

A PLAIN-LANGUAGE GUIDE

Causal Reasoning and Inference in Epidemiology

How epidemiologists move from “what is associated with what” to “what would happen if we intervened” — and what it takes to trust the answer.

Based on: Vanessa Didelez, “Causal Reasoning and Inference in Epidemiology,” Handbook of Epidemiology (Springer, 2025)

SETTING THE STAGE

From Describing the World to Changing It

A core aim of epidemiology is finding out how to protect or improve health — through behavior change, new treatments, or policy. That means asking not just how things currently are, but how they would be under a different choice: a “causal” question. Because true experiments are often impossible, most of this evidence has to come from observational data — cohort studies, health records, routine registries.

1. AN EXPLICIT TARGET

Say precisely what hypothetical comparison you actually mean.

2. STATED ASSUMPTIONS

Spell out what must be true for the data to answer that question.

3. MATCHING METHODS

Specialized statistics come last — and don't help without 1 and 2.

Without an explicit causal target and a clear statement of assumptions, no amount of statistical sophistication makes an analysis convincing.

THE SHAPE OF A CAUSAL ANALYSIS

Four Stages, From Question to Interpretation

Causal model: story & DAG

The research question, a target trial design, and the assumptions needed for identifiability



adapt — defensible?

Available data

Cohort studies, health records, registries — whatever evidence actually exists



Emulation & statistical analysis

Choose an estimator (weighting, g-formula, doubly robust) and run sensitivity checks: balance, positivity, negative controls



Interpretation

Statistical findings, what they imply for the causal question, and a critical look at the data and assumptions behind them

WHERE THIS DECK IS GOING

Eight Ideas, Building on Each Other

1 • Questions

Turning a vague question into a precise, hypothetical trial

2 • Models

Causal DAGs and the two main sources of structural bias

3 • Assumptions

What must hold for a causal estimate to be trustworthy

4 • Estimation

Core methods, with and without full adjustment for confounding

5 • Target trials

Designing an observational analysis to emulate an ideal trial

6 • Mediation

Direct and indirect effects — harder to define than they seem

7 • Discovery

Letting data help build the causal model itself

8 • Bottom line

From Hill's criteria to today's formal framework

FORMULATING CAUSAL RESEARCH QUESTIONS

Imagine the Trial You Wish You Could Run

Randomized controlled trials give causally interpretable results because of randomization — but also because of everything else about how they are designed: a clear population, clearly defined treatments, and a clear plan for analysis. Even when a real trial is impossible for ethical, financial, or practical reasons, it helps enormously to specify this hypothetical **target trial** before touching any data.

Hernán and Robins describe the spirit of causal inference as addressing “What if...?” questions — not just what happened, but what would have happened under a different, possibly hypothetical, intervention.

The thought process of specifying a target trial borrows the discipline of an RCT, even when no randomization is actually possible.

WHAT A TARGET TRIAL HAS TO SPECIFY

Six Questions Every Target Trial Must Answer

- 1 **Eligibility** — who is the population of interest?
- 2 **Treatment strategies** — exactly which options or interventions are being compared?
- 3 **Assignment** — how would people be assigned to a strategy?
- 4 **Follow-up** — for how long, starting when?
- 5 **Outcome** — defined precisely enough to apply in every circumstance
- 6 **Causal contrast & analysis plan** — the comparison being made, and how it will be estimated

Working through all six for an observational study borrows the strengths of an RCT — everything except the randomization itself.

A DETAIL RCTS GET FOR FREE

Alignment at “Time-Zero”

In a randomized trial, it is always unambiguous which measurements came before randomization and which came after. An observational analysis has to construct that same clean starting line by making sure three things line up at the same moment for every person:

i. Eligibility

Checked using only what was known up to this point



ii. Assignment

Which strategy a person’s data will represent



iii. Follow-up starts

The clock for the outcome begins

When these three drift apart — for instance, using information from after the clock has started to decide who counts as “exposed” — the analysis can look into the future by accident, a common and avoidable source of bias. A further catch: some exposure categories, like “former smoker,” are not well-defined interventions at all — you cannot randomly assign someone’s past, only a prospectively-defined strategy, such as starting to smoke in adolescence and then sustaining versus discontinuing it.

Every Target Trial Needs a “Causal Story”

Many exposures of interest in epidemiology — biomarkers, body mass index, birth weight, sleep duration — cannot themselves be directly randomized in a real trial. The causal story is the narrative that connects such an exposure to some hypothetical, and sometimes indirect, intervention that could plausibly change it.

A MODEST STORY

Sleep duration cannot be “set” directly, but a bedtime-routine program aimed at increasing it by 10% is a coherent hypothetical intervention to target instead.

A MORE STRAINED ONE

To explain why cancer and dementia are inversely associated, researchers have proposed a target trial with a hypothetical drug that raises a specific brain protein — even though no such drug yet exists.

The further the causal story strays from anything actually measured, the harder it becomes to map the question onto real data — but making that gap explicit is itself valuable.

TWO KINDS OF DEPENDENCE

“Is Linked To” Is Not “Causes”

Association

Describes how two things are observed to vary together. It is symmetric: if treatment is predictive of the outcome, the outcome is equally predictive of the treatment.

Causation

Describes what would happen under a hypothetical **intervention** that fixes the treatment from outside the system. It is directional, and need not match the association at all.

