

Article

Deferral-Rule Geometry, Not Multidimensional Reliability, Explains Selective-Prediction Gains in Clinical AI: A Three-Database Intensive Care Study

Yunguo Yu ^{1,†} , Bernardo Gabriele Collaco ^{2,†} , Srinivasagam Prabha ² , Carina Rosa Malena ², Nadia G. Wood ³ 
and Antonio Jorge Forte ^{1,2,4,*} 

¹ Center for Digital Health, Mayo Clinic, Rochester, MN 55905, USA

² Division of Plastic Surgery, Mayo Clinic, 4500 San Pablo Rd S, Jacksonville, FL 32224, USA

³ Department of Radiology of Artificial Intelligence & Information Technology, Mayo Clinic, Rochester, MN 55905, USA

⁴ Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL 32224, USA

* Correspondence: ajvforte@yahoo.com.br

† These authors contributed equally to this work.

Abstract

Multidimensional reliability scores have been proposed as better criteria than confidence for deferring clinical predictions to clinicians. We evaluated reliability-gated deferral for sepsis, mortality, and vasopressor prediction in intensive care patients from MIMIC-IV (68,546 stays), MIMIC-III (42,242), and eICU (119,757 stays; 206 hospitals), with patient-level splits, held-out hospitals, and MIMIC-IV-to-eICU transfer. Against a one-sided gate that defers the lowest predicted risks, a reliability composite reduced the Brier score of retained predictions 1.5- to 2.1-fold at 80% coverage. Against a symmetric gate, $\max(\tilde{p}, 1 - \tilde{p})$, which minimizes this error for a calibrated model, the composite offered no advantage in any of 12 database–task pairs; the symmetric gate was better in five. At matched predicted risk, missingness-aware reliability dimensions and an error model given the entire patient record showed no statistically detectable improvement in deferral beyond calibrated confidence in any of 75 planned tests (54 confirmatory), although injected error signals worth 0.004–0.007 in retained Brier score were detectable. After transfer between databases, calibration error reached 19.7% and some scores carried outcome information beyond predicted risk, but recalibration at the target site improved retained predictions far more than any score. Deferral signals should be evaluated against symmetric confidence at matched risk; shift calls for recalibration.

Keywords: selective prediction; learning to defer; calibration; Brier score; uncertainty quantification; distribution shift; clinical decision support; intensive care units; sepsis; MIMIC-IV; eICU



Academic Editor: Milan Toma

Received: 6 October 2026

Revised: 20 October 2026

Accepted: 21 October 2026

Published: 23 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Clinical prediction models for sepsis, mortality, and deterioration routinely report areas under the receiver operating characteristic curve (AUROCs) above 0.85 on held-out data [1–3]. Average performance, however, says little about any individual prediction. Models fail systematically on subgroups and new sites [4–6], and those failures are rarely signaled to the clinician who acts on the output [7,8]. A model’s self-reported confidence need not track whether a given output is correct [9], and clinician trust depends on knowing when not to rely on the model [10,11].

Selective prediction addresses this by letting the model abstain. The classical reject option [12] and its modern forms [13,14] threshold a confidence score. Learning-to-defer methods instead train the deferral decision jointly with, or on top of, the classifier [15,16], and clinical applications have shown that deferring to clinicians can improve joint human–AI accuracy [17]. Conformal methods add distribution-free coverage guarantees [18]. A common critique of confidence-based deferral is that confidence is one-dimensional: a model can be confident for the wrong reasons, for example when it extrapolates beyond its training distribution or rests on sparse inputs [19–21]. This critique motivates multidimensional reliability scores that supplement confidence with signals about data sufficiency, distributional familiarity, and prediction stability.

An earlier preprint of this work [22] proposed such a score for intensive care unit (ICU) prediction and reported large gains over confidence gating: the two criteria deferred almost entirely different patients, and the reliability gate halved the Brier score of retained predictions. Those comparisons used a one-sided confidence gate that defers to the patients with the lowest predicted risk ($\tilde{p} < t$). The textbook selective-prediction rule is symmetric: it defers predictions whose maximum-class probability, $\max(\tilde{p}, 1 - \tilde{p})$, is lowest. The two rules defer opposite ends of the risk distribution. Two further problems affect this comparison and, more generally, the evaluation of deferral signals. First, metrics computed on retained patients depend on whom each rule retains, and the Brier score depends on prevalence [23,24]; a rule can look better simply by retaining an easier population. Second, a reliability score adds value only if, among patients with the same predicted risk, it identifies the predictions that are wrong. That comparison, which holds case mix fixed, has not been made for clinical reliability scores, and it has not been made under distribution shift, where such scores are expected to matter most. Evaluations of confidence scoring outside clinical outcome prediction suggest that simple confidence is hard to beat [25,26], but none has tested this at matched predicted risk.

In the present study, we re-evaluated reliability-gated deferral for sepsis, mortality, and vasopressor prediction in three ICU databases (MIMIC-IV, MIMIC-III, and eICU) on cohorts split by patient, at eICU hospitals held out from training, and after transfer of the MIMIC-IV pipeline to eICU. To the best of our knowledge, this is the first study to test reliability-gated deferral against calibrated confidence at matched predicted risk in clinical outcome prediction, including under site and database shifts (Figure 1). We contribute (1) an analytical reference point, showing that for a calibrated model the expected squared error of predictions with risk $\tilde{p}$ is $\tilde{p}(1 - \tilde{p})$ and that, among patients with the same $\tilde{p}$, a score can find additional errors only if it carries information about the outcome that the model lacks; (2) evidence across 12 database–task pairs that a reliability composite’s 1.5- to 2.1-fold advantage over the one-sided rule disappears against the symmetric rule; (3) a risk-matched test, calibrated by positive controls, in which neither missingness-aware reliability dimensions nor an error model given the entire patient record improved deferral beyond confidence, in or out of distribution, within the sensitivity of the test; and (4) evaluation recommendations for clinical selective prediction, with public code and dated analysis plans. Table 1 summarizes how this study differs from the earlier preprint.

