

Advanced Predictive Modeling, Explainability, and Survival Analysis

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 8 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 a stronger model is not automatically a better model: if a clinician cannot understand when to trust it, or if it silently breaks when the clinical world shifts, it will never change practice.

PART 1: Beyond a Single Classifier

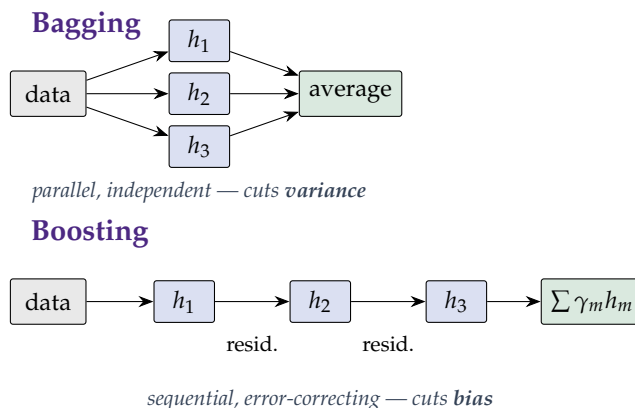
Ensemble Methods

One model is one opinion, and a panel is usually wiser than its smartest member. Ensemble methods build the panel: instead of one model you combine many weak learners, usually decision trees, into a composite predictor, and two families dominate tabular clinical data. A **random forest** trains hundreds of trees in parallel on bootstrap samples and averages them, a bagging strategy that cuts variance; picture it as polling many independent clinicians and averaging their reads. **Gradient boosting** (XGBoost, LightGBM, CatBoost) instead builds trees sequentially, each one fit to the errors of the ensemble so far, which cuts bias, like a case conference where each new clinician is handed only the cases the room has gotten wrong so far and told to focus there.

In symbols, the ensemble prediction is just a weighted sum of those base learners h_m :

$$F_M(\mathbf{x}) = \sum_{m=1}^M \gamma_m h_m(\mathbf{x})$$

Clinical relationships are rarely linear: mortality risk accelerates sharply after age 75, and two individually benign comorbidities can interact into serious risk. Trees capture these thresholds and interactions automatically, which is why gradient-boosted trees consistently top logistic regression and even deep nets on 50-500 feature EHR extracts. That performance has a price. **Watch out.** Ensembles trade interpretability for performance, so when regulators or clinicians demand coefficient-level transparency a locked logistic regression can still be the right deployment choice, no matter how many explanation plots you bolt on afterward. Neural networks rarely beat boosted trees on purely tabular data and cost far more to audit and maintain.



Two ensemble families. Bagging (random forest) averages independent trees to cut variance; boosting adds trees fit to the running errors, cutting bias.

Censoring: Why Classifiers Fail Here

Survival analysis models not whether an event happens but *when*, and its defining feature is **censoring**: a patient still alive at study end has not had the event, but their true event time is unknown. Dropping censored patients wastes data and biases results, while treating them as non-events is equally wrong. Clinical questions “what is one-year survival?”, “when will this graft likely fail?” are time-to-event questions that a yes/no classifier cannot answer.

ANALOGY

A censored patient is someone who left the movie theater before the ending. You know what happened up to that point, not how the story finishes.

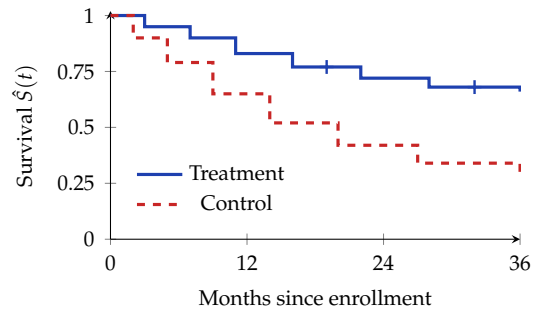
The Kaplan-Meier Estimator

The first question in any time-to-event study is simply this: what fraction of patients is still event-free at each point in time, given that some dropped out of view along the way? The Kaplan-Meier estimator answers it without assuming any shape for the curve. It walks forward in time, and at each moment an event occurs it multiplies in the conditional probability of surviving that instant (d_i events among the n_i still at risk), producing the descending step function you see in every oncology and transplant paper.

Formally, the survival function $S(t)$, the probability of surviving past time t , is the running product:

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

It makes no distributional assumptions and handles censoring naturally, and the log-rank test compares two such curves. Its limit is that it describes groups, not the effect of individual covariates.



Kaplan–Meier survival curves for two groups. Each drop is an event; the descending step function multiplies conditional survival probabilities, and + marks a censored patient. The log-rank test compares the curves.

The Cox Proportional Hazards Model

Kaplan-Meier describes whole groups, but the clinician usually wants the effect of one factor: how much does diabetes, or a given treatment, change *this* patient’s risk? The Cox model answers that. It says the hazard, the instantaneous event rate at time t , is an unspecified baseline hazard scaled up or down by an exponential function of the patient’s covariates.