Formally, epidemiologists write the intervention as a “do-operator”: comparing the outcome as observed, given a treatment value, against the outcome under an *imposed* intervention setting that same value. “Association does not imply causation” just means these two are, in general, different quantities.

THE BUILDING BLOCK OF CAUSAL EFFECTS

Potential Outcomes: What Would Have Happened

For each person, imagine two parallel possibilities: the outcome that would occur if they were treated, and the outcome that would occur if they were not. Epidemiologists write these as $Y(1)$ and $Y(0)$. Only one of the two is ever actually observed for any given person — the other stays forever hypothetical, or “counterfactual.”

$Y(1)$

Outcome if this person had been treated

$Y(0)$

Outcome if this person had not been treated

Under an assumption called **consistency** (covered shortly), the outcome we actually observe equals whichever potential outcome matches the treatment the person actually received.

Because both outcomes can never be seen for the same person at the same time, no data or study design — not even randomization — can reveal an individual's own causal effect without further, fundamentally untestable assumptions.

FROM ONE PERSON TO A POPULATION

Four Ways to Average a Causal Effect

Individual (ICE) $Y(1)$ minus $Y(0)$ for one person — never observable, since we never see both.

Conditional (CATE) The average effect within a defined subgroup, e.g. women aged 40–45 with low income.

Population (ACE) The average effect if everyone eligible were treated versus if none were — a weighted average of every CATE.

On the treated (ATT) The average effect just among those who would take the treatment anyway — the relevant number for many policy questions.

When the CATE differs across subgroups, that difference is called **effect modification** — the effect is not one number, but depends on who you are.

A LARGER FAMILY OF CAUSAL EFFECTS

Beyond a Single Treatment, a Single Number

Sequential & direct effects

With two treatments in sequence, we can ask about their joint effect, or the effect of one *holding the other fixed* — a controlled direct effect (more in the mediation section).

Causal interactions

Occurs when the effect of intervening on one treatment genuinely depends on what the other is set to — not just conditioning on a subgroup.

Attributable fractions

The share of disease in a population that a given intervention could prevent — e.g. an estimated 17% of childhood overweight attributable to not following healthy-behavior guidelines.

Hazard ratios, with care

Popular for time-to-event outcomes, but a population hazard ratio is not simply a weighted average of subgroup hazard ratios. Part of the problem is post-treatment selection: those still surviving at later times become a systematically different, healthier group under one treatment than under the other.

A further subtlety, principal stratification, restricts a causal effect to the subgroup who would survive or comply under either treatment — useful, but such subgroups are not fully identifiable from data.

CAUSAL MODELS AND ASSUMPTIONS

Why We Need an Explicit Causal Model

Causal inference is normally carried out inside a causal (or structural) model — a model for how variables and processes behave, both under intervention and left alone. The one notable exception is a randomized trial, where randomization alone is usually treated as enough to license a causal reading. But even a trial rests on an implicit causal model: the background reasoning that justified running it, and shaped its design, in the first place.

IN OBSERVATIONAL STUDIES

A causal model makes the current state of knowledge — and the assumptions behind a chosen analysis — explicit enough to scrutinize, question, or defend.

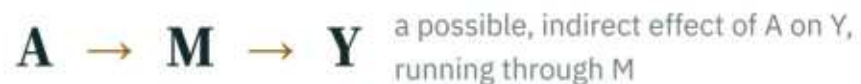
WHAT IT BUYS US

A link between the observable data and the assumed process that generated it — and a way to identify or rule out specific sources of bias.

A SHARED VISUAL LANGUAGE

Causal DAGs, in a Nutshell

A directed acyclic graph (DAG) draws variables as circles and possible direct effects as arrows. Hernán put it simply: “Draw your assumptions before you draw your conclusions.”



The twist: it is the **absence** of an arrow that carries the real weight. Drawing no arrow between two variables is a claim of near-certainty that no direct effect exists — a much stronger statement than drawing one in.

READING A DAG

Paths Can Be Open, or They Can Be Blocked

Any chain of arrows connecting two variables — not only a directed path — can be “open” (it induces an association between them) or “blocked” (it induces none). Whether a path is open or blocked can flip depending on which other variables are held fixed by conditioning, such as through stratification or statistical adjustment. The formal rule for checking this is known as d-separation; two patterns behind it explain most sources of bias in practice.

A shared-cause path

Open by default; blocked once you condition on the shared ancestor. This is the pattern behind confounding.

A shared-effect path

Blocked by default; opened if you condition on the shared effect (or something it causes). This is the pattern behind selection bias.

A related tool, single-world intervention graphs (SWIGs), makes potential outcomes explicit on the graph itself, without needing any of the untestable “cross-world” assumptions covered later.

