

Skewed Distributions and the Cost of Care

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 4 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. In healthcare cost data the average patient tells you almost nothing and the tail tells you almost everything, so your statistical and target choices decide whether a model helps patients or harms them.

PART 1: The Shape of the Curve

Why Costs Are Not Normal

Start with a fact from the ward, not a statistic. In any insured population most people cost almost nothing in a given year, a handful cost a fortune, and that handful drives the entire budget. U.S. healthcare spending is one of the most extreme right-skewed distributions in any industry: the mode sits near zero, the median insured adult spends \$3,000 to \$5,000, but the mean is dragged above \$12,000 by a long tail reaching into the millions. Report the "average" cost and you have described almost no one.

ANALOGY

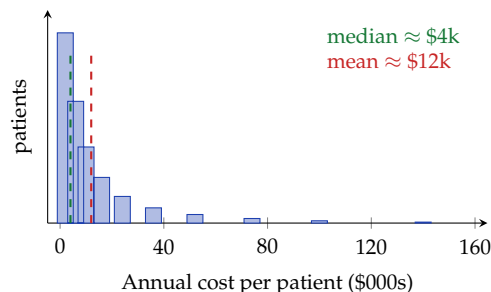
Most patients live on the flat prairie of the curve; a few costs shoot up like skyscrapers and drag the average skyline with them.

Only once you have seen that shape do the summary

statistics mean anything, and they are extreme. Formally, the distribution announces itself in three numbers:

skewness > 10 , kurtosis > 100 , mean $\gg$ median

The skew is not a cosmetic nuisance, it breaks standard tools. Ordinary least squares assumes normal, constant-variance residuals, so a linear model draws one compromise line and systematically underpredicts the expensive patients you most need to find while overpredicting the healthy majority. Summary means mislead, and t-tests and ANOVA fail because the tail keeps sampling distributions skewed even at large N.



The long tail of healthcare cost. A spike near zero and a thin tail of very expensive patients drag the mean far above the median, so "average cost" describes almost no one.

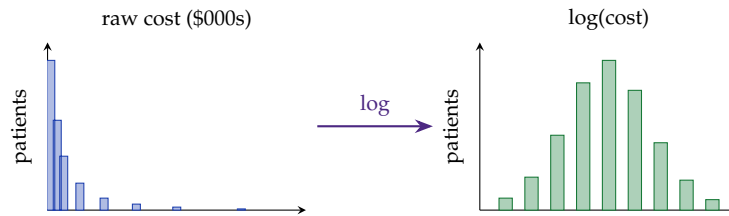
Modeling the Skew

So how do you fit a target shaped like that? Four workhorses handle right-skewed cost. Log-transform plus OLS compresses the tail but needs Duan's smearing estimator to back-transform without bias. Two-

part models split the probability of any cost (logit) from the amount given positive (continuous). Gamma or Tweedie GLMs model the skew directly. Gradient-boosted trees (XGBoost, LightGBM) split on rank order and ignore the absolute scale. **Watch out.** Even

trees need careful target specification. The choice of target, raw cost, log cost, cost quantile, or a binary high-cost flag, decides whether the model is clinically useful or merely statistically impressive. Before you

model anything, plot the distribution; if you see a bell curve, you have truncated the data or excluded the population that matters.



Log-transforming compresses the right skew into a roughly symmetric shape OLS can model; back-transform with Duan's smearing estimator to avoid bias.

The 5 Percent Rule

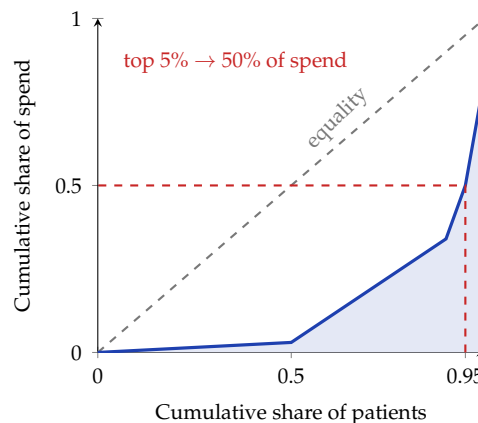
The single most useful fact about healthcare cost is that it concentrates: a small slice of patients drives most of the spend, so finding that slice is most of the value. It is twenty people at a dinner where one guest orders half the entire bill. Per MEPS, the top 5 percent of spenders account for about half of all expenditures (49.7 percent in 2022), averaging roughly \$73,000 each; the top 1 percent averages about \$147,000; the bottom 50 percent averages under \$500. In a \$5.3 trillion system, roughly one in twenty people drives about \$2.6 trillion. As a rule of thumb:

top 5% $\Rightarrow$ ~ 50% of spend

top 1% $\Rightarrow$ 22-24%

bottom 50% $\Rightarrow$ < 3%

But the tail is not a fixed roster. Only 30 to 35 percent of top-decile patients stay there year to year; the rest regress to the mean as an acute episode resolves. A naive model trained on last year's high-cost patients captures only a fraction of next year's, so the value of prediction is a ranking of who will benefit from intervention, not a dollar figure.



