

Calibration and Validation

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. Discrimination proves a model can sort patients; calibration proves you can trust the number it prints, and validation proves that number will still be true next quarter and at the next hospital, and only a model that clears all three earns a place in care.

PART 1: Two Different Questions a Number Must Answer

Discrimination vs. Calibration

Two very different questions hide behind the word "accurate." Discrimination is a sorting problem: does the model rank sicker patients above healthier ones? Calibration is a different problem entirely: when the model says 70 percent, does the event actually happen 70 percent of the time? A model can ace the first and fail the second, and the two failures reach the patient in opposite ways.

ANALOGY

A triage nurse who always puts the sickest patient at the front of the line has perfect discrimination, even if she cannot tell you whether that patient's risk is 30 percent or 80 percent. Calibration is knowing the actual number, not just the order.

Discrimination is the part clinicians already quote.

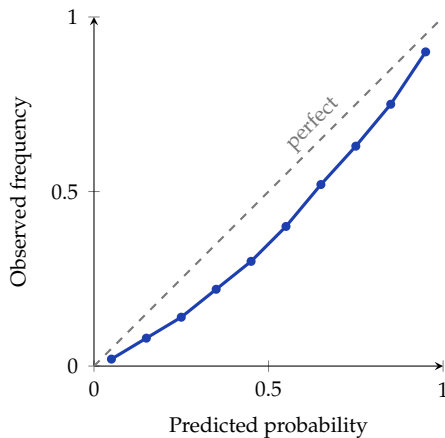
AUROC, the c-statistic, is the probability the model scores a random positive patient above a random negative one:

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

Van Calster and colleagues (2019, "Calibration: The Achilles Heel of Predictive Analytics," BMC Medicine) argued that calibration should take precedence over discrimination for clinical use, because a model that says 70 percent but is right only 30 percent of the time drives clinicians to over-intervene on people who were never that sick. AUROC is nearly useless on its own for deployment: it says nothing about whether the probabilities are correct, and nothing about performance at the single threshold you will actually use. Report it in the methods section, not to the people making decisions.

The Reliability Diagram

Sort every prediction, split into bins (usually 10 deciles), and plot each bin's mean predicted probability (x) against its observed event rate (y). Perfect calibration falls on the 45-degree diagonal; a curve bowing above the line means the model underpredicts (reality is worse than it says), and bowing below means it overpredicts (it cries wolf). It is a truth-telling audit for probabilities: line up every time the model said “about 30 percent” and check how often the event really happened, and if it was 30 percent, that point sits on the diagonal. The diagram shows what AUROC cannot, namely systematic over- or under-confidence at specific risk levels, which is exactly where a threshold gets set and a patient gets treated or sent home.



A reliability diagram: bin predictions by decile, plot mean predicted (x) against observed rate (y). This curve sits below the 45° line, so the model overpredicts risk at every level.

The Brier Score

Sometimes you want a single number that punishes a model for bad ranking and bad probabilities at once. The Brier score is that number: golf scoring for probabilities, where every prediction is penalized by how

far it lands from the truth and there is no way to post a good total without actually being close. Formally, it is the mean squared gap between predicted probability and outcome.

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

It is a proper scoring rule, rewarding accurate ranking and accurate probabilities together; zero is perfect, and random prediction under class balance scores 0.25. **Watch out.** It is base-rate dependent: at 15 percent prevalence, always predicting 0.15 scores about 0.13 and looks impressive while saying nothing. Anchor it before you believe it, since always predicting 0.5 scores 0.25.

Expected Calibration Error

A reliability diagram is a photograph of miscalibration; sometimes you need its one-line caption, a single number you can trend on a dashboard month over month. Expected calibration error is that caption. It collapses the whole diagram by taking, for each bin B_m , the gap between observed accuracy and mean predicted confidence, then averaging those gaps weighted by how many patients fall in each bin.