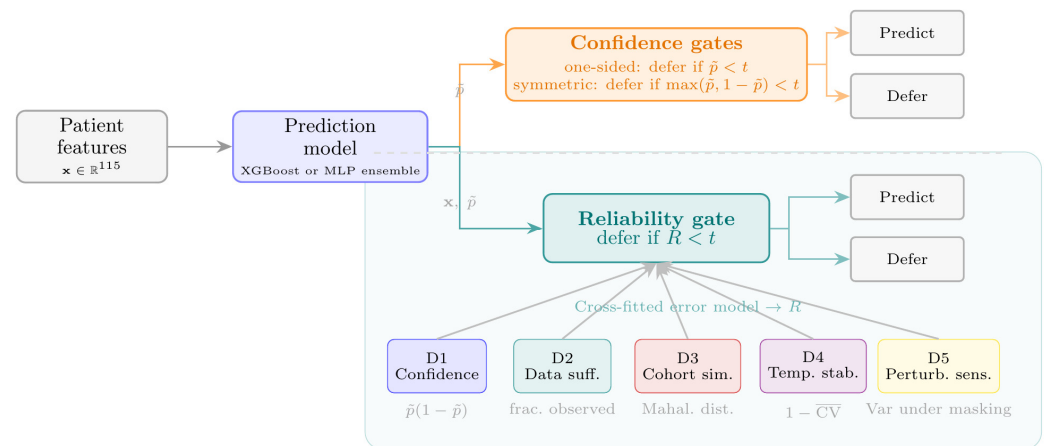


Figure 1. Deferral rules compared. **(Top):** two confidence gates computed from the calibrated predicted risk $\hat{p}$. The one-sided gate defers the lowest predicted risks. The symmetric gate defers predictions closest to $\hat{p} = 0.5$. **(Bottom):** the reliability gate defers patients with the lowest reliability R , the negative of a cross-fitted model’s predicted squared error given five candidate dimensions, each annotated with its operationalization (Section 2.4). All gates are compared at matched coverage on the same predictions.

Table 1. Relation of this study to the earlier preprint [22].

Aspect	Earlier Preprint	This Study
Comparator	One-sided gate ($\hat{p} < t$), labeled a confidence gate	One-sided and symmetric gates; temperature-scaled entropy, four other uncertainty scores, and SelectiveNet
Data and splits	MIMIC-IV (temporal split by de-identified year); MIMIC-III replication	MIMIC-IV, MIMIC-III, and eICU extracted with shared definitions; patient-level splits; 206-hospital eICU with held-out hospitals; MIMIC-IV $\rightarrow$ eICU transfer
Reliability score	Logistic composite of five dimensions	Cross-fitted error model; missingness-aware dimensions; error model on the full record; continuous sufficiency measure D2* (see footnote a)
Evaluation	Retained-set Brier score and NNR at fixed coverage	Paired tests with CIs; risk-matched test with positive controls; prevalence-adjusted skill; class-conditional, subgroup, clinical-score, and imperfect-review analyses
Theory	None	Expected-error identities, Equations (1) and (2)
Conclusion	Reliability is not reducible to confidence and outperforms it	No signal improved deferral beyond confidence at matched risk; the apparent gain reflected which tail of risk each rule defers

a D2*: variance of the primary model’s prediction over 50 imputations of missing values; replaces the saturated sufficiency ratio D2. Descriptive analysis.

2. Materials and Methods

2.1. Data Sources and Cohorts

We used three intensive care databases. MIMIC-IV v3.1 [1,27] (Beth Israel Deaconess Medical Center [BIDMC], 2008–2022) and MIMIC-III v1.4 [28] (BIDMC, 2001–2012) were processed with one shared extraction. The cohort was the first ICU stay of each hospital admission in adults (≥ 18 years) whose stay lasted at least 24 h and who had at least one observation in the feature window. This gave 68,546 stays from 54,551 patients in MIMIC-IV and 42,242 stays from 33,520 patients in MIMIC-III. The eICU Collaborative Research Database v2.0 [29] contributed 119,757 first ICU stays per hospitalization from

103,240 patients in 206 U.S. hospitals, with the same age and length-of-stay criteria and a known hospital discharge status.

Features were computed over the first 24 h after ICU admission from 23 raw signals (10 vital signs and 13 laboratory tests, including creatinine, lactate, hemoglobin, white blood cell count, platelets, electrolytes, and liver enzymes). Each signal was summarized by its mean, minimum, maximum, last value, and number of observations, giving 115 features with identical names in all three databases. One signal, temperature in degrees Fahrenheit, has no eICU equivalent; its four summary features are always missing there and its observation count is zero.

Each database was split into training, validation, and test sets by patient, so that no patient contributed to more than one set, using seven stratified group folds (one for test, one for validation; approximately 71/14/14%; seed 42), separately for each task. MIMIC shifts dates by a random offset per patient, so no split can be both patient-level and calendar-ordered; all splits are internal hold-out validation. To test transfer to unseen institutions, eICU was also split by hospital (eICU-H): whole hospitals were assigned to training, validation, or test, leaving 37, 33, and 19 test hospitals for the sepsis, mortality, and vasopressor tasks. Table 2 summarizes the test sets.