In symbols:

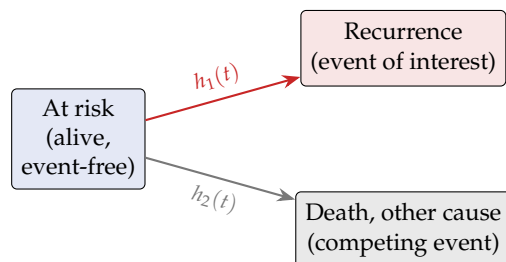
$$h(t | \mathbf{X}) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

The **proportional hazards** assumption says the hazard ratio between any two patients is constant over time. A hazard ratio of 2.0 is a patient who runs the same race but at double the risk-speed at every mile marker, at one month, six months, and five years alike. Cox output speaks the clinician’s language: “diabetes carries a hazard ratio of 1.45 (95% CI 1.22-1.71) for cardiovascular mortality” is immediately interpretable, and the Framingham score, MELD, and many cancer staging systems are Cox-derived. Test the proportional hazards assumption with Schoenfeld residuals; it breaks down over long follow-up when a treatment effect wanes, and then you need time-varying coefficients or accelerated failure time models.

Competing Risks and the Fine-Gray Model

Standard survival analysis assumes everyone would eventually have the event if followed long enough. **Competing risks** violate this: a patient who dies of

a heart attack can never have the cancer recurrence you were tracking. The Fine-Gray model targets the **subdistribution hazard** and reports the **cumulative incidence function (CIF)**, the honest real-world probability of the event. Treating competing deaths as mere censoring inflates estimated risk; Austin and Fine (2017) showed Kaplan-Meier can overstate 5-year recurrence by 10-15 percentage points in elderly populations, a gap that flips a decision between aggressive treatment and watchful waiting. If a competing event has meaningful incidence, roughly above 10 percent, you must use competing-risk methods or you will overstate risk and drive overtreatment.



A competing event removes the patient from risk, so it cannot be treated as ordinary censoring.

Competing risks. A patient exits the at-risk pool by the event of interest or by a competing event; Fine-Gray models the subdistribution hazard and reports the cumulative incidence function (CIF).

The Concordance Index

How do you grade a survival model when many patients are censored and you never learn their true event time? You fall back on ordering: among every pair of patients you *can* compare, did the model give the higher risk to the one who actually had the event sooner? That fraction is the concordance index, the survival analog of AUROC, with 0.5 a coin flip and 1.0 perfect ordering.

$$C = P(\hat{r}_i > \hat{r}_j \mid t_i < t_j)$$

It measures ranking under censoring, so it is the standard discrimination metric for Cox and survival models, but like AUROC it says nothing about calibration, so pair it with a calibration check before you trust absolute risk numbers.

Temporal Validation Is Non-Negotiable

A random train-test split is not sufficient validation for a clinical model. Random splitting makes the test set statistically exchangeable with training, so it flatters the model: it lets the model practice on Monday's exam and then grades it on shuffled photocopies of Monday's exam, whereas temporal validation changes the test. A 2024 JAMA Network Open deterioration model scored AUC 0.89 on a random split but 0.82 under temporal validation, a real degradation the random split would never expose. The validation hierarchy runs, weakest to strongest, random

split < cross-validation < temporal split < geographic split < prospective validation; temporal validation is the minimum before deployment, and geographic validation is strongly advised for multi-site use.

Distribution Shift Is the Expected Condition

The statistical link between inputs and outputs changes between training and deployment, and in healthcare this is the norm, not an edge case. COVID-19 was the canonical demonstration: Wynants et al. reviewed 731 pandemic prediction models and rated every one at high risk of bias, and well-validated pre-pandemic models broke overnight.

Shift type	Changes	Clinical example	How to detect
Covariate shift	$P(X)$	New referral network sends sicker patients	KS / Wasserstein test on features
Label shift	$P(Y)$	Readmission rate moves 15% to 22% post-pandemic	Predicted vs observed outcome rate
Concept drift	$P(Y X)$	New sepsis bundle cuts mortality at fixed risk	Calibration drift; outcome audit

Watch out. Concept drift is the most insidious of the three, because even a perfectly calibrated model with unchanged inputs produces wrong predictions. The defenses are operational, not optional: continuous drift monitoring, periodic retraining (quarterly or semiannually), and fairness auditing. Between 2022 and 2025, 60 authorized AI devices were tied to 182 recall events, 43 percent of them within a year of authorization, the “zombie algorithm” problem.

Off-Policy Evaluation

When a model is meant to *change* decisions, not just predict outcomes, historical data reflects the clinician's choices, not the model's. Off-policy evaluation (inverse probability weighting, doubly robust estimators) reweights historical cases to estimate what would have happened under the model's policy. A model that predicts well under the current treatment policy can still recommend badly, because the outcomes it trained on were generated by different decisions, so if your model recommends an action, evaluate it off-policy before deployment.

