

Causal Inference: From Correlation to Causation

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 13 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: what to know, why it matters at the bedside, and how models fail.

One-sentence thesis. Prediction tells you what tends to happen; causal inference tries to tell you what would change if you intervened, and confusing the two is how an accurate model produces a harmful decision.

PART 1: The Correlation Trap

The French Paradox and the Marker Problem

Observational data showed moderate red wine drinkers had 40 percent less coronary mortality. Randomized trials and biomarker studies (Semba et al., 2014) then found resveratrol did essentially nothing: wine drinkers were wealthier, better educated, more active, and less likely to smoke. The wine was a marker, not a cause.

ANALOGY

Roosters crow before sunrise every day, reliably.
Silencing the rooster will not keep the sun down.

Every supervised model in Chapters 7-8 is a prediction machine of exactly this kind. It learns associations; it does not tell you whether changing an input changes the outcome. Discharging patients on Wednesday instead of Friday will not lower readmissions if Friday discharges are just sicker patients.

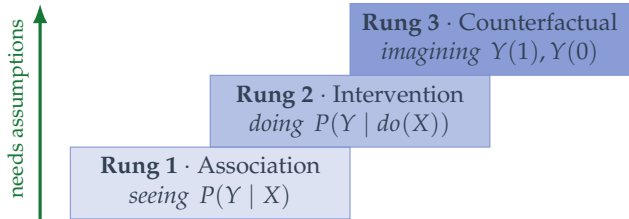
The Optum Algorithm: Prediction Masquerading as Causation

A widely used algorithm (Obermeyer et al., 2019, *Science*) picked patients for intensive care management by predicting future *cost*, assuming cost proxies need. Because Black patients historically incurred lower costs at the same illness severity (unequal access), the model rated them healthier; fixing the proxy raised the share of Black patients flagged from 17.7 percent to **46.5 percent**. It is like ranking firefighters by how much water they used last year, then sending help only to the ones who used the most and missing the neighborhoods that never got a truck.

This was not an algorithm bug; it was a causal-reasoning failure: a Rung 1 model (predict cost) used to answer a Rung 2 question (who benefits from intervention). Any prediction model used to guide action inherits this risk.

Three Rungs of Causation

Pearl’s central claim is that “cause” is not one question but three, stacked in a hierarchy you can only climb with help. Rung 1, *association* (seeing), is standard ML: conditional probabilities, what every supervised model in Chapter 7 already does. Rung 2, *intervention* (doing), asks what happens if I reach in and set X myself. Rung 3, *counterfactual* (imagining), asks whether this specific patient would have recovered had the drug been withheld. Seeing an elevated lactate predict sepsis (Rung 1) does not tell you whether pharmacologically lowering lactate reduces sepsis (Rung 2); the two look identical until a confounder pulls them apart. The entire discipline hangs on one inequality: doing is not the same as seeing.



Pearl’s ladder of causation. No amount of Rung-1 data lifts you to Rung 2 without structural assumptions, the causal hierarchy theorem, not a compute limit.

Formally, the do-operator marks that gap:

$$P(Y | do(X=x)) \neq P(Y | X=x)$$

ANALOGY

Rung 1 reads the thermometer; Rung 2 turns the thermostat; Rung 3 asks what the room would feel like had you never touched the dial.

Watch out. You cannot climb the ladder with data alone. No dataset, however large, moves you from Rung 1 to Rung 2 without structural assumptions; this is the causal hierarchy theorem, a mathematical impossibility, not a compute limit.

Rung	Reasoning	Clinical example	Tool
1	Association	$P(\text{sepsis} \text{lactate high})$	Conditional probability
2	Intervention	Effect of giving drug X	$P(Y do(X))$
3	Counterfactual	Would patient Z have recovered?	Potential outcomes $Y(1), Y(0)$

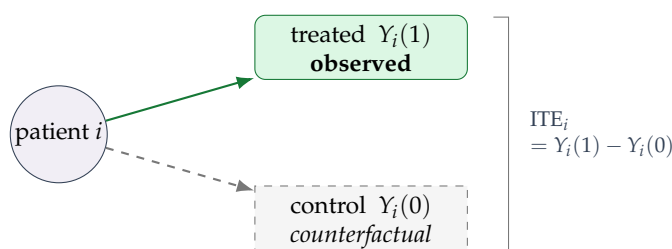
The Fundamental Problem of Causal Inference

To know whether a treatment helped one patient, you would have to see that same patient both ways at the same instant: treated and untreated. You never can. Each patient has two potential outcomes, one under treatment ($Y_i(1)$) and one under control ($Y_i(0)$), but you observe only one; the other is a counterfactual, gone forever, like a fork in the road you cannot un-walk to compare. That missing half is the fundamental problem of causal inference. This is why randomization works: it makes the treated and untreated groups exchangeable on average, so the group differ-

ence estimates the average treatment effect (ATE). The g-formula (Robins, 1986) mimics this from observational data: model the outcome given treatment and confounders, predict everyone as if treated and again as if untreated, then average the difference.