Table 2. Test cohorts and base-model performance (calibrated XGBoost ensemble, full test set, no deferral). M-ECE: expected calibration error with 10 equal-mass bins. ^a Held-out hospitals.

Database	Task	Test <i>n</i>	Prevalence	AUROC	Brier	M-ECE
MIMIC-IV	Sepsis	9792	16.9%	0.879	0.094	0.7%
	Mortality	9793	11.1%	0.858	0.076	0.8%
	Vasopressor	6946	5.5%	0.746	0.049	0.6%
MIMIC-III	Sepsis	6036	11.6%	0.867	0.076	1.1%
	Mortality	6034	11.3%	0.849	0.077	0.8%
	Vasopressor	4160	7.9%	0.743	0.066	1.1%
eICU	Sepsis	17,108	14.5%	0.861	0.091	0.7%
	Mortality	17,108	8.8%	0.825	0.065	0.4%
	Vasopressor	10,380	3.4%	0.743	0.032	0.6%
eICU-H ^a	Sepsis	17,017	14.7%	0.870	0.090	0.9%
	Mortality	17,080	8.9%	0.831	0.065	0.6%
	Vasopressor	10,414	3.3%	0.742	0.031	0.5%

2.2. Prediction Tasks

All predictions were made 24 h after ICU admission. Sepsis was defined in MIMIC by International Classification of Diseases (ICD)-9 codes 995.91, 995.92, and 785.52 and their ICD-10 equivalents (A41.9, R65.20, R65.21), and in eICU by admission diagnoses in the sepsis and septic-shock categories, excluding non-infectious systemic inflammatory response syndrome and septic arthritis. ICD-based labels are less specific than Sepsis-3 criteria [30], and the case-identification method changes sepsis incidence [31]; we therefore repeated the MIMIC-IV analyses with Sepsis-3 labels (Section 2.7). In-hospital mortality was taken from discharge status. Vasopressor initiation was the first infusion of norepinephrine, epinephrine, phenylephrine, vasopressin, or dopamine between 24 h and ICU discharge, among stays with no vasopressor before 24 h (48,615 MIMIC-IV, 29,120 MIMIC-III, and 72,660 eICU stays). In eICU, this task was restricted to the 146 hospitals that chart vasopressor infusions.

2.3. Base Models

The primary base model was a five-seed XGBoost v3.2 ensemble (up to 500 trees, maximum depth 6, learning rate 0.05, subsample and column subsample 0.8, minimum child weight 5, L2 penalty 1), trained on unimputed features with native missing-value handling and stopped early after 20 rounds without improvement in validation log loss. No class weighting was used: the analysis depends on predicted probabilities that mean what they say, and reweighting classes would distort them and require the isotonic step to undo the distortion. The ensemble mean was calibrated by isotonic regression on the validation set to give $\tilde{p}$. Training diagnostics for every model (stopping iteration 44–420; training–validation AUROC gap 0.056–0.166) are in Table S1; the test AUROC of each calibrated ensemble (Table 2) was within 0.026 of its members' validation AUROC. To test model dependence, we repeated the MIMIC-IV analyses with a five-seed multilayer perceptron (MLP; hidden layers of 128 and 64 units, dropout 0.1).

2.4. Reliability Dimensions and Composite

The composite combined confidence with four dimensions, each targeting a clinically motivated failure mode:

1. D1, confidence: $\tilde{p}(1 - \tilde{p})$.
2. D2, data sufficiency: The fraction of the 23 signals with at least one observation in the 24 h window. (The earlier preprint divided the fraction observed in the 24 h window by the fraction observed in its first 6 h, clipped to 1. Because the 6 h window lies inside the 24 h window, the ratio is at least 1 whenever any signal was observed early, and it was at its ceiling for 98.4–99.9% of test patients, so it was not used. A continuous alternative, D2*, is analyzed in Section 2.8).
3. D3, cohort similarity: $D_3 = \exp(-d(\mathbf{x})/\sigma)$, where $d(\mathbf{x})$ is the Mahalanobis distance of the standardized, median-imputed feature vector to the training mean under a Ledoit–Wolf shrinkage covariance [32], and σ is the median training-set distance.
4. D4, temporal stability: One minus the mean coefficient of variation in each signal within the window, over signals with at least two observations and nonzero mean (coefficients clipped to [0, 1]; 0.5 when no signal qualifies).
5. D5, perturbation sensitivity: The variance of the primary model's predicted probability over 50 passes in which a random 10% of the median-imputed inputs were set to zero.

The composite was a model of each patient's squared error. On the validation set, split in two halves by patient, isotonic calibration was fitted on one half and the target $(\tilde{p} - y)^2$ computed on the other, and vice versa, so that the calibrator and the error target never shared patients. A gradient-boosted regressor (XGBoost; depth 3, 300 trees, learning rate 0.05, subsample 0.8), constrained to be monotone increasing in D1, was then fitted to the pooled halves. The reliability score was $R(\mathbf{x}) = -\hat{\mathbb{E}}\left[(\tilde{p} - y)^2 \mid \mathbf{x}\right]$. This specification, which differs from the logistic composite of the earlier preprint, and its primary comparison were fixed in a dated analysis plan before any result was examined.

2.5. Deferral Rules

Every rule ranks patients by a score and retains the highest-ranked fraction; we use gate for a rule applied at a given coverage:

- One-sided confidence: score $\tilde{p}$, so the lowest predicted risks are deferred. This was the comparator in the earlier preprint.
- Symmetric confidence: score $\max(\tilde{p}, 1 - \tilde{p})$, so predictions closest to 0.5 are deferred.
- Reliability: score R .

