A \to M_0 \to M_1 \to \cdots \to M_K \to Y$, with treatment feeding back into subsequent mediator values, creates precisely the treatment-confounder feedback structure analyzed in Chapters 20 and 21. G-methods are therefore the appropriate tools for mediation analysis with time-varying mediators.
In this setting, the controlled direct effect is defined as the effect of A under an intervention that sets all $M_k = 0$ (or to some reference value) for the entire follow-up period. The indirect effect is the difference between the total effect and this controlled direct effect.

5 Summary

- **Classical mediation analysis** (Baron and Kenny) has serious limitations: it requires linearity and no $A \times M$ interaction, fails to identify confounders of the $M \to Y$ pathway, and does not extend to non-linear models.
- The **natural direct effect** (NDE) and **natural indirect effect** (NIE) are defined via nested cross-world counterfactuals:

$$\text{NDE} = \text{E}\big[Y^{a=1,M^{a=0}}\big] - \text{E}\big[Y^{a=0,M^{a=0}}\big],$$

$$\text{NIE} = \text{E}\big[Y^{a=1,M^{a=1}}\big] - \text{E}\big[Y^{a=1,M^{a=0}}\big].$$

- The NDE and NIE sum to the total effect and allow decomposition in non-linear models, but require cross-world consistency and the absence of A-affected M–Y confounders.
- The **controlled direct effect** (CDE) avoids cross-world counterfactuals and corresponds to an intervention that fixes M at a specific value:

$$\text{CDE}(m) = \text{E}\big[Y^{a=1,m}\big] - \text{E}\big[Y^{a=0,m}\big].$$

- The **interventionist** approach further resolves the cross-world problem by replacing the NDE/NIE with stochastic interventional analogues that correspond to realizable experiments.

- With **time-varying mediators**, mediation analysis becomes a special case of the g-methods framework for time-varying treatments.

6 References

Hernán, Miguel A, and James M Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC. https://miguelhernan.org/whatifbook.