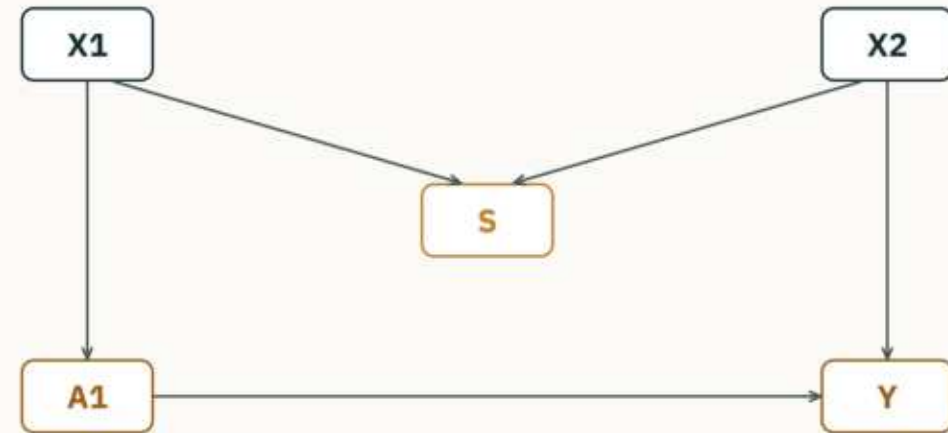
TWO SMALL DAGS, TWO DIFFERENT PROBLEMS

Same Four Variables, Opposite Fixes



Confounding

X1 opens a backdoor path from A1 to Y. Adjusting for either X1 or X2 blocks it and removes the confounding.



Selection (M-bias)

X1 and X2 are unrelated, and S has no effect on A1 or Y. But selecting on S — say, by design — opens a spurious path between A1 and Y out of nothing.

TWO WAYS AN ESTIMATE CAN MISLEAD

What “Structural Bias” Means

Bias is a **systematic** difference between the target you intend to estimate and what your analysis tends to produce on average. That is different from ordinary random variation due to sample size, which shrinks as data accumulates. Structural bias, by contrast, comes from problems with the study design, the database, or the analysis plan itself — and no amount of extra data fixes it.

Confounding

A common cause of both the treatment and the outcome distorts the comparison.

Selection bias

Conditioning — by design or by analysis choice — on the wrong kind of variable distorts it instead.

These two are the main structural biases specific to causal analysis; a further, trickier version — treatment-confounding feedback — arises specifically with treatments that unfold over time.

Confounding: A Shared Cause in Disguise

Confounding of a treatment's effect on an outcome occurs whenever some common cause links the two — whoever got the treatment already differed, in ways connected to the outcome, from whoever didn't. Randomization is powerful precisely because it severs every incoming arrow into the treatment node, balancing treated and untreated groups on every baseline characteristic, measured or not.

A subtlety worth keeping straight: this chapter defines **confounding** — a property of the whole causal structure — not “a confounder.” Deciding whether one particular variable counts as a confounder is not as simple as the usual rules of thumb suggest; it depends on the full causal model, including the other variables already being adjusted for.

THE LESS FAMILIAR CULPRIT

Selection Bias: When Who's Included Matters

Selection happens whenever the design of a study, or a choice made during analysis, conditions on certain characteristics — on purpose or by accident. That is not automatically a problem, but it always needs scrutiny. Case-control studies and so-called test-negative designs are built entirely around a deliberate selection step, which is precisely why they require such careful handling.

The usual way selection bias slips in is by conditioning on a **collider** — a variable that is itself a common effect of two other variables (or a descendant of one) — which opens a path between them that was otherwise safely blocked.

One well-known shape this takes is “M-bias”: an otherwise unremarkable variable used to define eligibility for a study — such as excluding people already diagnosed by the time screening begins — can quietly act as exactly this kind of collider.

REAL EXAMPLE · ADJUSTING FOR THE WRONG THING

Why Smoking Looked “Protective” Against COVID-19 Death

The analysis

Studying risk factors for COVID-19 death, some analyses adjusted for respiratory disease — and smoking came out looking protective.

The likely explanation

Smokers who don’t develop measured respiratory disease may still carry other unmeasured lung problems. Adjusting for respiratory disease — itself a consequence of smoking — opened a spurious path through those unmeasured factors.

A close cousin is the “birth weight paradox”: low-birth-weight infants of smokers can appear to have *better* survival odds than those of nonsmokers, once birth weight — itself affected by smoking — is conditioned on. In both cases, the mistake is the same: adjusting for a consequence of the exposure, not a cause of it.

WHEN TREATMENTS UNFOLD OVER TIME

Treatment-Confounding Feedback

With two treatments given in sequence, a covariate measured in between can play two roles at once: it confounds the second treatment's effect on the outcome, while itself being a consequence of the first treatment.

A classic example: in early HIV research, doctors decided when to start antiretroviral treatment partly based on a patient's CD4-count, monitored over time. CD4-count is a confounder of the later treatment decision — but it is also affected by the earlier treatment, making it a mediator too.

A single regression adjusting for this kind of variable gets tangled: it can block part of the real effect while opening a spurious path elsewhere. The good news is that the joint effect of the full treatment sequence can still be identified — just not by an ordinary regression, and only with methods built for exactly this situation (covered in the estimation section).

STRUCTURAL ASSUMPTIONS

What Has to Be True to Trust an Estimate

Structural assumptions are what make identifiability possible — they ensure that, aside from sample-size issues, the observable data are in principle enough to pin down the causal quantity. There are two broad routes to get there.

ROUTE ONE

Assume the data contain enough information to fully adjust for confounding.

ROUTE TWO

Acknowledge that confounding can't be fully measured, and lean on a natural experiment or extra modeling assumptions instead.

The next slides walk through consistency, exchangeability, the backdoor criterion, positivity, and no interference — the standard assumptions behind Route One — before turning to instrumental variables as an example of Route Two.

ASSUMPTION ONE

(Causal) Consistency

Consistency links the observational data to the hypothetical intervention: when a person is observed to receive treatment a_1 , their observed outcome must equal the potential outcome under that same value, $Y(a_1)$. In words — the treatment someone actually got has to mean the same thing as the treatment an intervention would have imposed.