SHAP: SHapley Additive exPlanations

A clinician handed a single risk score wants to know what drove it: why is *this* patient at 0.42 when the average patient sits at 0.18? SHAP answers by borrowing a fairness idea from cooperative game theory, dividing the prediction into one fair additive contribution per feature. It decomposes a single prediction into a base value ϕ_0 (the average prediction) plus one contribution ϕ_j per feature.

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^M \phi_j$$

If a patient's readmission risk is 0.42 against a 0.18 average, SHAP splits the 0.24 gap exactly: +0.08 prior readmission, +0.06 heart failure, and so on.

ANALOGY

SHAP is an itemized bill for risk. Instead of one mysterious total, each feature gets a line item you can inspect.

A 2025 meta-analysis of 62 studies found explainability is the single strongest predictor of clinician trust and adoption, outranking raw accuracy, and TreeSHAP computes exact values for tree ensembles fast enough for real-time use. The attributions are not fool-proof: they can mislead when features are strongly correlated, and a statistically valid explanation ("3 AM admission drives risk") that violates clinical intuition will be dismissed, rightly or not.



A SHAP contribution plot for one patient, showing the largest contributors only. Red bars push predicted risk up, green bars pull it down; summed over all features, the signed ϕ_j add to the base value ϕ_0 to reproduce this patient's prediction exactly.

LIME and the Limits of Explanation

LIME explains a prediction by fitting a simple linear model in the local neighborhood of that prediction, using perturbed copies of the input; it is model-agnostic but sacrifices SHAP's theoretical guarantees. **Watch out.** LIME explanations can be unstable, so querying the same patient twice can yield different explanations from the perturbation randomness alone, and in a clinical setting that instability destroys trust instantly.

Explainability Has Three Audiences, Not One

Explainability is not one deliverable but three, each with its own audience, the format that lands, and the failure mode that erodes trust.

Audience	Needs	Format that lands	Failure mode
Physician	Clinical coherence	SHAP + clinical narrative	"3 AM admission" as top driver erodes trust
Regulator	No protected-attribute reliance	Counterfactuals; subgroup metrics	Proxies (zip code, prior cost) for race/sex
Patient	Plain language	One or two reasons in prose	SHAP waterfall plot; jargon

The IBM Watson for Oncology Cautionary Tale

IBM invested over **\$5 billion** building Watson for Oncology, then sold Watson Health in 2022 for roughly \$1 billion. Concordance with expert tumor boards fell as low as **12 percent** for some cancers, and internal

documents cited "unsafe and incorrect" recommendations. Every lesson in this chapter was violated at once. It was trained on MSK physicians' *opinions*, not patient *outcomes*, so there was no survival data feeding back; it was validated only by concordance

("two wrong opinions that agree are still wrong"), never temporally or geographically, so global distribution shift went untested; and it was a black box, so every disagreement between Watson and a clinician was a dead end. The lesson is that physicians adopt a model only when they get transparency (they see the drivers), override authority (they can disagree and document why), and feedback loops (they see whether it was right over time). None of these depend on model architecture: a black-box net with SHAP, override tracking, and quarterly report cards earns more trust than a bare logistic regression.

Should You Build This?

Build it when you can ship the model with its deployment scaffolding: SHAP explanations a physician can

read, temporal and geographic validation instead of a random split, and a monitoring plan for distribution shift. When the clinical question is time-to-event, use censoring-aware survival models rather than a binary classifier that quietly discards incomplete follow-up. **Think twice when** the training data encodes expert opinion rather than patient outcomes, when validation stops at concordance, or when nobody has asked what happens after the environment shifts, the three counts on which Watson for Oncology failed at once. **Who bears the risk:** physicians are asked to stake their judgment on outputs they cannot inspect, patients inherit any unsafe recommendation that slips through, and the sponsoring organization carries the financial and reputational loss when the black box fails in public.

Quick Self-Test

1. Your gradient-boosted model beats logistic regression on AUC, but the deployment site demands coefficient-level transparency. What do you deploy, and why is that a deployment decision rather than a technical one?
2. Why does treating a competing death as "censoring" bias a cancer-recurrence estimate, and what does the Fine-Gray CIF fix?
3. A model scores AUC 0.89 on a random split. Why should that number not reassure you, and what validation do you run instead?
4. Name the three forms of distribution shift and one detection method for each.
5. In one sentence, what does the concordance index measure that a calibration plot does not?
6. Watson for Oncology reported high concordance with MSK oncologists. Why was that not evidence of accuracy?

If You Remember One Thing

A deployable clinical model must clear four bars: ensemble-grade accuracy, time-to-event honesty, robustness under shift, and an explanation a clinician can act on. A model can be excellent on the first and fail the rest, and the rest are what decide whether it helps a patient. Watson for Oncology failed all four, and the \$4 billion write-down made the cost of that failure visible.

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