

Causal Inference Essentials

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Concept guide, "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: Prediction Is Not Causation

The Correlation Trap

Observational data showed moderate red-wine drinkers had roughly **40 percent** less coronary mortality, the "French Paradox." Randomized trials and biomarker work (Semba et al., 2014) then found resveratrol did essentially nothing; wine drinkers were simply wealthier, better educated, more active, and less likely to smoke, so the wine was a marker, not a cause. Every supervised model in Chapters 7 and 8 is a prediction machine: it learns associations, and it does not tell you whether changing an input changes the outcome, so discharging patients on Wednesday instead of Friday will not lower readmissions if Friday discharges are just sicker patients.

ANALOGY

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

Watch out. A model can be superbly accurate and still be the wrong tool. Accuracy answers "what will happen?"; a treatment decision asks "what happens

if I act?" Those are different questions and no amount of AUROC bridges them.

The Optum Algorithm: A Prediction Plugged Into a Causal Question

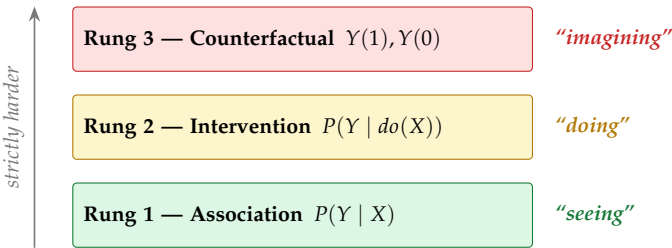
A widely used algorithm (Obermeyer et al., 2019, Science) picked patients for intensive care management by predicting future cost, assuming cost proxies need. Because Black patients historically incurred lower costs at the same illness severity (unequal access), the model rated them healthier; fixing the proxy raised the share of Black patients flagged from 17.7 percent to 46.5 percent and cut the disparity by **84 percent**. This was not a coding bug. It was ranking firefighters by how much water they used last year and then sending help only to the ones who used the most, missing the neighborhoods that never got a truck: a Rung 1 model (predict cost) used to answer a Rung 2 question (who benefits from intervention). Any prediction model wired to guide action inherits exactly this risk.

Three Rungs of Causation

Pearl sorts every causal question onto three rising rungs. Rung 1, association or “seeing,” is standard ML and conditional probabilities. Rung 2, intervention or “doing,” is what happens if I set X, written with the do-operator. Rung 3, counterfactual or “imagining,” asks whether this specific patient would have recovered without the drug. Rung 1 reads the thermometer, Rung 2 turns the thermostat, and Rung 3 asks what the room would have felt like had you never touched the dial. The whole reason the ladder matters is that seeing and doing are generally different quantities:

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

Seeing an elevated lactate predict sepsis (Rung 1) does not tell you whether pharmacologically lowering lactate reduces sepsis (Rung 2); the two look identical in the data until a confounder pulls them apart. And you cannot climb the ladder with data alone: no dataset, however large, moves you from Rung 1 to Rung 2 without structural assumptions, which is the causal hierarchy theorem, a mathematical impossibility, not a compute limit.



Pearl’s ladder of causation. Each rung asks a strictly harder question, and the causal hierarchy theorem says data alone cannot lift you from seeing to doing without structural assumptions.

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

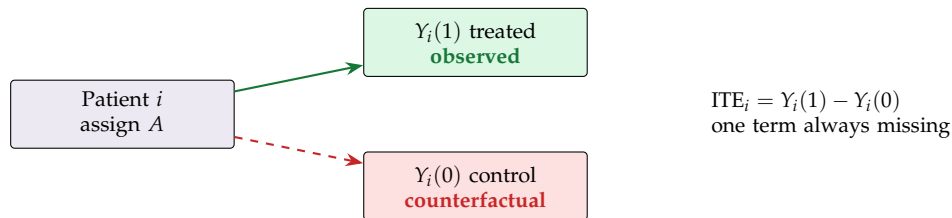
PART 3: Potential Outcomes and the Assumptions Everything Rests On

The Fundamental Problem of Causal Inference

To know whether a treatment helped a patient, you would have to see that same patient both treated and untreated at the same moment, and you never can. Each patient has two potential outcomes, one under treatment and one under control, but only one is ever observed; the other is a counterfactual, gone forever, like the fork in the road you cannot un-walk to compare. In symbols, the individual treatment effect (ITE) is the gap between the two, and the average treatment effect (ATE) is its population mean:

$$\text{ITE}_i = Y_i(1) - Y_i(0), \quad \text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

This is why randomization works: it makes the treated and untreated groups exchangeable on average, so the group difference estimates the ATE. The g-formula (Robins, 1986) mimics this from observational data by modeling the outcome given treatment and confounders, predicting everyone as if treated and again as if untreated, then averaging the difference.