Where this gets shaky: for an exposure like physical activity, it is implausible that everyone who happens to have low activity in the data would have had exactly the same low level in the total absence of any intervention program. Real-world “versions” of a treatment can vary in ways a single hypothetical intervention doesn’t capture.

Being explicit about exactly what intervention is intended — and how close the observed “versions” of treatment come to it — is the best defense available, since consistency itself cannot be tested directly.

ASSUMPTION TWO

Conditional Exchangeability

A set of pretreatment covariates X is **sufficient for adjustment** if, within every stratum defined by X , the treated and untreated are “exchangeable” — as good as randomly assigned. Treated and untreated subjects who share the same X values carry the same information about what would have happened to either group.

This same idea travels under several names in the literature: **no unmeasured confounding**, **conditional ignorability**, **unconfoundedness**, or “selection on observables” — all presupposing that X has actually been measured.

When X is empty, this reduces to unconditional exchangeability — exactly what randomization is designed to guarantee.

A GRAPHICAL SHORTCUT

The Backdoor Criterion

Rather than reasoning through exchangeability in the abstract, a causal DAG gives a mechanical rule for finding a sufficient adjustment set: choose covariates that are not affected by the treatment, and that together block every backdoor path — every path that starts with an arrow pointing into the treatment.

Not always unique

In the earlier confounding example, adjusting for either X1 or X2 alone is enough — both block the one open backdoor path.

Not always efficient

Where more than one valid set exists, they can differ in how precisely the effect gets estimated — some choices are simply better than others.

One other graphical route: the frontdoor criterion can identify an effect via an unconfounded mediator even when backdoor paths are blocked by unmeasured confounding — applied, for instance, to mobile stroke units.

The same logic extends to joint interventions on several treatments at once, and — in its full generality — can identify effects that are not captured by simply repeating the single-treatment version at each time point.

ASSUMPTION THREE

Positivity: Everyone Needs a Real Chance

Consistency and exchangeability are not quite enough. For every combination of adjustment-set values that occurs in the population, there also has to be a genuine possibility — a positive probability — of receiving each treatment level being compared. Without that overlap, part of the comparison is simply never observed.

In pharmacoepidemiology, some patients may never receive the more aggressive of two treatments, while others may never receive the milder one. For those groups, the comparison is not just empirically unavailable — it may not be medically meaningful either.

The same requirement extends to sequential treatments: every treatment value needs positive probability within every stratum of the covariates measured up to that point.

ASSUMPTION FOUR

No Interference Between People

A quiet, often unstated assumption: one person's treatment does not affect another person's outcome. It is easy to overlook because it is usually reasonable — but not always.

Infectious disease

Vaccinating some individuals can protect others around them — their grandchildren, for instance — who were never vaccinated themselves.

Social networks

Studying how peers shape adolescent smoking or drinking, it can matter whether people chose their friends because of shared attitudes, or grew to share them afterward.

For the remainder of this chapter's methods, no interference is simply assumed — but its violation is a genuine, active area of causal-inference research in its own right.

ROUTE TWO • A NATURAL EXPERIMENT

Instrumental Variables: Borrowing a Randomizer

When no set of measured covariates can plausibly rule out unmeasured confounding, an instrumental variable (IV) offers a different way in: a variable Z that behaves like an imperfect, naturally occurring randomization device.

Associated with treatment Z actually shifts who gets treated

No direct effect Z affects the outcome only by way of the treatment

Not itself confounded Z shares no hidden common cause with the outcome

Mendelian randomization uses a genetic variant as Z , since genes are fixed at conception, before any confounding could arise. In pharmacoepidemiology, a physician's personal prescribing preference has served the same purpose. Even without pinning down the exact effect size, a valid IV lets you test whether any causal effect exists at all.

WHAT CAN ACTUALLY BE VERIFIED

Most of This Cannot Be Tested — One Piece Can

Not directly testable

Consistency and exchangeability both involve potential outcomes — quantities that, by definition, are never all observed together. They can sometimes be falsified by an implied empirical pattern, but never confirmed outright, and a clean test result is not proof that the assumption holds.

Directly checkable: positivity

Because positivity concerns only the observed distribution of treatment and covariates, it can be inspected: histograms of covariates by treatment group, propensity-score overlap plots, and “Love plots” comparing covariate balance before and after adjustment.

These same diagnostic plots also give a basic impression of how much confounding adjustment is achieving — even though they can never confirm that confounding has been fully removed.

Three Ways to Put Exchangeability to Work

Once conditional exchangeability, consistency, and positivity hold, there is more than one way to actually produce a number. All three remove the same dependence between treatment and covariates — they just remove it from a different place.

Match

Pair similar people directly, balancing covariates by construction.

Model the treatment

Estimate each person's chance of treatment, and use it to remove the covariate imbalance.

Model the outcome

Regress the outcome on treatment and covariates, then average over the covariate distribution.

THE MOST DIRECT APPROACH

Matching: Pairing Like With Like

Matching pairs each treated person with one or more untreated people who share very similar covariate values, directly constructing independence between treatment and the adjustment set within the matched sample. It is attractive because it stays close to a nonparametric ideal — almost no modeling assumptions about how covariates relate to treatment or outcome.