Formally, the individual and average effects are:

$$ITE_i = Y_i(1) - Y_i(0), \quad ATE = E[Y(1) - Y(0)]$$



The fundamental problem: each patient has two potential outcomes but only one is ever observed; the other (dashed) is gone forever. Randomization makes the groups exchangeable, so the average difference estimates the ATE.

The Three Assumptions Everything Rests On

Every method here — the g-formula, matching, weighting — needs three assumptions to hold: consistency, exchangeability, and positivity. Exchangeability (no unmeasured confounding) is the one that almost always fails, because social determinants and physician preferences drive both treatment and outcome

and are rarely captured in the EHR.

Watch out. When exchangeability fails, the estimate is biased in an unknown direction. Observational causal inference is a complement to trials, never a substitute; your conclusions are only as good as your assumptions, and the assumptions are untestable from the data.

Assumption	Plain meaning	Failure mode
Consistency	Treatment means the same thing	"Aggressive fluids" varies by MD
Exchangeability	No unmeasured confounding given L	SDOH absent from the EHR
Positivity	Every profile can get each treatment	Some patients are never treated

Compressing the Chart to One Number

Matching a treated patient to a comparable untreated one sounds easy until you notice that “comparable” means matching on dozens of covariates at once, and no two charts ever line up. Rosenbaum and Rubin (1983) proved you can collapse all of them into a single scalar — the propensity score, the probability of treatment given covariates — and match on that instead. Inverse probability weighting (IPW) keeps every patient by reweighting them into a *pseudo-population* where confounders are balanced: instead of comparing two shoppers item by item, you summarize each by how likely they were to buy at all, then compare like-for-like, handing a megaphone to the surprising shoppers so they speak for the ones the data lacks. These methods let you extract treatment effects from routine EHR and claims data, which is

where value-based-care decisions and shared-savings dollars actually live.

In symbols, the score and the weights it produces are:

$$e(L) = P(A=1 \mid L)$$

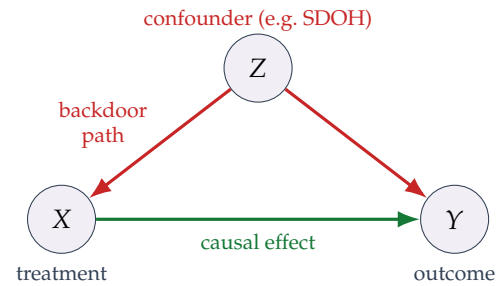
$$w_i = \frac{1}{e_i} \text{ (treated), } \frac{1}{1-e_i} \text{ (control)}$$

Watch out. When a propensity score nears 0 or 1 the weight explodes (a score of 0.01 yields a weight of 100), and one odd patient dominates the analysis — a positivity violation; fix it with weight truncation and stabilized weights. *Doubly robust* estimation combines the outcome model and the propensity model, staying consistent if *either* one is correct, but if both are wrong it is still biased, and no method survives an unmeasured confounder.

Method	Models	Strength	Fails when
G-formula	Outcome	Direct	Outcome model misspecified
Propensity matching	Treatment	Dimension reduction	Discards unmatched; <i>L</i> incomplete
IPW	Treatment	Keeps full sample	Weights explode near 0/1
Doubly robust	Both	Correct if either is	Both wrong; unmeasured confounder

Confounder, Mediator, Collider

A directed acyclic graph (DAG) encodes what causes what, and the three-variable shapes decide what you must and must not adjust for. A *confounder* is a common cause of treatment and outcome (a puppeteer pulling both strings); a *mediator* sits on the causal path between them (a relay runner passing the baton along); a *collider* is a common effect of two variables (a door that two hallways happen to share, and propping it open lets people from one hallway see into the other). Adjust for confounders — that is the whole game. Do *not* adjust for a mediator if you want the total treatment effect, or you erase part of the effect. Never condition on a collider: doing so manufactures a spurious association out of thin air.



