

Evaluation Metrics and Thresholds

Master reference card, "Healthcare Analytics and AI: Building Systems That Actually Work"

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Quick-lookup card: every evaluation metric in the book, its formula, its failure mode, and the number to remember.

How to read this card. Three questions, three metric families: does the model *rank* (discrimination), does “70 percent” *mean* 70 percent (calibration), and does using it *help the patient* (net benefit). A model can ace ranking and fail the other two, and the other two decide deployment. Notation: TP, FP, TN, FN are the confusion-matrix counts, N the sample size, p a predicted probability, y the observed label.

PART 1: Classification

Discrimination and Confusion-Matrix Metrics

	predicted –	predicted +
actual +	FN missed	TP caught
actual –	TN correct	FP false alarm

Confusion-matrix legend: the four counts every metric on this card is built from.

Metric	Formula	Measures	Pitfall	Typical value / threshold
Accuracy	$\frac{TP+TN}{N}$	Overall fraction correct	Lies under imbalance: “predict none” scores $1 - \text{prev}$	≈ 0.85 saying no on 15% readmission; never report alone
Sensitivity / Recall / TPR	$\frac{TP}{TP+FN}$	Share of true positives caught	High sensitivity inflates alert volume	Demand ≥ 0.90 for must-not-miss (sepsis, MI); retinopathy 91.4%
Specificity / TNR	$\frac{TN}{TN+FP}$	Share of true negatives cleared	Trades off against sensitivity	~ 0.60 when sensitivity forced to 0.90; retinopathy 95.4%
PPV / Precision	$\frac{TP}{TP+FP}$	$P(\text{disease} \mid \text{flagged})$	Collapses at low prevalence even with strong sens/spec	90/90 at 5% prev $\Rightarrow$ PPV ≈ 0.32 ; Epic sepsis 12%
NPV	$\frac{TN}{TN+FN}$	$P(\text{well} \mid \text{cleared})$	Looks great whenever disease is rare	Near 1.0 at low prevalence; rarely binding
F1	$\frac{2 \text{PPV} \cdot \text{Recall}}{\text{PPV} + \text{Recall}}$	Harmonic mean of precision and recall	Ignores TN; hides the sens/spec trade	Use for imbalance; report at a chosen threshold
AUROC (c-statistic)	$P(\hat{s}_+ > \hat{s}_-)$	Ranking: random + above random –	Blind to calibration; weak on very rare events	0.5 chance, 1.0 perfect; ≥ 0.80 good; Epic 0.63 vs 0.76–0.83
AUPRC	Area under PR curve	Ranking on the rare positive class	Baseline moves with prevalence	Random baseline = prevalence (≈ 0.05 – 0.07 sepsis)

Note: at 90% sensitivity and 90% specificity on a 5% condition, two of every three flags are false positives. This Bayes arithmetic, not model quality, is why good models still cry wolf.

Do the Probabilities Mean What They Say

Metric	Formula	Measures	Pitfall	Typical value / threshold
Brier score	$\frac{1}{N} \sum_i (p_i - y_i)^2$	Calibration + discrimination (proper score)	Base-rate dependent; hard to read alone	0 perfect, 1 worst; 0.25 random at balance; trust over AUROC
Expected calibration error (ECE)	$\sum_b \frac{n_b}{N} \text{acc}_b - \text{conf}_b $	Mean predicted-vs-observed gap, binned	Bin count and width change the number	Monitor over time; retrain when $\text{ECE} > \approx 0.10$
Calibration slope	Slope of observed on $\text{logit}(p)$	Over/under-confidence spread	AUROC-only reporting hides it	Ideal 1.0; < 1 means overfit / overconfident
Calibration intercept	Intercept of that regression	Systematic over/under-prediction (in-the-large)	Passes silently without a reliability plot	Ideal 0.0; $\neq 0$ shifts every prediction

Note: Van Calster and colleagues (2019) argued calibration should take precedence over discrimination for clinical use. A model that says 70% but is right 30% of the time makes clinicians over-intervene. Recalibrate (Platt, isotonic) on a held-out set, never the training or test set.