The catch: matching becomes genuinely difficult once covariates are continuous or numerous, forcing a trade-off between how closely people are matched and how much data survives the process. It also faces particular challenges once treatments unfold over time.

This difficulty is exactly what motivates the model-based alternatives on the following slides.

ONE NUMBER STANDING IN FOR MANY COVARIATES

The Propensity Score

The propensity score is simply each person's estimated probability of receiving treatment, given their covariates. Its remarkable property is that it is a **balancing score**: among people who share the same propensity score, the full covariate set carries no further information about who was treated — however many covariates went into calculating it.

In practice the propensity score is usually estimated with a logistic regression, though more flexible models are increasingly used. Its purpose is not to predict treatment as accurately as possible — it is to achieve good **balance** between the groups once it is used for matching or weighting.

Including unnecessary predictors in the propensity-score model — especially instruments — can actually make bias worse, a phenomenon called bias amplification.

THREE WAYS TO SPEND A PROPENSITY SCORE

Matching, Stratifying, or Weighting

PS matching

Pair people with similar scores, then analyze the matched sample almost as if it were a randomized trial.

PS stratification

Group people into bands of similar scores and compare treated versus untreated within each band.

IPTW

Weight each treated person by $1 \div \text{score}$, each untreated person by $1 \div (1 - \text{score})$, building a pseudo-population where covariates no longer predict treatment.

Propensity-score methods are popular partly for their simplicity and transparency: once treated and untreated are balanced, familiar tools — standard errors, survival curve estimators — can often be reused with only small adjustments.

DID THE ADJUSTMENT ACTUALLY WORK?

Diagnostics: Overlap and Balance

Because these diagnostics use only treatment and covariates — never the outcome — they can be examined freely, without any risk of tailoring the analysis to get a preferred answer.

Overlap

Plot the propensity score separately for treated and untreated. A positivity violation shows up as scores piling up near zero or one in one group but not the other.

Balance (a “Love plot”)

Compare the standardized difference in each covariate between groups, before and after adjustment — the after-values should sit much closer to zero.

A crucial limit: these plots can only speak to balance on *measured* covariates. They cannot confirm the untestable assumption that no unmeasured confounder remains.

EXTENDING IPTW TO SEQUENCES

Marginal Structural Models for Sequential Treatments

To model the joint effect of two treatments given over time — despite treatment-confounding feedback — a marginal structural model (MSM) specifies the effect of the treatment sequence directly, then fits it to the data reweighted by the inverse of each person's probability of following the treatment pattern they actually followed, given their history.

The recurring practical problem: weights can become extreme whenever the data come close to violating positivity, producing unstable estimates with very wide uncertainty. A common fix is to use **stabilized weights**, which additionally reweight by the overall treatment distribution in the population.

As with the single-treatment case, covariate balance in the reweighted data can and should be checked directly — using the same style of plot, adapted to check balance at every treatment time.

A DIFFERENT ROUTE TO THE SAME EFFECTS

The Parametric G-Formula: Model, Then Simulate

Rather than a single marginal model, the g-formula approach models every ingredient of the process directly — how each time-varying covariate depends on the treatment and covariate history so far, and how the outcome depends on the full history. Fitting all of these to real data, then simulating forward under a specified hypothetical strategy — “always treat,” “never treat,” or something more elaborate — produces a predicted potential outcome for that strategy.

A useful built-in check: simulating the “natural course” — no intervention at all — should approximately reproduce the trajectory actually observed in the data, which is a quick way to assess model fit.

The approach adapts naturally to complex, dynamic, or even probabilistic hypothetical strategies — such as intervening only when a measurement falls below some threshold — which is difficult to express any other way.

TWO CHANCES TO GET IT RIGHT

Doubly Robust Estimation

IPTW leans entirely on the treatment model; the g-formula leans entirely on the outcome model. A doubly robust estimator combines both — and needs only **one** of the two to be correctly specified in order to remain valid, not both at once.

Machine-learning models like random forests can, in principle, fit either piece flexibly without pre-specifying its exact form. Simply plugging naive machine-learning predictions into a causal-inference formula tends to perform poorly, because the regularization that keeps machine learning from overfitting introduces its own bias. **Double machine learning** corrects for this directly, and delivers on the promise of flexible, robust estimation.

A related strategy, structural nested models fitted by g-estimation, targets the treatment effect directly under a different set of modeling assumptions — another route toward the same double-robustness goal.

ESTIMATING WITHOUT FULL ADJUSTMENT

The Wald Estimator, and Other Natural Experiments

With a valid instrument Z , a simple and popular estimate of the treatment effect is the ratio of two ordinary regression coefficients: the association of the outcome with Z , divided by the association of the treatment with Z . This “Wald ratio” can also be used when outcome and treatment data are not even available on the same people.

Sometimes the data identify a range, not a number. The core instrumental-variable assumptions alone need not pin down one exact effect — they can instead bound it within a range. These identification bounds are a different thing from a confidence interval; turning the range into the single Wald-ratio point estimate takes additional structural assumptions.

Difference-in-differences

Compares how outcomes changed over time in a group exposed to a new policy versus a group that was not; synthetic control builds a similar comparison from a blend of untreated units.

Regression discontinuity

Exploits a sharp eligibility cutoff, comparing people just above and just below it.

Interrupted time series

Looks for a change in trend right at the moment an exposure or policy began.

All of these substitute a design-based “common trend” type assumption for full covariate adjustment — powerful when it holds, but often quite specific to the particular setting studied.