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} | \text{acc}(B_m) - \text{conf}(B_m) |$$

An ECE of 0.08 means predictions are off by about 8 percentage points on average. It is useful for tracking one number over time, though it can hide a large error in a sparse bin. On an MLOps dashboard (Chapter 25) a rising ECE is often the first quantitative sign a deployed model is drifting, before anyone notices harm.

Van Calster's calibration hierarchy

Level	Question it answers	Tool
In-the-large	Does mean predicted match overall observed rate?	Compare the two means
Weak	Is the calibration slope 1 and intercept 0?	Logistic recalibration fit
Moderate	Do observed rates track predictions across the range?	Reliability diagram, ECE
Strong	Correct for every covariate pattern?	Essentially unattainable

A calibration slope of 1.0 is perfect; a slope below 1 is the fingerprint of an overfit model (extreme predictions too extreme). The Hosmer-Lemeshow test is the classic goodness-of-fit companion.

PART 3: Fixing Calibration After the Fact

Platt Scaling and Isotonic Regression

Miscalibration is repairable without retraining the whole model. **Platt scaling** fits a logistic regression on the model's raw scores, squeezing them toward truth with a smooth S-curve. **Isotonic regression** fits a non-parametric monotonic step function, more flexible but hungrier for data. Both preserve the ranking (AUROC is unchanged); they only relabel the probabilities. A 2025 MIMIC-IV study found random-forest readmission models achieve good, generalizable calibration on internal and external data precisely when these techniques are applied, and skipping recalibration is a leading reason models pass internal validation and then fail at another hospital.

ANALOGY

The model is a thermometer that reads consistently five degrees high. You do not throw it out; you paste a corrected scale over the dial. Recalibration is that corrected scale.

Watch out. Fit the recalibration on a held-out calibration set that is separate from both training and test. Reuse the training set and you simply relearn the same overconfidence; reuse the test set and you have no honest estimate left. Platt suits small, roughly logistic errors; isotonic needs more data but corrects arbitrary monotonic distortion.

Train 60%

fit the model

Calib 20%

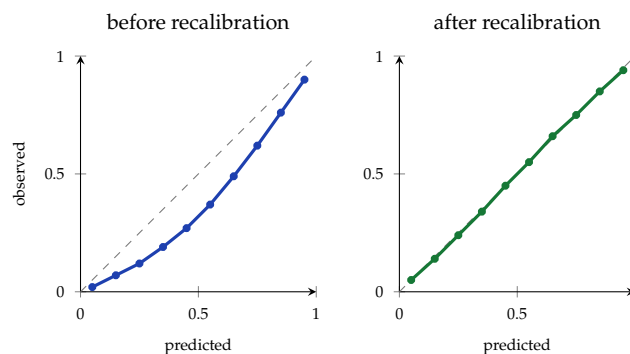
fit calibrator

Test 20%

held out

Recalibration needs its own slice. Fit the model on train, fit the calibration map on a separate calibration set, and keep test untouched, or the honest estimate is gone.

Method	Fits	Strength	Fails when
Platt scaling	Logistic curve on raw scores	Stable on small samples	Distortion is not S-shaped
Isotonic regression	Monotonic step function	Corrects any monotone error	Overfits on little data



Recalibration relabels probabilities without touching the ranking (AUROC is unchanged): the overconfident curve on the left is pulled onto the diagonal on the right.

Temporal and External Validation

Not all validation is equally honest, and the ranking of regimes is really a ranking of how hard each one makes it for the model to cheat. A random train-test split makes the test set statistically exchangeable with training, so it flatters the model. Temporal validation trains on the past and tests on a later period; external (geographic) validation tests at a different site; prospective validation runs live. Each step up removes a way the model could be quietly cheating, the way changing the exam and changing the room stops a student from practicing on a shuffled photocopy of the same test.

random split < cross-validation
< temporal < geographic < prospective

A 2024 JAMA Network Open deterioration model scored AUC 0.89 on a random split but **0.82** under temporal validation, real degradation the random split would never expose; temporal validation is the minimum before deployment, and geographic validation is strongly advised for any multi-site use.

