Topic 1: Validity and Prediction Evaluation

Six Ideas in AI Evaluation, First Term 2026

Yiqun T. Chen

1 A study, before we see any results

The running case: AEGIS at Riverbend Health. Riverbend Health, a four-hospital system, wants to know whether physicians' use of AI tools has improved patient care. Over the past two years some clinicians have started using assistants for note drafting, literature search, and diagnostic support. Leadership wants to know whether it is working before deciding what to buy next.

They have two kinds of data. Patient biomarkers are collected routinely and indicate health risk, and patients fill out satisfaction surveys after visits. The administration recruited study participants and enrolled 20 physicians who use AI tools and 20 who do not. They will compare the biomarkers and satisfaction scores of the two groups' patients.

Before we look at a single number, what do you think of this study?

No statistical analysis can repair a design that does not support the intended comparison. The key questions arise before the first outcome is observed.

2 Validity

Validity is the degree to which evidence and theory support the intended interpretation of a measurement. It is not a property of a number by itself; it concerns the inference from a number to a claim. The same satisfaction score may be a valid measure for one purpose and an invalid one for another.

In a stronger formulation, an evaluation is valid for a property when the property exists and variation in that property causes variation in the score. The property must be real, and it must be the reason the measure changes.

Limits of empirical evidence for validity

You cannot read validity off the data. Different explanations of what produces your measurements can imply exactly the same observed numbers, so the underlying structure is not identified. Validation is therefore a process of collecting evidence and making an argument, and the strongest evidence usually comes from decisions made before you collect data rather than from analysis afterwards.

We will use four forms of validity to examine Riverbend's study.

2.1 Construct validity

Patient care outcome is not something you can observe. It is a construct, meaning a quantity we believe exists and want to reason about but cannot measure directly. To study it we have to pick something we can measure, and Riverbend has several candidates.

- **Biomarkers.** They are collected routinely, they are cheap, and they are comparable across sites. A biomarker panel measures physiological risk rather than care quality, though, and it moves slowly. A physician who spends more time with a patient may change very little in an HbA1c over the study window.
- **Specific symptoms.** Symptoms are closer to what patients experience, but recording them depends on documentation, and documentation is exactly what an AI note-drafting tool changes. If the assistant writes more thorough notes, recorded symptom counts go up for reasons that have nothing to do with health.
- **Patient-reported satisfaction.** Satisfaction captures something patients care about, and it is the outcome most sensitive to the intervention. It is also sensitive to wait times, parking, and whether the physician made eye contact. In some settings satisfaction goes up when physicians order more tests and prescribe more freely, which is not better care.
- **Physician diagnosis.** A recorded diagnosis is close to what a diagnostic assistant should improve, but the physician being evaluated is the one who records it. If the AI suggested the diagnosis, the measure partly records the AI's output rather than checking it.

These are not interchangeable, noisy readings of a single quantity. They capture different aspects of care and may move in opposite directions: a tool that shortens visits could raise satisfaction while worsening biomarkers. Selecting an outcome is therefore a substantive claim about what better care means. State and defend that claim before the study, rather than selecting the outcome that changes.

Two kinds of evidence apply here. Content validity asks whether your chosen measures cover the whole construct, and satisfaction alone does not cover patient care outcomes. Convergent and discriminant validity ask whether measures of the same construct agree with each other and measures of different constructs do not. If satisfaction and biomarkers are uncorrelated in Riverbend's own historical data, then at most one of them measures what leadership means.

Takeaway. Ask what the construct is, then ask what would count as evidence that your measure tracks it. If the answer is that nothing you could check would count, you have a definition problem rather than a statistics problem.

2.2 External validity

The study includes 40 physicians from one health system who agreed to participate. Even with an otherwise sound design, its result would directly describe this group, not physicians in general.

The physicians volunteered, and clinicians who sign up for an AI evaluation study are more interested in the technology, probably younger, and probably working in services where adoption is easier. The four hospitals are also unlike each other, since an academic center and a rural critical access hospital have different patients, different case mix, and different documentation habits, so a result pooled across them may describe none of them. The tools themselves will be replaced during the study window, so a finding about AI use in 2025 may not describe the assistants available in 2027.

External validity asks whether a result transfers to the population and conditions of interest. We return to its site-to-site version in Topic 6.

2.3 Internal validity

This is the design's central weakness.