- Alternative uncertainty scores: ensemble variance (split-conformal score), Monte Carlo dropout variance, binary entropy after temperature scaling [33], and the posterior predictive variance of a Laplace-approximated Bayesian logistic regression [34].
- SelectiveNet [14]: a jointly trained prediction-and-rejection network with target coverage 0.8.

Temperature-scaled entropy ranks patients exactly as the symmetric rule applied to the uncalibrated ensemble output, because entropy is a monotone function of the distance of the logit from zero.

2.6. An Analytical Reference Point

For a binary outcome y with true conditional risk $\eta(\mathbf{x}) = \Pr(y = 1|\mathbf{x})$ and a prediction $\tilde{p}$, the expected squared error of one patient is

$$\mathbb{E}\left[\left(\tilde{p} - y\right)^2 \mid \mathbf{x}\right] = \tilde{p}^2 + \eta(\mathbf{x})\left(1 - 2\tilde{p}\right) \quad (1)$$

If $\tilde{p}$ is calibrated, $\mathbb{E}\left[\eta(\mathbf{x}) \mid \tilde{p}\right] = \tilde{p}$, and averaging Equation (1) over patients with the same $\tilde{p}$ gives

$$\mathbb{E}\left[\left(\tilde{p} - y\right)^2 \mid \tilde{p}\right] = \tilde{p}\left(1 - \tilde{p}\right) \quad (2)$$

Two consequences organize the study. First, the expected Brier score of the retained set at any fixed coverage is minimized by retaining the patients with the smallest $\tilde{p}(1 - \tilde{p})$, that is, by the symmetric rule. For tasks in which most predictions lie below 0.5, the one-sided rule does the opposite and removes the patients whose expected error is smallest; a gap between a reliability gate and the one-sided gate therefore says nothing about the reliability dimensions. Second, among patients with the same $\tilde{p}$, Equation (1) varies only through $\eta(\mathbf{x})$. A score can identify errors beyond confidence only if it carries information about the outcome that $\tilde{p}$ does not, that is, only if the base model left recoverable signal. This is what the risk-matched test in Section 2.8 measures.

2.7. Whole-Cohort Comparisons

The main analyses compared gates at 80% coverage (20% deferral) with a stable top- k selection (ties in the score keep their row order). No threshold was tuned: each gate retained a fixed fraction of patients ranked by its own score, so all gates were compared at identical workload, and full risk–coverage curves (50–100% coverage) show that conclusions do not depend on the chosen operating point. On retained patients we report the Brier score [35], AUROC, and outcome prevalence; on deferred patients, the misclassification rate at 0.5 and the number needed to review (NNR; deferred patients per deferred error [36]). The Jaccard index quantifies overlap between deferred sets. Gates were compared by a paired bootstrap: each of 2000 resamples (seed 42) drew test patients with replacement, and both gates were re-evaluated on the same resampled patients. We report 95% percentile confidence intervals (CIs) and two-sided bootstrap p -values with Holm–Bonferroni correction [37] across the nine database–task pairs, and separately across the three eICU-H pairs (smallest attainable adjusted p : 0.0045 and 0.0015). Because the retained populations of different gates differ in prevalence, and the Brier score depends on prevalence [23,24], we also report the retained-set Brier skill score, $1 - \text{Brier}/[\pi(1 - \pi)]$ with π the retained prevalence, and net benefit [38] on retained patients for thresholds 0.02–0.50.

Secondary analyses addressed specific questions. Class-conditional Brier score, sensitivity, and specificity at 0.5 were computed for each gate, and calibration was displayed in 10 equal-mass bins and summarized by the expected calibration error over the same

bins (M-ECE). A subgroup audit compared deferral rate, share of the group's errors that were deferred, retained Brier score, and the rate of deferring correctly classified patients across sex, age (18–39, 40–54, 55–64, 65–74, and ≥ 75 years), and five race/ethnicity groups. Expected errors under imperfect clinician review were computed for review accuracies of 70–100%, with the break-even accuracy below which deferral adds errors. MIMIC-IV analyses were repeated with Sepsis-3 labels (mimic_derived.sepsis3). For mortality, the first-day Sequential Organ Failure Assessment (SOFA) score (MIMIC-IV, recalibrated by logistic regression on the validation set) and the Acute Physiology and Chronic Health Evaluation (APACHE) IVa predicted mortality (eICU) were evaluated as clinical baselines and as inputs to the symmetric gate. The composite comparison was repeated over five patient-level split seeds. Stress tests applied frozen pipelines to perturbed test features: zeroing 10–50% of feature columns, which injects out-of-range values rather than simulating missingness, and shifting a renal (creatinine, blood urea nitrogen) or liver (alanine and aspartate aminotransferase, bilirubin) panel by 0.5–2.0 training-set standard deviations, with five replicates per condition. Reporting follows TRIPOD + AI [39].

2.8. Risk-Matched Deferral Test

Whole-cohort comparisons cannot separate a score's information about error from the case mix it retains. The risk-matched test holds predicted risk fixed. Test patients were sorted by $\tilde{p}$ (ties broken by a fixed random permutation) and cut into $K = 10$ equal-size bins. Within each bin, the 20% of patients with the lowest score were deferred, so every score retained the same distribution of predicted risk. The effect of a score was

$$\Delta = \bar{e}_{\text{random}} - \bar{e}_{\text{retained}}$$

where $e = (\tilde{p} - y)^2$, $\bar{e}_{\text{retained}}$ is the mean squared error of the retained patients, and $\bar{e}_{\text{random}}$ is its exact expectation under random deferral within bins. Residual variation of $\tilde{p}$ within a bin gives the symmetric rule a small advantage over random deferral, so the primary effect was the gain beyond the symmetric rule at matched risk, $\Delta_{\text{adj}} = \Delta(\text{score}) - \Delta(\text{symmetric})$. Inference used a paired bootstrap of 1000 resamples (seed 42) in which bins were recomputed on each resample, with percentile CIs and two-sided p -values floored at 0.001; Holm correction was applied within each setting across 15 tests (five scores $\times$ three tasks). $K = 20$ was a sensitivity analysis.