WHEN YOU STILL SUSPECT RESIDUAL CONFOUNDING

Quantitative Bias Analysis

Even after careful adjustment, doubt can remain about confounders that were never measured. Quantitative bias analysis uses explicit, domain-knowledge assumptions about that unmeasured confounding to correct or bound an estimate, rather than ignoring the doubt altogether.

Negative controls

An outcome (or exposure) that the real exposure could not plausibly cause, but that would still be affected by the same hidden confounders — a way to detect their fingerprints indirectly.

The e-value

Quantifies how strong an unmeasured confounder would have to be — on both the treatment and the outcome — to fully explain away the observed association.

These tools do not remove the doubt — they make it precise enough to reason about, alongside related strategies like triangulating across methods with different bias structures, or proximal inference, which uses proxies for the hidden confounder itself.

A QUESTION WORTH ASKING OF ANY REGRESSION TABLE

Does an Ordinary Regression Give a Causal Effect?

One treatment: often, yes

If the model is correctly specified, the adjustment set is sufficient, positivity holds, and no covariate in the model was itself affected by the treatment, the treatment's coefficient does describe the conditional causal effect. (Odds ratios and hazard ratios still need extra care — they don't average the same way across subgroups.)

Sequential treatments: usually not

One regression model cannot correctly represent the joint effect of two sequential treatments once treatment-confounding feedback is present. The variable needed to isolate one treatment's effect must be excluded to isolate the other's — a single model can't do both, sometimes called the "Table 2 fallacy."

Multivariable regression still matters here — it supplies the key ingredients for IPTW and the g-formula — but on its own, and left unexamined, it is not automatically a causal answer.

TARGET TRIAL EMULATION

Designing the Analysis to Match the Trial

Specifying a target trial is only half the job. The next step is deciding how each of its components — eligibility, treatment strategies, alignment, outcome, contrast, analysis plan — will actually be implemented using the data on hand. Confounding gets most of the attention in causal inference, but many *other* sources of bias are, by contrast, avoidable simply through careful design.

The rest of this section covers techniques for avoiding these largely **self-inflicted** biases — ones that come from how the emulation is built, not from any residual confounding.

DESIGN IT SO THE COMPARISON CAN EXIST

Eligibility and Positivity, in Practice

Positivity is not just a statistical footnote once a target trial meets real data — it should actively shape how eligibility is defined. The population should be one where being assigned to any of the trial arms is genuinely plausible, and the available database should offer empirical support for that.

It would not make sense to define a trial arm of “five years of continual drug use” if, in the actual data, nobody ever used the drug for more than three. Descriptive analysis of real drug-usage patterns over time should come first, and the trial arms adapted to match — exactly the approach taken in a study of second-line treatment for type 2 diabetes, where the arms were adjusted to fit the usage patterns actually observed.

Checking this before analysis begins avoids building a comparison the data were never going to be able to support.

REAL EXAMPLE · SCREENING COLONOSCOPY

When Time-Zero Has No Obvious Home

Some treatments have a natural time-zero for every person — the day of diagnosis, say. Others don't: screening colonoscopy can be attended, or not, at any point from around age 55 onward.

The tempting, wrong approach

Label as “exposed” anyone who ever attends screening across their lifetime, and “unexposed” everyone who never does. This quietly looks into the future to decide the past — classic immortal time bias.

The fix

Fix time-zero at the point eligibility is first met, or design a **sequence of emulated trials** with many valid starting points, pooled together and analyzed with bootstrap-based standard errors.

Studies of screening colonoscopy have shown just how much bias an unaligned design can introduce — and how much of it is avoidable by construction, before any statistical adjustment even begins.

A TECHNIQUE FOR SUSTAINED STRATEGIES

Clone, Censor, Weight

Comparing strategies like “never screen” versus “screen every year” runs into the same look-into-the-future problem: you cannot know today whether someone will keep complying with a strategy for years to come.

- | | |
|---------------|--|
| Clone | Assign each person's data to every strategy they are compatible with so far, instead of guessing one. |
| Censor | The moment a clone's real behavior departs from a strategy, stop counting it toward that strategy from then on. |
| Weight | Reweight the remaining, still-compliant clones so this artificial censoring doesn't reintroduce selection bias of its own. |

A simpler alternative avoids duplicating people across arms by randomly assigning each to just one strategy they currently comply with — at the cost of using the available data somewhat less fully.

REAL EXAMPLE · DIETARY GOALS AND 20-YEAR MORTALITY

Emulating a Trial of Sustained Healthy Eating

Chiu and colleagues (2021) estimated the 20-year risk of all-cause mortality under sustained adherence to several American Heart Association food-based goals — minimum servings of fruit and vegetables, a maximum of processed meat, and more — using three separate observational cohorts, evaluated separately and jointly across 14 treatment strategies.

Because each strategy only intervenes on someone whose diet is naturally falling short of the goal, and because earlier diet affects later confounders, the analysis needed the parametric g-formula rather than a simpler regression. The target trial framework helped keep the three cohort analyses as comparable as design allowed, even though their populations genuinely differed.

The result: full, sustained adherence to all six dietary goals considerably reduced estimated mortality risk, especially in the cohorts of older adults.

SOME CAVEATS

Emulating a Trial Is Not Replicating One

Trial **replication** means trying to reproduce the result of an existing RCT with observational data — typically restricted to the RCT’s own kind of population, and to an “intention-to-treat” style comparison, since randomization there only guarantees unconfounded inference for that specific contrast.