Net Benefit and Time-to-Event

Metric	Formula	Measures	Pitfall	Typical value / threshold
Net benefit (decision curve)	$\frac{TP}{N} - \frac{FP}{N} \cdot \frac{p_t}{1-p_t}$	Clinical utility vs treat-all / treat-none	Needs a threshold range, not one number	Beat both defaults across the p_t band (readmit $\approx 15\text{--}40\%$)
C-index (Harrell)	$P(\hat{s}_{\text{early}} > \hat{s}_{\text{late}})$ over comparable pairs	Discrimination for censored survival (AUROC analog)	Alignment, not accuracy; degrades with heavy censoring	0.5 chance, 1.0 perfect; Cox loses ~ 10 pts above 90% censoring
Time-dependent AUC	AUROC for event by time t	How discrimination shifts with horizon	One horizon can hide failure at another	Report a curve over t , not one value
Cumulative incidence (CIF) calibration	Fine-Gray subdistribution vs observed	Absolute-risk accuracy under competing risks	KM overstates risk 10–15 pts when competing events common	Use CIF , not $1 - \text{KM}$, in elderly cohorts

PART 4: Imaging Segmentation

Overlap Metrics

Metric	Formula	Measures	Pitfall	Typical value / threshold
Dice similarity coefficient	$\frac{2 A \cap B }{ A + B }$	Overlap of mask A and truth B	Rewards bulk overlap; weak on thin structures	0–1; federated BraTS 0.733 ET, 0.761 TC, 0.751 WT
IoU / Jaccard	$\frac{ A \cap B }{ A \cup B }$	Same overlap, harsher on errors	Always $\leq$ Dice; do not compare across metrics	Dice = $\frac{2IoU}{1+IoU}$; MedGemma 38% vs 3%

Note: for image *classification* (retinopathy, dermatology) use the Part 1 metrics, disaggregated by subgroup. Dermatology AI pooled AUROC ran ≈ 0.82 on darker skin vs 0.89 on lighter, a gap invisible in the aggregate.

Group Metrics and Gaps (report as the max gap across groups)

Metric	Formula	Measures	Pitfall	Typical value / threshold
Demographic parity	Equal $P(\hat{y}=1 \mid g)$ across groups	Equal selection / flag rate	Underpredicts higher-prevalence group; conflicts with calibration	Gap ideally 0; four-fifths ratio ≥ 0.80
Equalized odds	Equal TPR and FPR across groups	Errors distributed equally	Needs unbiased labels; conflicts with dem. parity	Report max TPR, FPR gaps; both $\rightarrow 0$
Equal opportunity	Equal TPR across groups	Equal miss rate on true positives	Ignores false-positive burden	Weaker equalized-odds variant; TPR gap $\rightarrow 0$
Predictive parity	Equal PPV across groups	Flag means the same per group	Insensitive to base-rate differences	PPV gap $\rightarrow 0$
Calibration by group	Score = actual prob. per group	Per-group probability honesty	Can hide disparate error rates	Slope/intercept per group; conflicts with eq. odds

Note: the impossibility result (Chouldechova 2017): when base rates differ across groups, you cannot satisfy calibration and equalized odds at once. Pick the metric that matches the clinical harm; in healthcare, equal sensitivity is usually the defensible choice.

Monitoring and Privacy Budgets

Metric	Formula	Measures	Pitfall	Typical value / threshold
PSI (population stability index)	$\sum_b (a_b - e_b) \ln \frac{a_b}{e_b}$	Feature shift (actual a vs expected e)	Bin choice drives it; flags input, not outcome, drift	< 0.1 stable, 0.1–0.25 moderate, > 0.25 retrain
Kolmogorov-Smirnov (KS)	$\max_x F_{\text{deploy}}(x) - F_{\text{train}}(x) $	Largest gap between two CDFs	Picks any change, not a harmful one	Alert when statistic (or p -value) crosses a preset bound
Jensen-Shannon divergence	Symmetrized, bounded KL	Distance between train and live distributions	A distance, not a performance measure	0 identical to 1 ($\log_2$); pair with a perf. check
Differential privacy ϵ	$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S] + \delta$	Privacy loss bound per release	Small ϵ kills utility; budget compounds over queries	$\epsilon \leq 1$ strong, > 10 weak; aggregates 0.5–2.0

Note: the DP failure term δ should be small (typically $\ll 1/N$). Google runs $\epsilon \approx 2$ with a lifetime cap of 8–9 per user; a single sum query in healthcare can force massive noise because one transplant stay ($\sim \$2.3\text{M}$) sets the sensitivity.

Read the Right Metric for the Job

Question you are answering	Use	Do NOT rely on
Which candidate model ranks best in development?	AUROC (AUPRC if rare)	Accuracy
Can I trust the probability at the bedside?	Calibration slope/intercept, Brier, ECE	AUROC
Should we deploy this at all?	Net benefit / decision curve	Any single ranking number
How many alerts, and how many are real?	Sensitivity, specificity, PPV at threshold	AUROC
Is it fair across groups?	Equalized-odds gap, calibration by group	Aggregate AUROC
Is the live model still valid?	PSI / KS on inputs, ECE + subgroup sensitivity	Training-time metrics

If you remember one thing. Discrimination, calibration, and net benefit are three different questions. The Epic Sepsis Model failed all three (AUROC 0.63, missed 67% of sepsis, 12% PPV when it fired) and still shipped. A confidently wrong model is worse than none, because it manufactures a false sense of safety.

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