Two references calibrate the test. The oracle score $-e$ gives the largest achievable effect. Positive controls $-(e + \epsilon)$, with noise ϵ normally distributed with standard deviation $k \text{SD}(e)$ and $k \in \{1, 2, 4\}$, are scores that carry a known amount of information about each patient's error; a test that cannot detect them cannot support a negative conclusion. Because these controls are strong (Section 3.5), we also estimated post hoc the smallest detectable gain by titrating k (Table S14).

The scores were constructed to give the reliability hypothesis its best chance. D3 and D5 as defined above operate on median-imputed inputs, which hides missingness from D3 and presents the NaN-trained model with zeros it never saw. We therefore also constructed missingness-aware versions: D3^m computes the same Mahalanobis similarity after appending binary missingness indicators, and D5^m is the variance of the primary model's prediction over 50 passes in which 10% of the observed entries were set to missing. C^m is a cross-fitted ridge model of squared error on $\tilde{p}(1 - \tilde{p})$, D2, D3^m, D4, D5^m, and the ensemble standard deviation. L is a cross-fitted gradient-boosted model of squared error (depth 2, 200 trees, learning rate 0.03, minimum child weight 50) given $\tilde{p}(1 - \tilde{p})$ and the entire 115-feature record, and L + D adds the dimensions to L. L tests whether any error signal recoverable from the patient record exists; L + D tests whether the dimensions

add to it. A further, descriptive score, D2, replaces the saturated sufficiency ratio with the variance of the primary model's (ensemble member 1) prediction over 50 completions of the patient's missing values by predictive mean matching [40]; it is large when the missing measurements would move the prediction and cannot be inferred from the recorded ones. We also report, without multiplicity adjustment, the dimensions and composite as defined above.

The test was run in five settings: MIMIC-IV, MIMIC-III, and eICU (patient splits); eICU-H (held-out hospitals); and MIMIC-IV $\rightarrow$ eICU, in which the MIMIC-IV pipeline (ensemble, calibrator, dimensions, and error models) was applied unchanged to the eICU test patients. The analysis plan was fixed and dated before its results were examined (Table S11). Because an exploratory check on in-distribution data preceded it, L and L + D are exploratory in the three in-distribution settings and confirmatory in the two shift settings, and D5^m is exploratory in MIMIC-IV; the remaining 54 of the 75 tests are confirmatory. The primary contrast was first specified relative to random deferral; after those results showed that composites containing $\tilde{p}(1 - \tilde{p})$ matched the symmetric rule's own within-bin gain, the plan was amended to make Δ_{adj} primary, before any result for the missingness-aware or learned scores was examined in the two shift settings. Both contrasts are reported.

Five analyses were added post hoc, after the planned results were known (Table S14): positive controls titrated to $k = 4, 8, 16$, and 32 , each injected 10 times, to estimate the smallest detectable gain; a hospital-cluster bootstrap for eICU-H and MIMIC-IV $\rightarrow$ eICU; a likelihood-ratio test of whether each score adds outcome information to a natural cubic spline of $\logit \tilde{p}$, which uses all patients and absorbs any monotone recalibration (Holm correction within setting); isotonic recalibration of the transferred model on eICU validation patients; and recovery by L of a planted feature carrying error information. Four further post hoc analyses (Table S15) examined per-hospital heterogeneity at held-out eICU hospitals (random-effects pooling of within-hospital gate differences), decision loss at clinical thresholds $t = 0.05\text{--}0.50$ (false positives costing t and false negatives $1 - t$), transfer with harmonized features (observation counts and temperature removed from all model inputs, leaving 84 features), and sepsis defined by Sepsis-3 onset after the 24 h prediction time in MIMIC-IV (stays with earlier onset excluded).

2.9. Software and Reproducibility

Analyses used Python 3.11, XGBoost 3.2, scikit-learn 1.9, pandas 3.0, PyTorch 2.14, and statsmodels 0.15. Seeds were 0–4 for ensemble members and shift replicates and 42 for splits, cross-fitting, and bootstrap resampling. Computing D1, D3, and D5 took under 0.1 ms per patient on one CPU core (batched, Apple M1 Pro, Apple Inc., Cupertino, CA, USA). Code is available as described in the Data Availability Statement.

3. Results

3.1. Reliability Versus One-Sided Confidence

Against the one-sided gate, the reliability composite reproduced the headline finding of the earlier preprint in all three databases and at held-out hospitals (Table 3). Deferred sets were nearly disjoint (Jaccard ≤ 0.02 in all 12 pairs), and the retained-set Brier score was 1.54- to 2.12-fold lower under the composite. All 12 Brier differences were statistically significant after Holm correction (adjusted $p \leq 0.0045$). NNR fell from 59 to 280 under the one-sided gate to 2.7–14.9. Taken at face value, this suggests that reliability identifies patients that confidence misses. The next analyses show that the disjointness is a property of the one-sided comparator and that the Brier advantage does not survive the symmetric one.

Table 3. Three deferral rules at 80% coverage. Brier score and AUROC are computed on retained patients. NNR is the number of deferred patients per deferred misclassification. *J* is the Jaccard index between each gate’s deferred set and the composite’s deferred set. eICU-H: held-out hospitals.