Target trial emulation’s real strength is different: it is often the only available option for decision-relevant questions where no RCT exists or ever will — early COVID-19 vaccine effectiveness, or drug safety during pregnancy, are both examples where this was the only path to an answer in time to matter.

Even when the available data fall short of a fully convincing emulation, going through the exercise still pays off: it surfaces exactly where the bias risks are, and points to what future data collection should prioritize.

WHERE TARGET TRIAL EMULATION ACTUALLY HELPS

Design-Related Biases, Compared

Potential limitation	RCT	Observational, with TTE	Observational, without TTE
Prevalent-user bias	Low — randomization marks the start	Low, if aligned at time-zero	High risk
Immortal time bias	Low — assignment is randomized	Low, if aligned at time-zero	High risk
Unclear research question	Low — arms are defined by design	Low — the trial made it explicit	High, if left undescribed

Baseline confounding, time-dependent confounding, and external validity remain real risks in observational data whether or not a target trial is emulated — emulation mainly protects against the design-related biases shown here.

A Simple Question, a Surprisingly Deep One

It comes naturally to talk about “direct” and “indirect” pathways — does diet affect health only through the gut microbiome, or does it matter in other ways too? The formal treatment of this intuitive question turns out to be one of the most genuinely complicated corners of causal inference.

The most concrete starting point is the **controlled direct effect (CDE)**: the effect of treatment A1 on outcome Y that remains after a hypothetical intervention holds a suspected mediator A2 fixed at a chosen value — a special case of the joint, sequential-treatment effects covered earlier.

THE COLLIDER RISK, APPLIED TO MEDIATORS

Adjusting for a Mediator Can Manufacture an Effect

It seems natural to just add the suspected mediator to a regression model and see what remains of the treatment's coefficient. This can badly mislead.

Suppose a treatment truly has **no effect at all** — direct or indirect — on an outcome. If some unmeasured factor happens to influence both the suspected mediator and the outcome, simply adjusting for that mediator opens a spurious path between treatment and outcome through the collider it creates. The analysis can manufacture the appearance of a direct effect that was never there.

A valid controlled direct effect requires the same careful exchangeability assumptions for the mediator's relationship with the outcome as any sequential treatment does — adjusting for a mediator is never a free move.

THE TRADITIONAL APPROACH

Structural Equation Models and Path Analysis

Long before the potential-outcomes language above, mediation was — and often still is — analyzed by fitting a structural equation model: one regression equation per variable in the causal chain. In a linear model, the indirect effect through a mediator is simply the product of the two path coefficients along the way, and the direct and indirect pieces add up neatly to the total.

Where it works well

Easy to fit, easy to explain, and exact in simple linear settings with no treatment-mediator interaction.

Where it strains

Relies on restrictive parametric assumptions — including no effect modification between treatment and mediator — and the resulting “path effects” do not always correspond to any actionable hypothetical intervention.

Many readers meet mediation analysis through this route first — the causal framework above is best understood as clarifying exactly what assumptions this familiar approach was making all along.

WHY THE ARITHMETIC DOES NOT STOP THERE

Three Limits of the Controlled Direct Effect

It can move without mediation

A CDE can differ from the total effect purely through effect modification — the treatment's effect changing with the mediator's level — even when nothing is actually being mediated at all.

No clean “indirect” counterpart

Total effects split neatly into direct plus indirect. Controlled direct effects have no corresponding “controlled indirect effect” that completes the same equation.

It needs a meaningful intervention

A CDE is defined by holding the mediator fixed at some value. If intervening on that mediator is not a coherent idea, the resulting number is not a useful causal quantity.

Together, these are the author's central caution against treating mediation as simple arithmetic — “total effect minus direct effect” is rarely the whole story.

WHY PEOPLE REACH FOR MEDIATION ANALYSIS

Four Different Questions, One Label

Where best to intervene?	Often really just comparing two separate point-treatment effects — not a direct/indirect question at all.
Many related outcomes	A treatment affects several downstream biological markers; each can be studied on its own without invoking mediation.
“Is there a direct effect?”	Often a fairness question in disguise — would a health disparity shrink if a mediator, like access to care, were equalized across groups?
Understanding mechanisms	The classic motivation — but it demands strong, often unstated assumptions about the mechanism to even get started.

Being clear about which of these four is actually being asked changes which analysis — and which assumptions — are appropriate.

THE CLASSIC, AND CONTESTED, FORMALIZATION

Natural Direct and Indirect Effects

Pearl's natural direct and indirect effects compare the outcome under one treatment value to a hypothetical blend: the same treatment, but with the mediator held to whatever value it *would have taken* under a **different** treatment value. That blend mixes two hypothetical worlds at once — a “cross-world” quantity that could never actually be produced by any single experiment, even an ideal one.

Critics have called natural effects “non-manipulable” and unscientific for exactly this reason, and they rest on an extra, empirically unverifiable independence assumption that specific structures — where the treatment itself confounds the mediator's relationship with the outcome — can violate. Despite this, they remain some of the most widely applied mediation tools in the field.

TWO WAYS AROUND THE CROSS-WORLD PROBLEM

Alternatives That Avoid Cross-World Assumptions

Randomized interventional effects

Instead of the mediator's cross-world "natural" value, set it to a value randomly drawn from its own real, observable distribution under a chosen treatment level. This needs only ordinary-style assumptions, no cross-world leap — at the cost of a slightly different, less intuitive quantity.