A confounder Z opens a backdoor path $X \leftarrow Z \rightarrow Y$ (red) that mimics a treatment effect. Adjusting for Z , and only Z , closes it.

Watch out. The PC algorithm can *learn* graph structure from conditional-independence tests, but it outputs a hypothesis, not proof, and struggles with non-linearity, high dimensions, and missing data; treat its output as a starting point for expert review.

Structure	Shape	Rule
Confounder	$X \leftarrow C \rightarrow Y$	Adjust for it
Mediator	$X \rightarrow M \rightarrow Y$	Do not adjust (for total effect)
Collider	$X \rightarrow Z \leftarrow Y$	Never condition on it

From Average Effects to This Patient (ITE)

Medicine is not practiced on averages. Causal forests (Athey and Wager) split on covariates that maximize *heterogeneity* of the treatment effect, estimating a conditional average treatment effect (CATE) with confidence intervals — a weather app that stops reporting the national average temperature and instead tells you the forecast for your exact street. Frontier work adds NLP-extracted features from clinical notes

(symptom severity, hedging language, stated preferences). This is the machinery of precision medicine: matching the right treatment to the right patient rather than to the population mean.

Watch out. NLP feature extraction introduces measurement error, and error in covariates biases causal estimates, so a note-derived feature can help or quietly poison the analysis.

From Weather Forecast to Flight Simulator

Static methods estimate an effect at one time point. A clinical *world model* learns an action-conditioned dynamics function that simulates how a patient evolves under different interventions over time; a November 2025 survey rates them on four levels. A weather forecast tells you it will rain; a world model is a flight simulator that lets you practice landing in the storm, test a different approach angle, and see what happens. Instead of “35 percent risk of deterioration,” it can say “35 percent under current care, 22 percent if you switch to vasopressin and raise fluids” — the form of information that actually changes clinical behavior. Johns Hopkins’ Medical World Model for liver-cancer TACE planning improved protocol-selection F1 by **13 percent**.

Watch out. Most systems reach only Levels 1-2, use a single modality, cover short horizons, and are weakly validated against real interventional data. The ambition is wide of the current reality.

Should You Build This?

Build it when the decision on the table is an intervention rather than a ranking, whether to fund a

care-management program or whether a treatment actually moves the outcome, and you can name the confounders you are adjusting for and defend consistency, exchangeability, and positivity out loud. Causal machinery earns its complexity exactly where a correlational model would be misread as a recommendation. **Think twice when** unmeasured confounding is plausible and untestable, since doubly robust estimation insures against a misspecified model, not a missing variable; when propensity scores crowd toward 0 or 1 and the weights explode; when the apparent effect could be regression to the mean (Chapter 4’s Camden lesson); and when a graph discovered by the PC algorithm gets promoted from hypothesis to proof. **Who bears the risk:** a miscalibrated baseline moves shared-savings dollars in one direction or the other, and patients bear the downstream consequence when an ineffective program is scaled or an effective one is cancelled on a bad estimate. A causal number carries more authority than a predictive one and is harder to falsify at a glance, which is why the assumptions belong in the write-up.

Level	Capability	Maturity
1	Temporal prediction	Common
2	Action-conditioned prediction	Most systems
3	Counterfactual rollouts	Rare (e.g. JHU MeWM)
4	Planning and control	Almost none

Quick Self-Test

1. A readmission model flags high-risk patients accurately. Why can a post-discharge phone-call program still waste money on exactly the patients it flags?
2. State the do-operator distinction in one line, and name the theorem that says data alone cannot bridge Rung 1 to Rung 2.
3. Which of the three core assumptions almost always fails in EHR data, and why?
4. Your IPW analysis is dominated by three patients with weights near 100. What happened, and what are two fixes?
5. You want the total effect of a drug on mortality. There is a mediator on the pathway. Do you adjust for it? What happens if you condition on a collider by mistake?
6. What can a world model tell a clinician that a risk score cannot?

If You Remember One Thing

Prediction estimates what is likely; causal inference estimates what would change if you act. Healthcare decisions are causal, but most healthcare data and most healthcare AI are correlational, and the Optum algorithm proves that an accurate prediction plugged into a causal question can systematically harm the patients it was meant to help.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552