Physicians chose whether to use AI, because nobody assigned them. The two groups therefore differ in everything that predicts adoption, including age, training, subspecialty, comfort with technology, panel size, and almost certainly baseline quality and patient population. Any difference in outcomes mixes the effect of the tools with the effect of being the kind of physician who adopts them, and the design gives you no way to separate the two. Adjusting for the covariates you happen to have does not fix it.

The direction of causation is also unsettled, since physicians whose practices already run efficiently may have the slack to try new tools, rather than the tools creating the slack. Patients are not randomly assigned to physicians either, because referral patterns, insurance, and geography determine who sees whom, so the two groups of physicians see different patients.

The unit of analysis is ambiguous as well. Outcomes are measured on patients, but the exposure is at the physician level, and patients within one physician's panel are correlated. Treating patients as independent will badly overstate precision, and the effective sample size is closer to 40 than to the thousands of patient records. Forty clusters is a small study for detecting a modest effect on a noisy outcome even when the design is otherwise sound.

Takeaway. The strongest evidence would come from randomizing physicians, or clinics, or the order in which sites switch on the tools. Where you cannot randomize, the study should say which comparison it makes and what would have to be true for that comparison to be causal. What it should not do is report a difference in satisfaction and call it an effect.

2.4 Consequential validity

Suppose the study runs, satisfaction is the primary outcome, and satisfaction goes up. Leadership then adopts patient satisfaction as the metric for AI deployment, and later for evaluating physicians.

Satisfaction is now a target. A physician who wants to score well can order more tests, prescribe more freely, and avoid difficult conversations about smoking, weight, or whether the patient is taking their medication. Each of those raises satisfaction, none of them improves care, and some of them make care worse. Within a couple of years the correlation between satisfaction and care quality is weaker than it was, because people have optimized against the measure.

Goodhart's law describes what happened, and consequential validity is the kind of evidence that

asks about it. It asks whether the real-world results of using scores this way line up with what you intended. It is the category people leave out most often, and it is the only one that asks what the measurement does to the world rather than what the world does to the measurement. The same cycle runs on benchmarks, where it is fast enough to watch.

3 From studies to predictions

We now narrow the focus from a study to a prediction system. Riverbend is considering a commercial tool called AEGIS, sold by Corvus Health. It reads the chart hourly and returns a sepsis risk score for every admitted patient, and a second component reads the notes and summarizes findings for the clinician. AEGIS is the system the rest of this course evaluates, and every later topic returns to it. The validity questions still apply, because a prediction target is also a construct; but the system's output gives us a specific object to evaluate.

Consider three predictions that are deliberately unalike.

- **Will this patient be in remission at six months?** A yes or no answer about a future event that has not yet been determined.
- **Is this image a cat or a dog?** A yes or no answer about a fact that is already settled, so the uncertainty is in the model rather than in the world.
- **What will this patient's HbA1c be at six months?** A number rather than a label.

How would you evaluate each one? All three are prediction problems, and the natural answers differ more than you might expect.

4 Binary outcomes

Write $Y \in \{0, 1\}$ for the ground truth and $\hat{Y} \in \{0, 1\}$ for the prediction. The simplest loss is the zero-one loss, $\mathbf{1}\{\hat{Y} \neq Y\}$, whose average is the error rate. One minus the error rate is accuracy.

Accuracy is the right summary when the classes are roughly balanced and the two kinds of mistake cost about the same. Both hold for cats and dogs. Neither holds for most clinical problems.

Calculation 4.1 (Accuracy compares the system to nothing)**.** Sepsis occurs in 2% of Riverbend admissions, so a system that always says no is 98.0% accurate. Now take a real model with sensitivity 0.60 and specificity 0.98, per 10,000 admissions.

	Sepsis (200)	No sepsis (9800)
Predict $\hat{Y} = 1$	TP = 120	FP = 196
Predict $\hat{Y} = 0$	FN = 80	TN = 9604

Accuracy is $9724/10000 = 97.2\%$, which is lower than always saying no, even though the model is far more useful. A summary that ranks a useful model below a useless one is not measuring what we want.

One response is to report accuracy separately within each class. For binary outcomes, these quantities have standard names:

$$\text{sensitivity (recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 0.60, \qquad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = 0.98.$$

You want class-wise accuracy whenever the classes are imbalanced or the two errors cost different amounts, which covers most cases where a decision follows the prediction. Screening for a rare cancer is one example, where missing a case is far more expensive than a false alarm, and a single accuracy figure averages the two into something meaningless.