The fundamental problem of causal inference: each patient has two potential outcomes but only one is ever observed. The other is a counterfactual, gone forever, so no individual effect is seen directly.

The Three Assumptions

Every method here, the g-formula, matching, and 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 physi-

cian 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 alone.

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

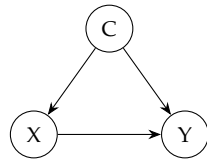
Confounder, Mediator, Collider

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 mediator sits on the causal path between them; a collider is a common effect of two variables. 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, which is how “selecting on who got admitted” invents fake correlations between admission

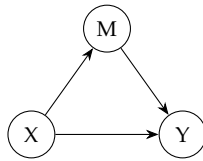
causes. Algorithms like PC can learn graph structure from conditional-independence tests, but they output a hypothesis, not proof, and struggle with nonlinearity, high dimensions, and missing data, so treat the output as a starting point for expert review, not an answer.

ANALOGY

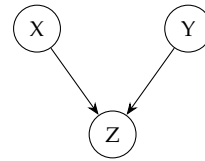
A confounder is a puppeteer pulling both strings; a mediator is a relay runner passing the baton along; a collider is a door two hallways happen to share, and propping it open lets people from one hallway see into the other.



Confounder
adjust for it



Mediator
do not adjust



Collider
never condition

The three-variable shapes that decide adjustment. Adjust for a confounder; leave a mediator alone for the total effect; never condition on a collider, which manufactures a spurious association from nothing.

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

Propensity Scores and IPW

Balancing treated and untreated patients on dozens of confounders at once is hopeless; Rosenbaum and Rubin (1983) proved you do not have to. You can match on a single scalar, the propensity score, the probability of treatment given covariates. Inverse probability weighting (IPW) uses that same score differently: instead of matching, it keeps every patient and reweights them into a pseudo-population where confounders are balanced, handing a megaphone to the surprising shoppers so they speak for the ones the data lacks.

$$e(L) = P(A=1 | L)$$

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

These methods extract treatment effects from routine EHR and claims data, which is where value-based-care and shared-savings decisions actually live, and the same machinery drives off-policy evaluation in Chapter 8: when a model is meant to change decisions, IPW and doubly robust estimators reweight historical cases to estimate what would have happened under the model's policy, not the clinician's. **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

With messy observational data you rarely trust any single model to be right. Doubly robust (augmented IPW) estimation hedges that bet: it combines an outcome model ($\hat{\mu}$) and a propensity model ($\hat{e}$), and its remarkable property is that the estimate stays consistent if either one is correctly specified, so you get two chances to be right instead of one, like a climber clipped into two anchors where either rope alone will hold. In symbols, it adds a propensity-weighted correction to the outcome-model difference:

$$\hat{\tau}_{\text{DR}} = \frac{1}{n} \sum_i \left[(\hat{\mu}_1 - \hat{\mu}_0) + \frac{A_i(Y_i - \hat{\mu}_1)}{\hat{e}_i} - \frac{(1 - A_i)(Y_i - \hat{\mu}_0)}{1 - \hat{e}_i} \right]$$

That is why doubly robust methods are the workhorse of modern causal machine learning and off-policy policy evaluation. But if both models are wrong it is still biased, and no method here survives an unmeasured confounder: doubly robust is insurance against misspecification, not against violated assumptions.

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

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 does its failure bias the estimate in an unknown direction?
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 and there is a mediator on the pathway. Do you adjust for

- it? What happens if you condition on a collider by mistake?
6. What does “doubly robust” buy you, and what does it explicitly not protect against?

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. The Optum algorithm is the proof: an accurate prediction plugged into a causal question systematically harmed the patients it was meant to help, and no method, propensity score, IPW, or doubly robust, can rescue you from an assumption that never held.

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