

Exploratory Analysis and Clinical Visualization

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 5 of "Healthcare Analytics and AI: Building Systems That Actually Work"

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

One-sentence thesis. A clinical chart earns its place only if a time-pressed clinician can extract the one decision it supports in the seconds it takes to walk between exam rooms; a statistically precise chart no one acts on is still a failure.

PART 1: The Data Is Not Bell-Shaped

Skewed Costs and the Long Tail

Before you choose a chart, look at the shape of what you are charting, because healthcare cost is not a hill. It is a spike near zero with a long thin tail of very expensive patients: the median annual cost for a commercially insured adult sits around \$3,000 to \$5,000, but the mean runs two to three times higher because a few patients cost over \$100,000, and the skewness coefficient routinely exceeds 10. Plotting that is less like drawing a bell curve and more like drawing a city skyline, one flat block of low-rise houses and a handful of skyscrapers. The visualization trap is that default histograms and linear axes cram 95 percent of patients into the first bar and shove the costliest 5 percent, the ones you most want to intervene on, into invisible specks at the margin. Use log axes, box plots with explicit outlier marks, and quantile groupings (deciles) so you show the distribution you actually have, not the one you wish you had.

Censored Survival Times

When the outcome is time-to-event (death, readmission, recurrence), some patients leave the study before the event happens: they move, switch plans, or are still alive when it ends. You know only that the event had not occurred yet, which is right censoring, and it is everywhere in healthcare data. Think of a book you had to return to the library before finishing: you know the hero was alive on the last page you read, not whether they survived the ending. **Watch out.** If you exclude censored patients your cohort looks artificially sicker; if you treat "last seen alive" as "died" it looks artificially healthier. The fix is survival analysis,

which models censoring explicitly, and on any time-to-event plot the small tick marks are not decoration, they mark where each patient left the observed story.

Cohort Definition and Immortal-Time Bias

The subtlest way a chart lies is through how the cohort was assembled. Immortal-time bias occurs when person-time during which a patient could not possibly have had the event is wrongly credited to the treatment group, shrinking that group's rate. It is giving one racer a four-week head start and then announcing they are faster: to enroll in a cardiac rehab program that starts four weeks after discharge, a patient must first survive those four weeks, while the control group still contains the people who died in week one. In symbols, the quantity being distorted is a rate per unit of time at risk:

$$\text{incidence rate} = \frac{\text{events}}{\text{person-time at risk}}$$

Rigorously defined by Suissa (2008) and popularized by Levesque and colleagues in the BMJ (2010) using statins and diabetes progression, cohort errors produce confident, reproducible, wrong results with no error message. Align follow-up with eligibility (not treatment start), use time-varying covariates or landmark analysis, and always annotate your cohort timeline so the audience can see the temporal structure. A related trap: patients with complete records are healthier and more engaged, so informative missingness is itself a signal your charts must acknowledge.

Interrogation, Not Exploration

Data scientists explore; clinicians interrogate. A hospitalist rounding on twenty patients wants three answers fast (who is deteriorating, who can go home, what critical result have I not seen), so the chart is a newspaper headline, not a research notebook: the headline must deliver the story on its own, with detail waiting in the article for those who choose to read on. The principle that makes that possible is Tufte's data-ink ratio, the fraction of the graphic actually devoted to data:

$$\text{data-ink ratio} = \frac{\text{ink encoding data}}{\text{total ink on the page}}$$

Maximize it, and strip decoration, redundant labels, and chartjunk. One academic center shipped a beautiful Tableau sepsis dashboard with 23 metrics; usage logs showed clinicians touched exactly two, the SOFA trend and time since last antibiotic. The design principle is subtraction, not addition: every panel that does not answer a question the clinician actually asks is



A clinical color grammar. Red is reserved for values that demand action now, never for "merely elevated"; missing data is gray, never blank.

Lead with the Exception

Show what changed, not what is normal. Sort a patient list by acuity with the sickest on top, rather than an alphabetical census with severity badges scattered through it; the model is an emergency triage board, not a filing cabinet, where the most urgent case sits at the top and is readable from across the

not neutral, it drains the scarce attention examined in Chapter 6.

Color Is Semantic, Not Aesthetic

In a clinical display red means "act now," not "above average" or "interesting," so reserve it for values that demand immediate action. A workable palette: red = critical (potassium above 6.5, systolic BP below 80), yellow = attention within hours, green = normal, gray = missing or not yet resulted.

ANALOGY

Red is a fire alarm, not a highlighter. Pull it for a fire, not to mark a sentence you found neat.