Sensitivity and specificity both condition on the truth, but the person acting on the prediction conditions the other way. A nurse sees an alert and wants to know how likely this patient is to be septic given that the alert fired, which is precision, also called positive predictive value.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{120}{316} = 0.380.$$

Report precision and recall together, because each one alone is easy to game. Predict positive for everyone and recall is 1. Predict positive only for the single most obvious case and precision is 1.

Precision and prevalence. By Bayes' rule, $\text{PPV} = \pi\,\text{TPR} \,/\, \big(\pi\,\text{TPR} + (1-\pi)\,\text{FPR}\big)$, where π is the prevalence. Hold the model fixed at sensitivity 0.80 and specificity 0.95, and change only the population.

Prevalence π	10%	1%	0.1%
PPV	0.640	0.139	0.016
False alerts per true alert	0.6	6.2	62.4

The model, its sensitivity, and its specificity are the same in all three columns. At a prevalence of 0.1% a clinician chases 62 false alarms for every real case, and staff will stop reading the alerts within a month. Report false alerts per true alert, which is $(1/\text{PPV}) - 1$, because that is what users experience.

5 Why predict a probability

Thus far the system has returned a label. Many useful systems instead return a number between 0 and 1:

- **The world is not determined.** Whether a patient reaches remission is not a fact waiting to be looked up, and two similar patients can have different outcomes. A model forced to output 0 or 1 has to pretend otherwise, while a probability describes the uncertainty directly.
- **It separates the prediction from the decision.** A label has a threshold built into it, and therefore a cost ratio, chosen by whoever trained the model rather than by whoever lives with the consequences. With a probability, the ICU and the general ward can act on the same score at different thresholds.
- **It supports triage.** If you can follow up thirty patients today, you want them ordered by risk rather than split into two piles.

- **You can change the threshold later.** Staffing changes, prevalence changes, and a new therapy changes the cost of missing a case. With probabilities you move the threshold, and with labels you retrain.
- **It combines with other information.** You can combine a probability with costs, with another model, or with a clinician's own judgment, and you cannot do that with a label.

This flexibility creates an evaluation problem: how should we compare a forecast between 0 and 1 with an outcome that is either 0 or 1?

6 Ways to score a probability

Write $s = f(X)$ for the predicted probability. Summaries in common use include the following.

- Threshold at 0.5 and compute accuracy, or sensitivity and specificity at a chosen operating point.
- Brier score, $(s - Y)^2$, which is squared error on the probability scale.
- Log loss, $-Y \log s - (1 - Y) \log(1 - s)$.
- AUC, the probability that a random case is ranked above a random control.
- Average precision, or the area under the precision and recall curve.
- Calibration error, meaning some measure of whether the numbers mean what they say.

They capture different properties. Choosing among them requires deciding which properties matter for the application. The following example isolates one important distinction.

6.1 Two systems you can tell apart, and a summary that cannot

Calculation 6.1 (Ten patients, two systems)**.** Ten patients are evaluated for remission at six months. System A and System B score them as follows, and Y records what actually happened.

Patient	A says	B says	Y	$(s_A - Y)^2$	$(s_B - Y)^2$
1	0.9	0.6	1	0.01	0.16
2	0.9	0.6	1	0.01	0.16
3	0.9	0.6	1	0.01	0.16
4	0.9	0.6	0	0.81	0.36
5	0.9	0.6	0	0.81	0.36
6	0.1	0.4	1	0.81	0.36
7	0.1	0.4	1	0.81	0.36
8	0.1	0.4	0	0.01	0.16
9	0.1	0.4	0	0.01	0.16
10	0.1	0.4	0	0.01	0.16
			mean	**0.330**	**0.240**

Threshold either system at 0.5. Both call patients 1 to 5 positive and patients 6 to 10 negative, so both get patients 1, 2, 3, 8, 9, and 10 right and the other four wrong. Both are 60% accurate. Both also rank the first five patients above the last five, so they have the same AUC.

Look at what actually happened in each block. Three of the first five patients reached remission, so the rate in that group is 0.6, and System B said 0.6. Two of the last five reached remission, so that rate is 0.4, and System B said 0.4. System A said 0.9 and 0.1 for the same two groups.

The Brier scores are 0.330 for A and 0.240 for B.

The systems are not equally useful. System A tells a patient in the first group that remission is a 90% prospect, although the observed rate for comparable patients is three in five; that claim could reasonably affect a treatment decision. Accuracy misses this difference, as does AUC, because AUC uses only the ordering of scores. The Brier score distinguishes them because it also reflects calibration.

