

Supervised Learning and Clinical Prediction

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 7 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. A well-calibrated model that a nurse trusts and can act on is worth more than a sharper-ranking model that no one looks at.

PART 1: The Supervised Framework

The Learning Problem

Start with the question, not the equation. You have a pile of patient data, labs, vitals, codes, prior utilization, and an outcome you care about, say whether someone is readmitted within 30 days. Supervised learning is just the machinery for turning the first into a prediction of the second: it learns a mapping from the features to the outcome, tuned by a *loss function* that decides how badly each kind of mistake counts.

ANALOGY

A resident learning to triage: the features are what they observe, the target is what actually happened, and the loss function is how badly they get scolded for each kind of mistake.

In symbols, the model learns a function f that maps the feature matrix $\mathbf{X}$ onto the target y :

$$f(\mathbf{X}) \approx y$$

The loss is where the medicine lives, not the math. A symmetric loss says a missed sepsis case and a false alarm are equally bad, a claim that is almost never defensible, so the loss must encode the asymmetry, through **cost-sensitive learning** or **threshold adjustment**. (Two terms to keep straight: **classification** predicts a category, readmit or not, while **regression** predicts a number, like length of stay.)

Why Accuracy Lies

Thirty-day readmission runs roughly 12 to 20 percent, so a model that predicts “no readmission” for everyone is about 85 percent accurate and completely useless, a smoke detector that never goes off, right on almost every day of the year and worthless on the one that matters. Accuracy hides the rare positives that carry all the clinical weight. Judge imbalanced models on ranking (AUROC, AUPRC), on probability quality (calibration, Brier), and on net benefit, never on raw accuracy.

Resampling and Cost-Sensitive Learning

Two techniques make a model care about the rare class. **SMOTE** synthesizes new minority examples between existing ones; **cost-sensitive learning** leaves the data alone and raises the penalty for missing a positive (in scikit-learn, `class_weight`).

Watch out. SMOTE can invent implausible “patients” in feature space and distort calibration. Prefer cost-sensitive weights or threshold tuning when calibrated probabilities matter, and evaluate on the metrics above rather than on accuracy.

Before any formula, fix the picture every metric is built on: the four ways a single prediction can land.

		Predicted	
		-	+
Actual	+	FN missed costly	TP caught
	-	TN correct	FP false alarm

In healthcare the four cells almost never cost the same; the loss function must say so.

Discrimination vs. Calibration

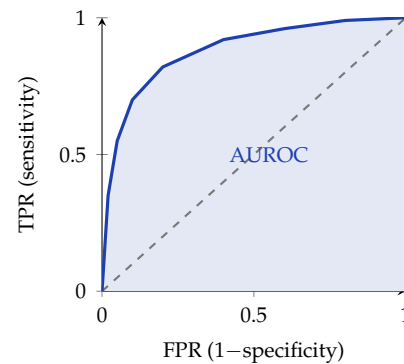
Two very different questions hide behind the word “accurate.” Discrimination is a **sorting** problem: does the model rank sicker patients higher? Calibration is a **measuring** problem: does “70 percent” actually happen 70 percent of the time? A model can ace one and fail the other, and a model that says 70 percent but is right only 30 percent makes clinicians over-intervene. Van Calster and colleagues (2019) argued calibration should take precedence over discrimination for clinical use.

AUROC (the c-statistic)

The practical question is simple: you have two models, which one ranks sicker patients higher? That is all AUROC measures, the probability that the model scores a random positive patient above a random negative one, with 0.5 a coin flip and 1.0 perfect ranking; clinicians call it the c-statistic.

$$\text{AUROC} = P(\hat{s}_{\text{positive}} > \hat{s}_{\text{negative}})$$

Good for comparing models during development, it is nearly useless on its own for deployment, because it says nothing about calibration, nor about performance at the one threshold you will actually use.



The shaded area is AUROC; the diagonal is chance.

The Brier Score

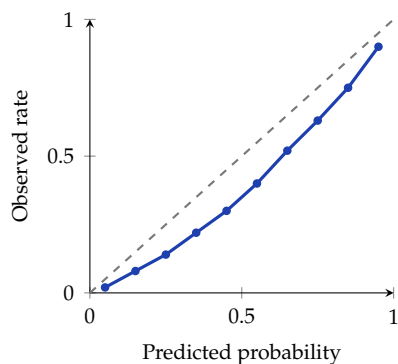
Sometimes you care whether the predicted probabilities are actually right, not just ranked right. The Brier score punishes both at once: it is the mean squared gap between predicted probability and outcome.

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2$$

Zero is perfect. But it is base-rate dependent: at 15 percent prevalence, always predicting 0.15 scores about 0.13 and “looks” good, so anchor it, since always predicting 0.5 scores 0.25. (The classic Murphy decomposition is *reliability, resolution, uncertainty*.)

Calibration and the Reliability Diagram

The way to see calibration is to draw it. Bin predictions into deciles and plot mean predicted probability against observed rate; the 45-degree line is perfect, bowing above it means underprediction, below means overprediction. Recalibrate with **Platt scaling** or **isotonic regression** on a held-out calibration set separate from both training and test. Skipping this is why models pass internal validation and fail at another hospital.