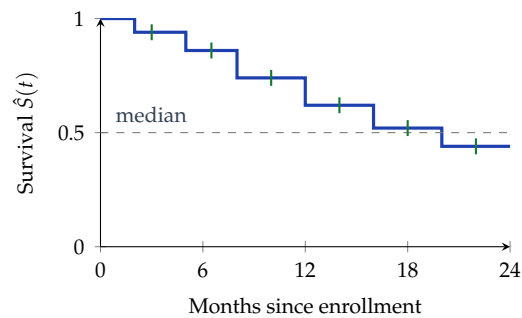
Miscoloring a merely elevated value red triggers a trained threat response and burns trust. **Never leave gaps blank.** Clinicians read blank as "normal," which can be lethal, so color missing data gray. Annotate for context too: "12 percent readmission" is meaningless until you add a benchmark line and a plain-language tag like "above national average."

room. Eliminate interaction: the default view, before any click or filter, must answer the primary question. Offer drill-down, never require it, because every click, dropdown, and filter is a tax on cognitive load. If the instant-load view does not answer the clinician's most important question, the dashboard is abandoned within a week regardless of how good the underlying data is.

Kaplan-Meier Curves

The clinical question is simple: what fraction of patients are still event-free at each point in time? The Kaplan-Meier estimator answers it as a stepwise survival scoreboard that ticks down only when something actually happens. Formally, it multiplies the surviving fraction across every event time:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$



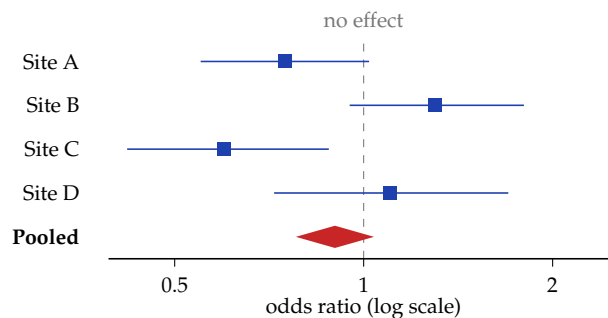
It steps down at each event (d_i events among n_i still at risk) and stays flat between events; median survival is read where the curve crosses 0.5, and if it never crosses, median survival was not reached, which is good news. Censoring ticks and a “number at risk” row are essential, and the far-right tail is unreliable once few patients remain. Comparing two groups with a log-rank p-value is the standard way to show a treatment effect.

A Kaplan-Meier curve. It steps down only at events; green ticks mark censored patients, and the far-right tail is unreliable once few remain at risk.

Forest Plots

When you need to know whether an effect holds up across studies or sites, the forest plot is a jury poll drawn as a chart: you see not just the verdict but how much each juror agreed. Each row is a study or subgroup, a horizontal line is its confidence interval, the square (sized by weight) is the point estimate, and a vertical line at 1.0 (ratios) or 0 (differences) marks

“no effect”; the pooled diamond at the bottom summarizes, and if it clears the no-effect line the overall finding is significant. Their real power is exposing heterogeneity. When you validate a model across sites, plotting each site as a “study” instantly shows the spread: an AUC of 0.82 overall but ranging 0.71 to 0.93 tells a very different story than a tight 0.80 to 0.84.



A forest plot. Each square is a site’s estimate (size = weight) with its confidence interval; the diamond’s width is the pooled interval. Spread across sites exposes heterogeneity a single number hides.

Patient Timelines (Swimmer Plots)

To see temporal structure, give each patient a lane. A swimmer plot draws one horizontal bar per patient across their observation window, with events (admis-

sions, surgeries, med changes, abnormal labs) marked along it; stacked and sorted by outcome, patterns jump out, earlier drug administration, adverse-event clustering, gaps in care. Because they lay time out explicitly, timelines make immortal-time bias visible:

if the treatment group's bars all start later than the control group's, the bias is obvious to anyone reading the plot.

$$\text{rate} = \frac{\text{events}}{\text{population}} \times 100,000$$

Population Heatmaps and Sankey Trajectories

To see where a burden concentrates, use a heatmap, a weather map for disease that encodes a metric across regions or units by color intensity. The critical choice is normalization, because raw counts mislead: larger areas have more cases by size alone, so use rates or age-adjusted rates:

A dark-red county is alarming; a dark-red county labeled "42.3 percent vs. national average 11.6 percent" is actionable. Sankey (alluvial) diagrams are the companion for flow, a river delta for treatment pathways where the ribbon width is patient volume, and they surface unwarranted variation: if two hospitals treating the same cancer produce completely different flow diagrams, it raises immediate protocol questions.