7 Calibration

A system is calibrated if, among all the cases it scored v, the outcome happens a fraction v of the time.

$$\mathbb{E}[\,Y \mid f(X) = v\,] = v \qquad \text{for every value } v \text{ the system reports.}$$

This definition is empirically testable. It requires only an average among cases assigned a given score; it does not require access to any individual patient's unobservable true probability.

Several increasingly demanding versions are useful:

- **Calibration in the large.** The average prediction equals the overall event rate. It is one number, and errors in opposite directions cancel out.
- **Weak calibration.** Regressing Y on $\operatorname{logit}(f(X))$ gives a slope of 1 and an intercept of 0. A slope below 1 is the signature of an overfit model, and System A above is exactly that.
- **Moderate calibration.** The whole curve $v \mapsto \mathbb{E}[Y \mid f(X) = v]$ lies on the diagonal, which is the definition given above.
- **Group-conditional calibration.** The same holds within each of a set of groups you name in advance. It is necessary, much harder, and the subject of a later week.

Calibration does not establish discrimination

Calibration does not mean the model is useful. If half of patients reach remission, a system that always says 0.50 is perfectly calibrated and tells you nothing about any individual. A system that says 0.9 for the patients who reach remission and 0.1 for those who do not is also calibrated and is far better. Report calibration next to a measure of discrimination, never instead of one.

Calibration can also fail in a way that no pooled summary will show you.

Calculation 7.1 (A calibrated system that is wrong for everybody in it)**.** AEGIS reports a risk of 0.30 for a set of 1,000 admissions, and 300 of them become septic. The score is exactly right, and at that value the system is calibrated by the definition above.

Now split the same 1,000 patients by the language their chart was documented in.

	Patients	Events	Observed rate
English-documented	800	200	0.25
Spanish-documented	200	100	0.50
All	1,000	300	0.30

The reported 0.30 is too high for one group and too low for the other, and because the first group is four times the size of the second, the two errors cancel exactly. A clinician acting at a threshold of 0.35 works up neither group, although half of the second one is septic.

The pooled calibration curve cannot show this, because the pooled curve is the weighted average that hid it. Group-conditional calibration is the version that would have caught it, and it requires naming the groups in advance. We take that up in Topic 5.

7.1 Measuring calibration

The usual diagnostic is a calibration curve, or reliability diagram, which plots mean prediction against the observed event rate. It can be estimated by binning scores, isotonic regression, or a smooth fit on the logit scale; the latter can also provide a confidence band.

The standard scalar is the expected calibration error, $\mathrm{ECE} = \mathbb{E}\big|f(X) - \mathbb{E}[Y \mid f(X)]\big|$, usually estimated by binning and averaging $|\bar{s}_b - \bar{Y}_b|$ across bins. It misbehaves in two ways at once.

Sampling noise pushes it up. Within a bin holding n_b cases, the observed rate differs from the mean prediction by noise of order $n_b^{-1/2}$, and taking the absolute value turns that noise into a positive contribution. A perfectly calibrated system therefore has a positive expected estimated ECE, and the estimate grows as you add bins.

Binning pushes it down. Within a bin, over-prediction and under-prediction cancel before you take the absolute value, so coarse bins hide real miscalibration.

No choice of bin count fixes both, so an ECE reported without its bin count and sample size cannot be interpreted, and ECE values from different papers are not comparable. Prefer the plotted curve. If you need a number, report the calibration intercept and slope, whose standard errors you can interpret.

8 Proper scoring rules

We also want a score that rewards calibrated forecasts and cannot be improved by strategic reporting. A scoring rule $S(s, Y)$ is proper if a forecaster minimizes expected loss by reporting their genuine belief, and strictly proper if that report is uniquely optimal.

Proposition 8.1 (Strict propriety of the Brier score)**.** *If $Y \sim \mathrm{Bernoulli}(q)$, then the expected Brier score of a reported probability s is uniquely minimized at $s = q$.*

Proof. If the true probability is q, then

$$\mathbb{E}\big[(s-Y)^2\big] \;=\; (s-q)^2 + q(1-q),$$

which is uniquely minimized when $s = q$. □

Log loss is also strictly proper.

Accuracy is not proper. A forecaster who believes the risk is 0.40 and one who reports 0.001 are both classified as negative at a threshold of 0.5, and both are right 60% of the time, so accuracy cannot tell an honest forecast from a badly overconfident one. The F_1 score is not proper either, and it is also asymmetric between the classes and not comparable across studies with different prevalence.