Concentration (Lorenz) curve of U.S. health spending. The farther the curve bows below the equality line, the more spend concentrates in a few patients: the top 5 percent account for roughly half.

Concentration of U.S. health spending (MEPS, 2024 levels)

Spending tier	Share of total	Avg annual cost	Approx absolute dollars
Top 1%	22–24%	~\$147,000	~\$1.2T
Top 5%	~50%	~\$73,000	~\$2.6T
Top 10%	~66%	~\$48,000	~\$3.4T
Bottom 50%	<3%	<\$500	~\$150B

■ Super-Utilizers and the Camden Lesson

The tail has a name in the delivery system: super-utilizers, patients with five, ten, or twenty hospital visits a year whose charts read as everything upstream failing at once, uncontrolled chronic disease, behavioral health needs, housing instability, and no one coordinating any of it. The most famous attempt to act on cost concentration began with them. In Camden, New Jersey, the physician Jeffrey Brenner mapped hospital claims block by block and found roughly **1 percent** of patients driving about **30 percent** of the city's hospital costs. His Camden Coalition wrapped nurses, social workers, and care coordinators around the highest utilizers, and "hotspotting" became a national model.

Then the model met a randomized controlled trial. In January 2020, Finkelstein and colleagues published in the *New England Journal of Medicine* the results of randomizing 800 super-utilizers to the Coalition's signature care-transition program or to usual care:

the 180-day readmission rate was about **62 percent in both arms**. The program that defined a movement produced no detectable effect on its primary outcome.

Watch out. The failure was statistical before it was clinical. Patients were enrolled at the peak of their utilization, and, as the churn numbers above imply, people selected at an extreme drift back toward their own average no matter what anyone does; the control group improved almost exactly as much as the treatment group. Every uncontrolled before-and-after evaluation of a tail-targeted program is flattered by that regression to the mean, which is why so many care-management pilots report savings that vanish under randomization. Two lessons follow. If you select patients because they are extreme, your evaluation needs a real control group. And prediction is not tractability: identifying expensive patients is the easy half, and the value lives in the harder question of who is impactable, whose trajectory an available intervention can actually change.

Informative Missingness

Intro statistics treats missing data as noise to impute or delete, assuming it is missing completely at random. In clinical data that is almost never true. A physician who orders a troponin and chest X-ray but skips the metabolic panel has encoded a clinical hypothesis; the gap is a decision, not an empty box. In a detective's notes, the questions the detective chose not to ask reveal the theory they were already working. Che and colleagues' GRU-D architecture treats the missingness mask and time-since-last-observation as explicit features, and missingness indicators alone can rival observed values for some outcomes. A missing A1c may mean the patient has not seen a doctor in two years, not that diabetes is controlled. **Never drop missing indicators.** Add a binary "was this ordered" feature alongside any imputation; in readmission and cost models these often rank among the top predictors.

Informative Censoring

Censoring is when a patient exits the observation window before the event, and the standard tools assume it happens for reasons unrelated to risk. Kaplan-Meier and Cox regression treat censoring as non-informative, which is like losing a plane because it left your radar screen, not because it left the sky. In fragmented U.S. data that assumption routinely breaks: a patient readmitted to a hospital 60 miles away, a patient who dies at home, a patient who switches in-

surers, all vanish from your denominator for reasons tied to their outcome. **Watch out.** Building a model only on patients observable for the full window introduces survivorship bias and underestimates risk for the highest-need patients. Use inverse-probability-of-censoring weighting, competing-risk models (Fine-Gray, since a patient who dies cannot be readmitted), or newer neural survival methods (DySurv) that gain about 10 percent in concordance when censoring exceeds 90 percent. Linking data across payers and providers shrinks the censoring rate.

The Young Patent: Time-Varying Risk

A risk score computed once at discharge treats a patient as if their danger were frozen. It is not. U.S. Patent Application 18/323,518 moves beyond a single binary discharge-time prediction to a risk score that updates as post-discharge signals arrive. Readmission hazard peaks in the first 72 hours, declines through two weeks, then rises again as short-term resources run out; the system flags "risk trajectory divergence" when a patient deviates from their expected recovery path. A static "32 percent readmission risk" at discharge is useful, but detecting at day four that a specific patient jumped from 32 percent to 58 percent because they never filled their cardiac medications is operationally transformative. Prediction alone is insufficient; the point is to trigger a call, a home visit, or a medication reconciliation at the moment it can change the outcome.