Cohort/Task	Gate	Brier	AUROC	Retained Prev.	NNR	<i>J</i> vs. <i>R</i>
MIMIC-IV sepsis	One-sided	0.116	0.849	21.0%	151	0.00
	Symmetric	0.059	0.869	10.0%	2.6	0.83
	Composite	0.061	0.870	10.1%	2.7	–
MIMIC-IV mortality	One-sided	0.094	0.822	13.8%	280	0.00
	Symmetric	0.043	0.811	5.3%	3.2	0.76
	Composite	0.045	0.828	5.8%	3.3	–
MIMIC-IV vasopressor	One-sided	0.059	0.711	6.6%	93	0.00
	Symmetric	0.032	0.687	3.4%	7.1	0.65
	Composite	0.033	0.699	3.5%	7.3	–
MIMIC-III sepsis	One-sided	0.094	0.837	14.4%	201	0.00
	Symmetric	0.041	0.785	4.6%	3.0	0.74
	Composite	0.044	0.822	5.5%	3.2	–
MIMIC-III mortality	One-sided	0.095	0.813	14.0%	201	0.00
	Symmetric	0.046	0.776	5.2%	3.4	0.84
	Composite	0.047	0.790	5.6%	3.5	–
MIMIC-III vasopressor	One-sided	0.079	0.709	9.4%	59	0.01
	Symmetric	0.044	0.639	4.7%	5.0	0.54
	Composite	0.051	0.688	5.5%	6.1	–
eICU sepsis	One-sided	0.113	0.826	18.0%	171	0.00
	Symmetric	0.057	0.800	6.7%	2.7	0.81
	Composite	0.057	0.807	6.9%	2.7	–
eICU mortality	One-sided	0.079	0.797	10.7%	95	0.00
	Symmetric	0.038	0.750	4.4%	4.1	0.76
	Composite	0.039	0.749	4.5%	4.2	–
eICU vasopressor	One-sided	0.038	0.710	4.1%	160	0.00
	Symmetric	0.019	0.654	2.0%	11.0	0.64
	Composite	0.022	0.692	2.3%	12.5	–
eICU-H sepsis	One-sided	0.111	0.837	18.3%	200	0.00
	Symmetric	0.055	0.820	7.0%	2.8	0.91
	Composite	0.056	0.829	7.3%	2.8	–
eICU-H mortality	One-sided	0.079	0.801	10.9%	122	0.00
	Symmetric	0.038	0.749	4.3%	4.1	0.85
	Composite	0.039	0.761	4.4%	4.2	–
eICU-H vasopressor	One-sided	0.037	0.728	3.9%	99	0.02
	Symmetric	0.019	0.656	2.0%	11.4	0.48
	Composite	0.024	0.702	2.5%	14.9	–

Bold: symmetric gate statistically significantly better than the composite (Holm-adjusted paired-bootstrap $p < 0.05$). CIs for every pair are in Table S2. One-sided NNR rests on few deferred errors (for MIMIC-IV sepsis, 13 among 1958 deferred patients) and is unstable.

3.2. The One-Sided and Reliability Gates Defer Opposite Tails of Predicted Risk

Figure 2 shows the calibrated risk of the patients each gate deferred on MIMIC-IV. The one-sided gate deferred the lowest predicted risks (median $\tilde{p}$ 0.008, <0.001 , and 0.008 for sepsis, mortality, and vasopressor), where almost no events occurred (deferred prevalence 0.7%, 0.4%, and 1.1%). The composite deferred patients were higher in the distribution (median $\tilde{p}$ 0.44, 0.33, and 0.12; deferred prevalence 44.2%, 32.7%, and 13.1%), as did the symmetric gate (0.44, 0.33, and 0.14; 44.7%, 34.4%, and 13.7%). A Jaccard index near zero

between the one-sided and reliability gates is therefore a geometric consequence of where they cut, not evidence that they capture different kinds of unreliability.

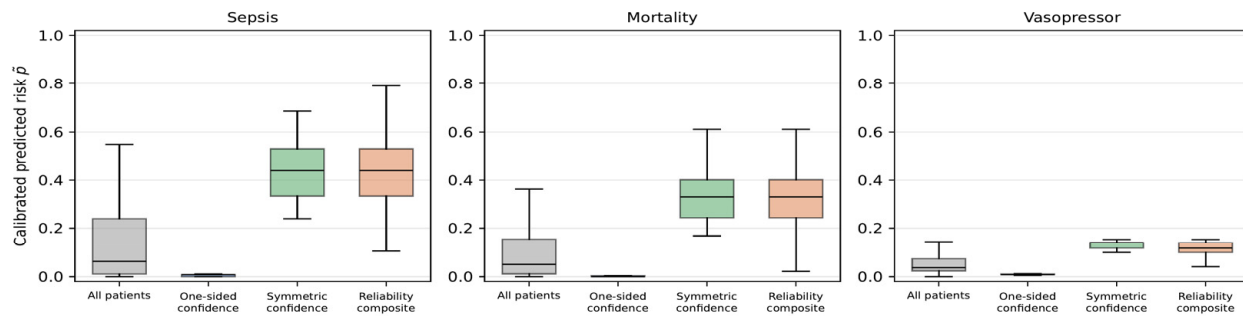


Figure 2. Calibrated predicted risk of the patients deferred by each gate at 20% deferral (MIMIC-IV test set). The grey box shows all patients. Boxes span the interquartile range, and whiskers extend to $1.5 \times$ the interquartile range. The one-sided gate defers the lowest-risk patients; the symmetric and reliability gates defer largely the same higher-risk patients. - Grey: all patients (reference). Blue: deferred by the one-sided confidence gate. The box is a thin sliver near $\tilde{p} \approx 0$, because this gate defers the lowest-risk patients. Green: deferred by the symmetric confidence gate. Orange: deferred by the reliability composite.