Between the two proper rules, the Brier score is bounded and stable in small samples, while log loss is unbounded, so one confident mistake dominates the average, and it is undefined if the system ever reports exactly 0 or 1. Use the Brier score by default for evaluation reports on modest samples, and use log loss when you care about the extremes of the probability scale.

8.1 What a proper score is made of

The Brier score splits into three parts, and the split explains the ten-patient example. Write $\rho(v) = \mathbb{E}[Y \mid f(X) = v]$ for the true rate among cases scored v, and π for the overall rate.

$$\underbrace{\mathbb{E}(f(X)-Y)^2}_{\text{Brier}} = \underbrace{\mathbb{E}\big(f(X)-\rho(f(X))\big)^2}_{\text{miscalibration}} - \underbrace{\mathbb{E}\big(\rho(f(X))-\pi\big)^2}_{\text{resolution}} + \underbrace{\pi(1-\pi)}_{\text{intrinsic uncertainty}} .$$

Calculation 8.1 (Checking the split on Systems A and B)**.** Half the ten patients reached remission, so $\pi = 0.5$ and the intrinsic uncertainty is 0.25. Both systems split the patients the same way, into groups whose true rates are 0.6 and 0.4, so both have resolution $0.5(0.1)^2 + 0.5(0.1)^2 = 0.01$.

System B is calibrated, so its miscalibration is 0, and its Brier score is $0 - 0.01 + 0.25 = 0.240$, matching the table.

System A has miscalibration $0.5(0.9-0.6)^2 + 0.5(0.1-0.4)^2 = 0.09$, and its Brier score is $0.09 - 0.01 + 0.25 = 0.330$, which also matches.

All of the 0.09 gap between the two systems is miscalibration. Their resolution, which is the actual information content and the part that required a good model, is identical.

The terms answer distinct questions. Miscalibration can often be corrected after fitting; resolution is the useful variation in risk and requires a more informative model. Intrinsic uncertainty is a property of the population, not the system. Raw Brier scores are therefore not comparable across sites with different prevalence. Instead report a skill score, $1 - \text{Brier}/\text{Brier}_{\text{base}}$, relative to each site's always-predict-π baseline.

9 Recalibration, and the ranking scores

Recalibration estimates the map $v \mapsto \hat{\rho}(v)$ on held-out data, using either a logistic model of the score or isotonic regression, and then reports $\hat{\rho}(f(X))$ in place of $f(X)$. It removes the miscalibration

term while preserving resolution and intrinsic uncertainty, so it cannot worsen the Brier score on the population to which it applies. It requires a few hundred observed outcomes and leaves the underlying model unchanged; in the example, it maps System A to System B.

Two ranking-based summaries stay in wide use, so it helps to state what each one does.

AUC equals the probability that a random case is ranked above a random control, which is what you want when you are triaging a worklist of fixed size. It depends only on the ordering of scores, so multiplying every prediction by 0.1, squaring it, or replacing it with its rank leaves AUC unchanged. AUC is therefore blind to exactly what calibration measures, which is why it could not separate System A from System B. AUC also conditions on Y, so prevalence does not change it, which makes it transfer between populations more stably than precision does and makes it unable to tell you the alert burden.

Average precision does respond to prevalence, because it is built from precision. Its baseline is π rather than 0.5, so an average precision of 0.20 at a prevalence of 1% is a twenty-fold enrichment and may be excellent, while the same number at a prevalence of 30% is worse than guessing.

9.1 From predictions to decisions

A prediction is not a recommendation. A prediction summarizes uncertainty about an outcome; a decision chooses an action. Decision theory makes the missing ingredients explicit: an action set $\mathcal{A}$, a loss $L(a, Y)$ for each action–outcome pair, and a rule that chooses the action with smallest conditional expected loss,

$$a^*(x) \in \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}\big[L(a, Y) \mid X = x\big].$$

There is therefore no universally correct threshold. The appropriate action depends on the consequences of acting, the consequences of not acting, and the alternatives available to the clinician.

For a binary action—intervene or do not intervene—suppose an unnecessary intervention has loss $c_{\mathrm{FP}} > 0$ and a missed event has loss $c_{\mathrm{FN}} > 0$; give correct decisions loss zero. If s is the patient's event probability, the expected-loss-minimizing rule is to intervene when

$$s > \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}.$$

To see why, acting has conditional expected loss $(1 - s)c_{\mathrm{FP}}$, while not acting has loss sc_{FN}. The threshold is simply where those two quantities are equal. An ICU may therefore use a lower threshold than a general ward for the same predicted risk: missing a case has a different consequence, and the two settings may have different capacity to absorb false alerts.