SDOH Often Beat Lab Values

Here is a fact that unsettles people who trust the lab panel above all else: where a patient lives often predicts their health better than what their bloodwork says. The WHO attributes 30 to 55 percent of health outcomes to social determinants; a widely cited breakdown puts clinical care at only about 20 percent, with behaviors (30 percent), social and economic factors (40 percent), and environment (10 percent) driving the rest. Zip code, housing instability, food insecurity, and transportation frequently outpredict individual lab values for utilization, readmission, and cost.

ANALOGY

A person's zip code often predicts their health more accurately than their genetic code.

The catch is that SDOH is badly captured in structured data; ICD-10 Z-codes (Z55-Z65) appear in under 2 percent of encounters. The signal lives in free-text notes ("third-floor walk-up, no elevator; cannot afford medications"). NLP systems now close the gap: SDOH-NLI frames extraction as entailment, SynthSBDH uses GPT-4-generated passages (yielding up to 62.5 percent macro-F1 gains), and instruction-tuned LLMs reach F1 above 0.9. Adding SDOH lifted a sepsis-readmission model's AUROC from 0.75 to 0.80.

The Optum Precedent: The Target Is Ethical

The most important lesson in this chapter is not a method, it is a warning about the target variable. In 2019 Obermeyer and colleagues showed in *Science* that a widely used Optum care-management algorithm, which predicted cost as a proxy for health need, systematically under-identified Black patients. Because structural barriers cause Black patients to receive less care at the same illness level, they generate lower costs, so the model inferred they needed less help; at a given risk score, Black patients were substantially sicker. The algorithm was not broken, it predicted cost perfectly. The failure was the choice of target. Cost measures consumption, not need, and

consumption is shaped by access and bias. Retraining on direct health measures raised Black-patient identification from **17.7 percent** to **46.5 percent** and cut the disparity by 84 percent. XGBoost, neural nets, and transformers all faithfully reproduce the bias baked into the label; in healthcare, "what are we predicting?" is never purely technical.

Risk Management vs. Risk Selection

The same cost model can point in two opposite directions. Used one way it directs resources to patients who need help (risk management); used another it helps a plan avoid enrolling and spending on expensive patients (risk selection). Under risk management, accurate SDOH is a feature; under risk selection, it becomes a tool for discriminatory benefit design. **Watch out.** A readmission model that accurately finds high-risk patients but is used to avoid admitting them has been weaponized against the population it was meant to help. You will not always control how your model is used, but you control the target variable, the feature set, and whether you validate performance separately across racial, ethnic, and socioeconomic subgroups. Those are ethical decisions disguised as technical ones.

Should You Build This?

Build it when the prediction feeds risk management: someone owns an intervention (care-management enrollment, transportation, discharge support) aimed at the small fraction of patients who drive most spending, and the target measures health need rather than consumption alone. **Think twice when** the same scores could serve risk selection, or when the target is cost, since the Optum algorithm predicted spending faithfully and so ranked Black patients as healthier than equally sick white patients because they had received less care. **Who bears the risk.** Misclassified patients absorb the error, excluded from help they needed or targeted for avoidance they never see, and the burden falls hardest on the populations whose access gaps the cost proxy encodes.

Quick Self-Test

1. Why does OLS on raw cost data underpredict exactly the patients you most need to identify?
2. State the 5 percent rule, and explain why year-to-year churn makes a naive high-cost model miss most of

next year's high-cost patients.

3. The Camden hotspotting trial found roughly the same 180-day readmission rate in both arms. What statistical mechanism explains that, and what does it imply about evaluating any program targeted at the tail?
4. A lab column is 40 percent missing. Why might dropping it destroy the most informative signal, and what should you do instead?
5. Give one example of informative censoring in claims data and name a method that corrects for it.
6. Why did the Optum algorithm under-identify Black patients even though it predicted cost accurately, and what single change fixed most of the disparity?
7. Distinguish risk management from risk selection using the same cost model and the same SDOH features.

If You Remember One Thing

The tail is the whole game, and the tail is where standard tools break. Plot the distribution, respect the skew, treat missingness and censoring as signals rather than nuisances, and remember that the SDOH forces driving the far right of the cost curve are exactly what the clinical record captures worst. Above all, the target variable is an ethical choice: predict cost when you meant to predict need and, like Optum, you will encode the system's inequities into every prediction the model makes.

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