Watch out. Distribution shift is the expected condition, not an edge case. Wynants and colleagues reviewed 731 COVID-era prediction models and rated every one at high risk of bias; well-validated pre-pandemic models broke overnight. Validate for the world you will deploy into, not the one you trained in.

The three shifts, and how to catch them

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

Concept drift is the most insidious: inputs look unchanged, yet a once-calibrated model is now wrong.

Decision-Curve Analysis

Calibration and discrimination tell you the model is well built; neither answers the only question that matters at deployment: is using this model better than doing nothing, or than treating everyone? Decision-curve analysis answers exactly that, measured in true-positives-per-patient, and the threshold p_t encodes the cost ratio, so at $p_t = 5$ percent you tolerate 19 false positives per true positive and at 50 percent only one.

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

Vickers and Elkin (2006) showed it does not require pricing harms in dollars, only naming the threshold probability at which you would act, and for readmission programs the clinically relevant window is roughly 15 to 40 percent. Report the decision curve and the reliability diagram to stakeholders; leave AUROC in the appendix. Net benefit needs a threshold range, not a single point, and that range is a clinical judgment about how bad a false positive is versus a miss, because a model can have the best AUROC in the room and still show no net benefit anywhere in the range that matters.

Drift Detection and the Monitor to Retrain Ladder

A model at 0.92 AUC on Tuesday can hit 0.78 on Thursday when the case mix shifts (flu season, an EHR coding change, a new bundle). Production systems watch two things: input drift, via the Population Stability Index, Kolmogorov-Smirnov, and Jensen-Shannon divergence on feature distributions; and performance drift, via rolling AUC, ECE, subgroup sensitivity, and PPV as labels mature. This is pharmacovigilance for algorithms: a deployed model must be watched for drift the way an approved drug is watched for adverse events. The ladder has three rungs, in order. **Monitor** continuously. When calibration slips, **recalibrate** (a cheap Platt or isotonic refit on fresh data). Only when both distribution and performance have moved do you **retrain**, and a retrained model never auto-deploys: it passes back through validation and governance gates first. Mayo, for example, requires a 90-day shadow period before a model touches decisions. The “zombie algorithm”

problem is real: between 2022 and 2025, 60 authorized AI devices were tied to 182 recall events, **43 percent** within a year of authorization. A model that is never monitored is not stable; it is unwatched.

Reporting Discipline: The Model Card

Before a model reaches a patient, a model card discloses intended use, training-data demographics, performance stratified by subgroup, known failure modes, and regulatory status. An aggregate number hides what matters: a model that is 92 percent accurate overall but 74 percent for Black patients is not a 92 percent model. The Epic Sepsis Model is the cautionary tale: external validation (Wong et al., 2021) found AUROC 0.63 versus the 0.76 to 0.83 claimed, alerts on 18 percent of patients, **67 percent** of sepsis missed, and a true-positive only 12 percent of the times it fired. It failed discrimination, calibration, and net benefit at once, and it still shipped, because no one demanded external validation or an honest model card.

Quick Self-Test

1. A model has AUROC 0.88 but a reliability diagram bowing far below the diagonal. What is wrong, what will clinicians do, and which metric caught it?
2. Why must Platt scaling or isotonic regression be fit on a held-out calibration set rather than the training or test set?
3. Your model scores AUC 0.89 on a random split. Why should that not reassure you, and what is the minimum validation you run before deployment?
4. Name the three forms of distribution shift and one detection method for each. Which is the hardest to see, and why?
5. In one sentence, what does decision-curve analysis add that calibration and discrimination cannot?
6. Describe the monitor-to-retrain ladder, say where recalibration sits, and why detected drift never auto-redeploys.

If You Remember One Thing

Discrimination, calibration, and net benefit are three separate questions, and validation asks whether the answers survive time and place. A model can top the AUROC chart yet be miscalibrated, useless at every clinical threshold, and broken by next quarter’s case mix. The Epic Sepsis Model failed all four at once, quietly, for years, because its owners trusted one number instead of measuring all four. Trust a number only after you have made it earn it.