This also clarifies what to evaluate. For a fixed threshold, the relevant score must be calibrated in the deployment population; otherwise the threshold no longer represents the stated tradeoff. AUC alone cannot justify a decision rule, because it contains no information about the risk attached to a score. For a capacity-constrained task—for example, calling only the thirty patients who can be followed up today— ranking matters: with equal intervention benefits and costs, the highest-risk patients should be first. In practice, report both calibration and discrimination, then evaluate the proposed action rule over the range of thresholds decision-makers would actually consider.

Decision thresholds and net benefit. Neither ranking score determines whether action is worthwhile. At a threshold t, the implied cost ratio is $c_{\mathrm{FP}}/c_{\mathrm{FN}} = t/(1 - t)$. Making that tradeoff

explicit gives net benefit,

$$\mathrm{NB}(t) = \frac{\mathrm{TP}(t)}{n} - \frac{t}{1-t} \cdot \frac{\mathrm{FP}(t)}{n},$$

which you plot against t alongside treating everyone and treating no one. Models with respectable AUCs often turn out to be worse than treating everyone across the whole plausible range of t, and no discrimination measure would ever show that.

10 A framework for other outcome types

The same structure applies beyond binary outcomes. An evaluation specifies a prediction format (a label, probability, or distribution), a scoring rule that compares prediction with outcome, and an aggregation across cases. The outcome type determines the first two components. The table summarizes the cases that remain.

Outcome	Prediction	Proper score	Calibration check
Binary, $Y \in \{0, 1\}$	$p \in [0, 1]$	Brier, log loss	reliability diagram
Ordinal, $Y \in \{1, \ldots, K\}$	$(p_1, \ldots, p_K)$	ranked probability score	$K - 1$ reliability diagrams
Continuous, $Y \in \mathbb{R}$	a distribution $\hat{F}$	CRPS	PIT histogram, coverage
Preference over i, j	$\mathbb{P}(i \succ j)$	Brier, log loss	reliability on comparisons

10.1 Ordinal outcomes

Many clinical outcomes are ordered categories, such as cancer stage I to IV, a four-level severity grade, or a five-point pain scale. Write $Y \in \{1, \ldots, K\}$ where the labels have an order but the gaps between them have no fixed size.

Accuracy is wrong here for a reason that has nothing to do with imbalance, which is that it treats every error as equally bad. Predicting stage II when the truth is stage IV is a worse mistake than predicting stage III, and accuracy scores them the same.

For a point prediction $\hat{Y}$, the fix is a loss that grows with the distance between the prediction and the truth, such as $|\hat{Y} - Y|$ or $(\hat{Y} - Y)^2$. The version used in medical grading is quadratic weighted kappa, which applies those weights and then corrects for the agreement you would get by chance,

$$\kappa_w \;=\; 1 \;-\; \frac{\sum_{a,b} w_{ab}\, O_{ab}}{\sum_{a,b} w_{ab}\, E_{ab}}, \qquad w_{ab} = \frac{(a-b)^2}{(K-1)^2},$$

where O_{ab} is the proportion of cases with true grade a and predicted grade b, and E_{ab} is what that proportion would be if the two were independent with the same margins. The chance correction is useful and it has a cost, because E_{ab} depends on how often each grade occurs, so two studies with different grade distributions produce kappas you cannot compare.

For a probabilistic prediction you report $(p_1, \ldots, p_K)$ with $\sum_k p_k = 1$, and you score the cumulative probabilities $\hat{F}(k) = \sum_{a \le k} p_a$ rather than the individual ones. The ranked probability score applies

the Brier idea at every cut point,

$$\text{RPS} \;=\; \sum_{k=1}^{K-1} \Big(\hat{F}(k) - \mathbf{1}\{Y \le k\} \Big)^2,$$

often divided by $K-1$ so that scores from scales of different lengths can be compared. It is strictly proper, and it charges you more for putting probability far from the truth than for putting it nearby, which is what having an order means. When $K = 2$ the sum has one term, $\hat{F}(1) = p_1$ and $\mathbf{1}\{Y \le 1\} = \mathbf{1}\{Y = 1\}$, so the ranked probability score is exactly the Brier score.