Below the 45° line: this model overpredicts risk.

PPV and the Bayes Trap

The question a clinician actually asks is not “how sensitive is it” but “if it flagged my patient, how likely is that real?” That is positive predictive value, and Bayes’ theorem is brutal about it.

$$\text{PPV} = \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec})(1 - \text{prev})}$$

At 90 percent sensitivity and 90 percent specificity on a 5 percent prevalence condition, PPV is only about **32 percent**: two of every three flags are wrong, purely from the base rate. This is why strong models still drown clinicians in false alarms.

ANALOGY

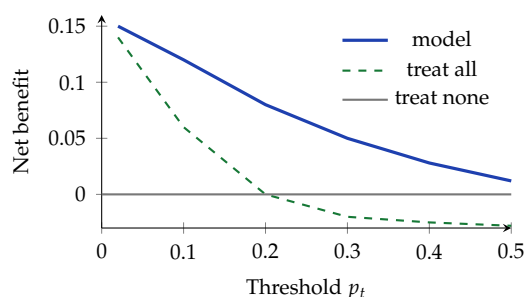
Airport screening for a rare disease: even a near-perfect scanner mostly flags healthy travelers, because almost everyone is healthy. The test didn’t fail, the base rate did.

Net Benefit and Decision Curves

None of the metrics above answers the only question that matters at deployment: is using this model better than doing nothing, or than treating everyone? Decision-curve analysis does, measured in true-positives-per-patient.

$$\text{Net Benefit} = \frac{\text{TP}}{N} - \frac{\text{FP}}{N} \cdot \frac{p_t}{1 - p_t}$$

The threshold p_t encodes the cost ratio (at 5 percent you tolerate 19 false positives per true positive; at 50 percent, one). Vickers and Elkin (2006) showed it answers the deployment question that discrimination and calibration cannot, without pricing harms in dollars. For readmission the clinically relevant window is roughly 15 to 40 percent, so report the decision curve and the reliability diagram to stakeholders and leave AUROC in the methods.



The model earns its keep only above both defaults.

Metric	Measures	Pitfall	Prioritize for
AUROC	Ranking	Blind to calibration; weak on rare events	Model comparison
AUPRC	Ranking on rare positives	Baseline shifts with prevalence	Rare-event detection
Brier	Calibration + discrimination	Base-rate dependent	Final deployment call
PPV	$P(\text{disease} \mid \text{flagged})$	Collapses at low prevalence	Talking to clinicians
Sensitivity	True-positive rate	Drives alert volume	Must-not-miss (sepsis, MI)
Net benefit	Clinical utility	Needs a threshold range	Deploy / do-not-deploy

PART 4: Why Models Get Ignored (Deployment Reality)

The Override Problem

A model can be right and still be useless if no one acts on it. About **90 percent** of clinical decision support alerts are overridden, and a 2022 Michigan Medicine study found only 16 percent of physicians considered ML sepsis predictions helpful. These are system-design failures, not model-quality failures. A model earns its place only if it is calibrated, arrives at the right moment in the workflow, and attaches to an action someone can take.

The Epic Sepsis Cautionary Tale

The clearest warning in this whole chapter is a real product. In external validation (Wong et al., 2021) the Epic Sepsis Model scored **AUROC 0.63** against the 0.76 to 0.83 claimed, alerted on 18 percent of patients, missed **67 percent** of sepsis, and when it fired the patient truly had sepsis only **12 percent** of the time, so a clinician following the alert would be wrong seven times out of eight. A confidently wrong model is worse than no model: it manufactures a false sense of safety and trains clinicians to ignore it. Vendor-run

internal validation is not enough; demand external, temporal validation.

Should You Build This?

Build it when the prediction feeds an intervention the institution can actually execute: a calibrated readmission score surfaced 24 to 48 hours before discharge, paired with a specific care pathway such as a pharmacist-led medication reconciliation call, in a hospital with real readmission-penalty exposure. A decision curve showing net benefit over treat-all and treat-none is the entry ticket. **Think twice when** the plan is a standalone risk score validated only on random splits: the Epic Sepsis Model shipped exactly that way, and no workflow, no recalibration plan, and no retraining schedule leaves you with noise on a dashboard. **Who bears the risk:** clinicians absorb the false alarms and learn to dismiss the alert that matters, patients absorb the missed sepsis cases and preventable readmissions, and the safety-net hospital, penalized for social factors it cannot control, absorbs the bill.

Quick Self-Test

1. Your model is 92 percent accurate on a 12 percent readmission rate. Why is that not reassuring, and what do you report instead?
2. What does a reliability diagram show that AUROC cannot?
3. Why must recalibration use a held-out set rather than the training or test set?
4. A model has 90 percent sensitivity and 90 percent specificity on a 5 percent condition. Roughly what is its PPV, and why?
5. In one sentence, what does decision-curve analysis add beyond calibration?

If You Remember One Thing

Discrimination, calibration, and net benefit are three different questions. A model can ace the first and fail the other two, and the other two are the ones that decide whether it helps a patient. The Epic Sepsis Model failed all three, and it still shipped.

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