Separable treatment effects

Imagine the treatment split into two components that could, in principle, be manipulated independently — like a hypothetical cigarette that delivers nicotine but not the other toxins linked to hypertension. This isolates the mediated pathway without any cross-world assumption, provided such separate manipulation is a coherent idea in the first place.

A placebo-controlled, blinded trial already does something similar in practice — separating the "awareness of being treated" component from the active ingredient itself. Extending any of these ideas to several mediators, or to mediators that change over time, remains active methods research.

CAUSAL DISCOVERY

Letting the Data Suggest the DAG

Everything so far assumes a causal DAG built from expert knowledge. But sometimes a research question is at too early a stage for that — this is where causal discovery (or structure learning) contributes: constructing a plausible DAG directly from data.

The underlying idea reverses the usual logic: a causal structure implies specific patterns of statistical (in)dependence. Searching the data for exactly those patterns can help recover — at least partially — the DAG that produced them.

A MORE FUNDAMENTAL LIMIT THAN INSTABILITY

Sometimes There Is No Single “Right” DAG

Even with perfect, infinite data, several genuinely different DAGs can imply exactly the same set of observed conditional independencies. No amount of data can then tell them apart — discovery recovers a whole **equivalence class** of DAGs, not one uniquely correct answer.

Without latent variables

The equivalence class can still leave some arrow directions genuinely ambiguous, even under the simplifying assumption that every relevant variable was measured.

With latent variables

Allowing for unmeasured common causes widens the equivalence class further — more structures become observationally indistinguishable from one another.

This is not a flaw in any particular algorithm — it is a ceiling on what observational data alone can ever settle, no matter how much of it is collected.

AN EXPLORATORY TOOL, NOT AN ORACLE

What Makes Discovered DAGs Trustworthy

Choosing the variables

The search only works on a carefully chosen set of meaningful variables — raw, high-dimensional sensor data first needs distilling into concepts a causal graph can actually represent.

Faithfulness

Assumes every independence in the data reflects the true causal structure, and that different pathways never accidentally cancel each other out.

Causal sufficiency

Assumes no unmeasured common cause exists for any pair of variables in the search — rarely fully plausible, and relaxing it is computationally demanding.

Stability

Search algorithms can be unstable — small changes to the data can produce a noticeably different DAG — so results should be checked by resampling or bootstrap.

Given these caveats, the most defensible conclusions from causal discovery tend to be about the **absence** of a direct or indirect link, rather than confident claims about any one specific edge.

REAL EXAMPLES FROM LIFE-COURSE EPIDEMIOLOGY

What Data-Driven DAGs Have Actually Found

Depression, across a lifetime

A discovery analysis of Danish men followed to age 65 surfaced some plausible directed links — such as childhood social class shaping later intelligence scores — that domain experts had not specifically predicted in advance.

Childhood behaviors and BMI

A European children's cohort study found unexpectedly few and weak *direct* links from early diet, activity, or sleep to adolescent BMI — a pattern experts read as support for a broader “systems hypothesis,” where the general environment matters more than any single targetable habit.

A related analysis of cardiac-surgery recovery times found that many long-assumed predictors of a longer hospital stay were, in fact, indirect rather than direct causes.

WRAPPING UP

From Hill's Criteria to a Formal Framework

For most of the twentieth century, causal interpretation in epidemiology leaned on Hill's famous 1965 criteria — an inductive checklist for weighing whether an association looks causal. Since the 1970s, a parallel, deductive framework has grown alongside it: potential outcomes in the 1970s, methods for time-varying treatments in the 1980s, and causal DAGs in the 1990s.

The two approaches share more than they differ: a stronger association is harder to explain away by confounding under either one, and Hill's call for "specificity" echoes today's use of negative controls. What both agree on is this — without a clear causal question, embedded in an explicit causal model, it is very hard to know which associations are worth trusting.

This formal framework has considerably strengthened observational epidemiology since the 1990s, and remains one of the field's most active areas of research.

WRAPPING UP

Bottom Line

- 1 A causal question needs an explicit target and explicit assumptions before any statistical method can make it convincing.
- 2 Causal DAGs make assumptions about confounding and selection bias visible — and it's the absent arrows that do the real work.
- 3 Positivity can be checked; consistency and exchangeability mostly can't — they have to be reasoned through instead.
- 4 A plain regression rarely suffices for treatments given over time — that is exactly what IPTW, the g-formula, and doubly robust methods are for.
- 5 Emulating a target trial's full design — not only adjusting for confounding — avoids a whole further family of self-inflicted biases.
- 6 Mediation analysis and causal discovery extend the same core logic further — but both demand extra assumptions that deserve to be stated plainly.

SOURCE & ATTRIBUTION

Didelez, V. (2025). “Causal Reasoning and Inference in Epidemiology.”

In: Handbook of Epidemiology, W. Ahrens & I. Pigeot (eds), Springer Nature.

Prepared by Omokhodion Alfred Eriakha · September 19, 2026

An original plain-language educational synthesis based on the source cited above. Diagrams and charts throughout are author-created illustrations of the chapter's concepts, not reproductions of the original published figures. Created for scholarly and educational portfolio purposes. This presentation is not affiliated with or endorsed by the original authors or publisher.