To produce ordinal probabilities in the first place, the standard model is the cumulative logit, also called proportional odds,

$$\operatorname{logit} \mathbb{P}(Y \le k \mid X = x) \;=\; \alpha_k - x^\top \beta, \qquad \alpha_1 \le \alpha_2 \le \cdots \le \alpha_{K-1},$$

which gives each cut point its own intercept and shares one slope across all of them. Calibration checks follow the same pattern. Because the score is built from $K-1$ cumulative probabilities, you check $K-1$ reliability diagrams, one for each cut point, asking in each case whether the outcome falls at or below grade k as often as the model said it would.

10.2 Continuous outcomes

For the HbA1c question the prediction is a number, $Y \in \mathbb{R}$. The usual losses for a point prediction $\hat{y}$ are squared error $(Y - \hat{y})^2$, absolute error $|Y - \hat{y}|$, and the pinball loss at level τ,

$$\rho_\tau(u) \;=\; u\,\big(\tau - \mathbf{1}\{u < 0\}\big), \qquad u = Y - \hat{y},$$

which charges $\tau|u|$ when you predict too low and $(1-\tau)|u|$ when you predict too high. Choosing among them is not a matter of taste, because each one is minimized by a different summary of the outcome distribution, and the derivations are one line each.

For squared error, differentiate under the expectation, $\frac{d}{d\hat{y}}\,\mathbb{E}(Y-\hat{y})^2 = -2\,\mathbb{E}[Y - \hat{y}]$, which is zero when $\hat{y} = \mathbb{E}[Y]$. For absolute error the derivative is $\mathbb{P}(Y < \hat{y}) - \mathbb{P}(Y > \hat{y})$, which is zero when $\hat{y}$ is the median. For the pinball loss the same calculation gives $F(\hat{y}) = \tau$, so the best prediction is the τ quantile.

Loss functions elicit different summaries

A loss does not only rank predictions. It decides which summary of the outcome distribution is best to report. Squared error is smallest when you report the mean, absolute error is smallest when you report the median, and the pinball loss at level τ is smallest when you report the τ quantile. If you train and evaluate with absolute error and then describe the output as an expected value, you have misdescribed your own system.

The same thing happens in the binary case, where the Brier score and log loss are smallest when you report the true probability. In both cases you choose the score so that reporting the quantity you actually want is the best a forecaster can do.

The scale of these losses depends on the population, exactly as the Brier score did. The usual repair is $R^2 = 1 - \mathrm{MSE}/\operatorname{Var}(Y)$, which is a skill score measured against the system that always predicts the mean, and it plays the same role that $1 - \mathrm{Brier}/\mathrm{Brier}_{\mathrm{base}}$ played for binary outcomes. It also inherits the same limitation, since $\operatorname{Var}(Y)$ differs between populations, so an R^2 of 0.4 in a homogeneous clinic and an R^2 of 0.4 in a mixed one describe different levels of predictive accuracy.

A single predicted HbA1c throws away the uncertainty, so the better output is a whole predictive distribution $\hat{F}$. The generalization of the Brier score to that case is the continuous ranked probability score,

$$\mathrm{CRPS}(\hat{F}, y) \;=\; \int_{-\infty}^{\infty} \big(\hat{F}(z) - \mathbf{1}\{z \geq y\}\big)^2 \, dz,$$

which compares the predicted cumulative distribution to the step function at the observed value. It is strictly proper, it is the continuous limit of the ranked probability score above, and when $\hat{F}$ puts all its mass at one point it equals the absolute error $|\hat{y} - y|$. There is an equivalent form that is often easier to compute,

$$\mathrm{CRPS}(\hat{F}, y) \;=\; \mathbb{E}\big|X - y\big| \;-\; \tfrac{1}{2}\,\mathbb{E}\big|X - X'\big|,$$

where X and X' are independent draws from $\hat{F}$. The first term rewards putting mass near the observed value and the second penalizes a distribution that is too wide, so the score balances accuracy against sharpness.

Calibration also generalizes. For a numeric outcome you compute the probability integral transform, $u_i = \hat{F}_i(y_i)$, which is where the observed value falls in the predicted distribution. If the predictive distributions are correct then the u_i are uniform on $[0, 1]$, so you plot their histogram. A U-shaped histogram means the distributions are too narrow, because outcomes land in the tails more often than predicted, and a histogram with a hump in the middle means they are too wide. The equivalent statement in the language people use in practice is that a 90% prediction interval should contain the truth 90% of the time, which you check by plotting observed coverage against the stated level. The uniform histogram and the coverage plot are the continuous versions of the reliability diagram.