The same geometry explains the risk–coverage curves (Figure 3). As coverage fell from 100% to 50%, the one-sided gate’s retained-set Brier score rose (sepsis 0.094 to 0.166, mortality 0.076 to 0.137, vasopressor 0.049 to 0.080), and its retained prevalence rose with it (16.9% to 31.6%, 11.1% to 20.8%, and 5.5% to 9.0%), as Equation (2) predicts: the rule removes patients whose expected squared error is near zero. The symmetric and reliability gates retained increasingly low-risk populations instead.

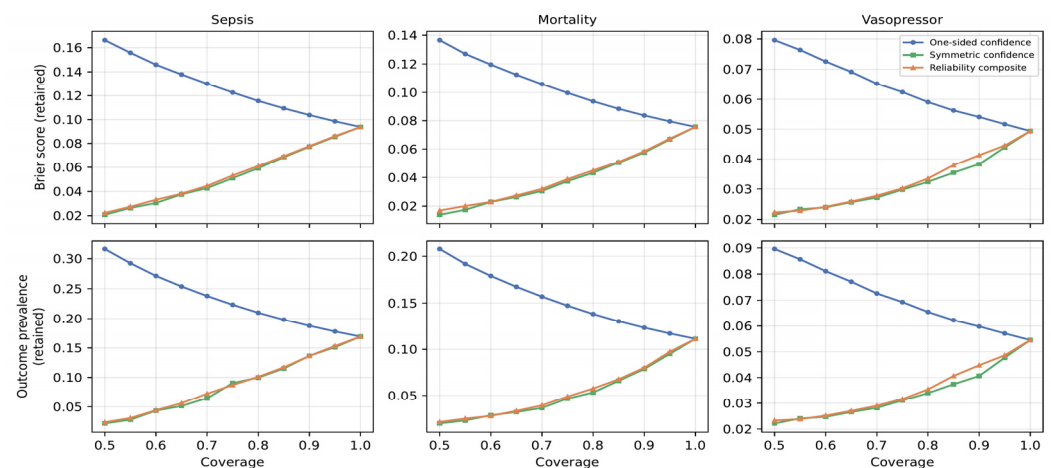


Figure 3. Risk–coverage curves on MIMIC-IV. (**Top**): Brier score of retained predictions. (**Bottom**): outcome prevalence among retained patients. The one-sided gate’s Brier score rises as coverage falls because it retains an increasingly high-risk population. The symmetric and reliability gates retain increasingly low-risk populations and track each other closely.

3.3. A Symmetric Confidence Rule Matches or Exceeds the Composite

Against the symmetric gate, the composite offered no advantage in any pair (Table 3). Retained-set Brier scores did not differ statistically significantly in seven of 12 pairs, and the symmetric gate was better in the other five: MIMIC-IV sepsis (difference -0.0018 ; 95% CI -0.0030 to -0.0006), MIMIC-III sepsis (-0.0033 ; -0.0055 to -0.0012), MIMIC-III vasopressor (-0.0069 ; -0.0104 to -0.0038), eICU vasopressor (-0.0024 ; -0.0038 to -0.0011), and eICU-H vasopressor (-0.0047 ; -0.0064 to -0.0031). The composite was never better: across five split seeds, it was statistically significantly better in none of 60 runs

and worse in 25 (Table S12). The two gates deferred largely the same patients (Jaccard median 0.76, range 0.48–0.91), whereas the symmetric and one-sided gates were disjoint (Jaccard 0.00 in all 12 pairs).

The composite’s retained AUROC was higher than the symmetric gate’s in 11 of 12 pairs (statistically significant after correction in nine). This does not indicate better deferral. At these prevalences the symmetric gate removes more of the highest-risk patients, among whom positives are concentrated (its retained prevalence was lower than the composite’s in all 12 pairs), so the population it keeps spans a narrower range of risk and is harder to rank. Retained-subset AUROC, like retained-subset Brier score, depends on whom each gate keeps. Section 3.7 returns to this point.

3.4. What the Composite Learns

The error model relied mainly on confidence. D1 accounted for 44–85% of the model’s split gain across the 12 pairs; D2–D5 shared the rest roughly equally (3–15% each). The composite score correlated with $\tilde{p}(1 - \tilde{p})$ at Spearman ρ from -0.64 to -0.96 . Removing D2–D5 entirely (D1 alone) gave a lower retained Brier score in all nine primary pairs (point estimates), and the full model predicted held-out squared error worse than D1 alone in all nine (mean squared error 0.6–5.4% higher).

Where the composite departed from the symmetric gate, the departure cost accuracy (Table 4). The patients deferred only by the composite carried less error than those deferred only by the symmetric gate in all three MIMIC-IV tasks (squared error 0.134 vs. 0.210 for sepsis, 0.126 vs. 0.174 for mortality, and 0.088 vs. 0.108 for vasopressor), and their data sufficiency and similarity were similar (Table 4). Split-gain shares, rank correlations, and held-out error-prediction results are in Table S12. The dimensions changed which patients were reviewed at the margin, without directing review toward errors.

Table 4. Patients deferred by only one of the two gates (MIMIC-IV, 20% deferral). Error is misclassification at 0.5; squared error is $(\tilde{p} - y)^2$. D5 is shown as the perturbation variance ($\times 10^{-3}$). The two groups are equal in size by construction.