Chart decision guide

Chart	Clinical question	Key annotation	Pitfall
Kaplan-Meier	Fraction event-free over time?	Censoring ticks; number-at-risk row	Late tail unreliable when few remain
Forest plot	How consistent across sites/studies?	Pooled diamond; per-row CI	Heterogeneity hidden if rows omitted
Patient timeline	When did events occur per patient?	Treatment-start anchor	(Feature) reveals immortal time
Population heatmap	Where is disease burden concentrated?	Age-adjusted rate, not raw count	Bigger population reads darker
Sankey / alluvial	How do patients flow between treatments?	Cohort size at each node	Rare paths visually exaggerated

Interesting vs. Actionable

The distinction that decides a dashboard's fate is whether it is actionable, meaning the person viewing it can do something different, right now, within their scope of authority. "Heart-failure readmissions are 18 percent higher on Friday discharges" is interesting to an administrator who can adjust staffing, and useless to the 7:00 AM charge nurse who cannot change the day of the week. It is a weather forecast that only tells you last year's average rainfall: true, tidy, and no help deciding whether to carry an umbrella today. Both interesting and actionable have value, but they need different audiences and delivery, and confusing the two is why so many dashboards become administrative reporting mislabeled as clinical decision support.

Failure mode	Mechanism	Fix
Outside workflow	Requires separate app or URL	Embed in EHR sidebar, single sign-on
Never pruned	Stale metrics pile up	Quarterly panel-retirement review
Yesterday's question	Overnight refresh in acute care	Near-real-time feed for high-acuity panels
No feedback loop	Behavior change unmeasured	Pre/post outcome metric per panel

The COVID-19 Lesson at Global Scale

The clearest proof of these principles ran at global scale. The Johns Hopkins dashboard (Dong, Du, and Gardner, 2020) won through radical simplicity, three numbers and a map with no filters, a single standardized source of truth, transparent methodology, and a design aimed at one audience's one question: "how bad is it, and is it getting worse?" It logged over 3.6 billion page views. Hundreds of state and county dashboards failed through feature creep (15 to 30 panels), inconsistent definitions of "cases" and "deaths," unlabeled update lag, and raw counts with no "so what."

ANALOGY

A single well-lit exit sign beats a wall covered in a hundred blinking notices when the building is on fire.

The parallel to the charge nurse's dashboard is exact: stakeholders will always demand more metrics, and the gap between "comprehensive" and "usable"

The Four Failure Modes

Dashboards go unused for structural reasons, not technical ones. They live outside the EHR workflow, they are never pruned, they answer yesterday's question (an overnight refresh in an acute setting), and they have no feedback loop measuring whether they changed behavior, so they become exercise equipment that turns into a coat rack, installed with enthusiasm, never measured, never removed. A 2022 systematic review (Rabiei and Almasi) found the literature rich on building dashboards and nearly silent on clinical impact. Because dashboards compete for the same scarce attention as alerts and pages (Chapter 6), an unused panel is not neutral, it degrades the attention available for the panels that matter.

widens every sprint. Comprehensiveness is not a virtue. Clarity is.

Should You Build This?

Build it when a named user faces a recurring decision inside their scope of authority and the panel can answer it within the 30-second window, embedded where the decision happens, with data fresh enough to act on. **Think twice when** the finding is merely interesting (readmissions run higher for Friday discharges, but the charge nurse cannot change the calendar), or when the dashboard would live outside the EHR, refresh overnight for acute decisions, or join a wall of panels nobody prunes. Shipped without a feedback loop, you will never know whether it changed behavior, the most consistent gap in the dashboard literature. **Who bears the risk.** An unused dashboard is not neutral: it spends scarce clinician attention and degrades the panels that do matter, so the cost of a design failure lands on the clinicians it distracts and the patients behind the decisions it failed to support.

1. Why does a default histogram of per-patient cost visually erase the patients who matter most, and what three fixes address it?
2. On a Kaplan-Meier curve, what do the small tick marks mean, and why is the far-right tail often unreliable?
3. A cardiac-rehab program shows 20 percent lower one-year mortality. What bias should you suspect first, and what makes it visible on a patient-timeline plot?
4. Why is red a semantic choice rather than an aesthetic one, and why must missing data be gray rather than blank?
5. Give one metric that is “interesting” to an administrator but not “actionable” for a bedside nurse, and explain the difference.
6. Name two things Johns Hopkins did that the failing COVID dashboards did not.

If You Remember One Thing

Healthcare data breaks the assumptions of intro statistics (skew, censoring, cohort bias) and clinicians read charts under brutal time pressure. The discipline that separates a dashboard that changes behavior from one that joins the archive is not building the chart, it is deciding what to remove, what to emphasize, and what to annotate. Comprehensiveness is not a virtue; clarity is.

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