11 Preferences, rankings, and text

Riverbend's assistant also drafts text, and the framework needs more work here. There is no ground truth for a clinical summary, because many summaries are acceptable and you cannot list the good ones.

Three common approaches make progressively stronger demands on raters:

- **Pairwise preference.** Instead of scoring outputs, compare them. Show a rater two summaries and ask which is better, recording that A beat B. Raters find this much easier than assigning a number, and it is the basis of most current language model evaluation.
- **Rankings.** Extend the comparison to an order over several outputs. Each judgment gives you more information, it costs more to collect, and you have to be careful about which comparisons you actually make.
- **Proxy scoring.** Have a human rater, a language model acting as a judge, or a trained reward model assign a score directly, and treat that score as the outcome.

11.1 Turning comparisons into scores

Write $Y_{ij} = 1$ when option i is preferred to option j and 0 otherwise. The standard model gives each option a latent quality θ and says the comparison depends only on the difference between the two qualities,

$$\mathbb{P}(i \succ j) \;=\; \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} \;=\; \sigma(\theta_i - \theta_j), \qquad \sigma(u) = \frac{1}{1 + e^{-u}},$$

which is the Bradley-Terry model. Only differences in θ affect the probabilities, so the parameters are identified only up to adding a constant to all of them, and you fix this by setting $\sum_i \theta_i = 0$ or by holding one option at zero.

Fitting is ordinary logistic regression on the comparisons. The log-likelihood is

$$\ell(\theta) \;=\; \sum_{(i,j)} \Big[Y_{ij} \log \sigma(\theta_i - \theta_j) \;+\; (1 - Y_{ij}) \log \big(1 - \sigma(\theta_i - \theta_j)\big) \Big],$$

and its derivative with respect to θ_i is $\sum_j \big(Y_{ij} - \sigma(\theta_i - \theta_j)\big)$, which is the observed number of wins minus the number the model expected. Taking one gradient step per comparison gives

$$\theta_i \;\leftarrow\; \theta_i + K\big(Y_{ij} - \sigma(\theta_i - \theta_j)\big),$$

which is the Elo update used in chess and on model leaderboards. Elo is online gradient ascent on the Bradley-Terry likelihood, which is why Elo ratings and Bradley-Terry fits usually agree, and why Elo depends on the order in which the comparisons arrive while the fitted model does not.

Ties need a decision. The simplest treatment scores them as half a win to each side, and the Davidson model instead adds a parameter for how often ties occur.

11.2 Evaluating a preference model

Once you have $\mathbb{P}(i \succ j)$ the outcome is binary again, so everything from the earlier sections applies without change. Hold out a set of comparisons, and score the predicted probabilities with the Brier score or log loss. Calibration means what it meant before, which is that among the comparisons where the model said 0.70, the first option won about 70% of the time. Accuracy on held-out comparisons has the same weakness it had before, since two models can agree on every winner and disagree about how confident to be.

The model does make an assumption you should test, which is that a single number per option explains every comparison. Real preferences can be intransitive, so raters may prefer A to B, B to C, and C to A, and no assignment of θ reproduces that pattern. Check for it by comparing observed win rates against fitted ones for each pair and looking for pairs where the model is systematically wrong.

For rankings rather than pairs, compare a predicted order to a reference order with Kendall's τ, which counts the fraction of pairs placed in the same relative order, or with Spearman's rank correlation. Both reduce a ranking comparison to a number between -1 and 1.

11.3 Proxy scores

The third response assigns a quality score $s(\text{text})$ directly, using a human rater, a language model, or a trained reward model. Doing so returns you to the ordinal or continuous case, and every tool

from the previous section applies to the resulting numbers.

These approaches relocate the measurement problem rather than eliminating it: preference judgments and proxy scores become the measurements to validate. Does a rater's preferred summary reflect clinical usefulness, fluency, or length? Do raters agree? Does an LLM judge prefer the first response, the longer response, or prose that resembles its own? These are construct-validity and reliability questions about the evaluator. A preference study can be carefully run yet still measure the wrong property. We return to these questions in the second term.

Discussion questions

1. Redesign Riverbend's physician study with the same budget. What do you randomize, what do you measure, and which of the four validity questions does your design still fail?
2. System A and System B have the same accuracy and the same AUC. Construct a third system with better accuracy than both and a worse Brier score than both, and say what it is doing.
3. You are asked to evaluate a system that drafts discharge instructions. Write down the construct, then two things you could measure that would rank systems differently, and say which one you would defend to a patient safety committee.