Task	Deferred Only By	<i>n</i>	Error	Sq. Error	Median $\tilde{p}$	D2	D3	D5
Sepsis	Composite	185	15.1%	0.134	0.24	0.78	0.32	9.8
	Symmetric	185	29.2%	0.210	0.24	0.81	0.35	12.2
Mortality	Composite	266	14.7%	0.126	0.15	0.79	0.35	13.1
	Symmetric	266	22.2%	0.174	0.17	0.79	0.30	13.7
Vasopressor	Composite	295	9.8%	0.088	0.07	0.79	0.29	1.9
	Symmetric	295	12.5%	0.108	0.12	0.80	0.30	2.2

3.5. At Matched Risk, No Score Showed a Detectable Deferral Improvement Beyond Confidence

The risk-matched test removed case mix from the comparison (Figure 4; Table 5; full results in Table S6). The oracle gained $27.9\text{--}54.4 \times 10^{-3}$ over random deferral. The positive controls were detected in all 15 setting–task combinations, but they are strong: with noise twice the standard deviation of the error ($k = 2$), the control gained $17.3\text{--}36.4 \times 10^{-3}$ over the symmetric gate (60–93% of the oracle gain), and with $k = 4$, 31–57%. The symmetric gate itself gained $0.7\text{--}4.8 \times 10^{-3}$ over random deferral, or 2.4–10.7% of the oracle gain, confirming that the bins held risk nearly constant.

No score showed a statistically detectable improvement in deferral beyond the symmetric gate. Across the 75 planned tests (54 confirmatory), none showed a positive Δ_{adj} after Holm correction, and 20 of 75 point estimates were positive, all with CIs spanning zero. The learned error model *L*, given the entire patient record, gained between -1.9 and

$+1.1 \times 10^{-3}$, with CIs spanning zero in 14 of 15 combinations. Adding the dimensions to L did not help (L + D minus L: -0.9 to $+1.0 \times 10^{-3}$; smallest adjusted $p = 0.28$). The continuous sufficiency measure D2* did not exceed the symmetric gate in any setting (Table S5). D3^m and D5^m fell below the symmetric gate in several settings (seven tests with adjusted $p < 0.05$). These scores ignore $\tilde{p}$ and so forgo the symmetric gate's residual within-bin advantage; relative to random deferral, they ranged from -1.8 to $+2.2 \times 10^{-3}$ (CI excluding zero in 8 of 30 tests above and 1 below, unadjusted).

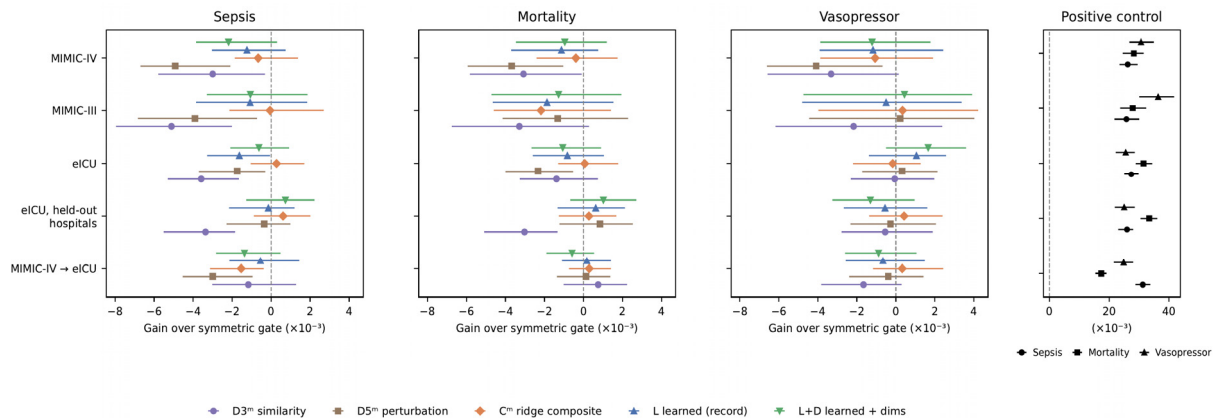


Figure 4. Gain over the symmetric confidence gate at matched predicted risk (Δ_{adj} , $\times 10^{-3}$; points with 95% paired-bootstrap CIs), by task and setting. Positive values would mean that a score defers more error than the symmetric gate among patients with the same predicted risk. **(Right):** the positive control ($k = 2$), an error signal of known strength, on its own scale. MIMIC-IV $\rightarrow$ eICU: MIMIC-IV pipeline applied unchanged to eICU test patients.

Table 5. Risk-matched test: gain over the symmetric gate at matched predicted risk (Δ_{adj} , $\times 10^{-3}$) across 15 setting–task combinations (MIMIC-IV, MIMIC-III, eICU, eICU-H, and MIMIC-IV $\rightarrow$ eICU; three tasks each), $K = 10$ bins.

Score	Gain ^c	CI Spans 0	Range of Δ_{adj}
D3 ^m , missingness-aware similarity	0/15	9/15	−5.1 to +0.8
D5 ^m , missing-value perturbation	0/15	8/15	−4.9 to +0.8
C ^m , ridge composite	0/15	14/15	−2.2 to +0.6
L, error model on full record	0/15	14/15	−1.9 to +1.1
L + D, L plus all dimensions	0/15	15/15	−2.2 to +1.7
D2*, imputation uncertainty ^a	0/15 ^b	5/15	−5.3 to +0.9
Positive control ($k = 2$)	15/15 ^b	0/15	+17.3 to +36.4

^a Descriptive analysis. ^b CI excludes zero (no multiplicity adjustment). ^c Number of combinations with $\Delta_{adj} > 0$ and Holm-adjusted $p < 0.05$; Holm correction was applied within each setting across the five primary scores and three tasks.

Three post hoc analyses bound this result (Table S14). First, titrated controls show what the test can detect: signals that gained about $4\text{--}7 \times 10^{-3}$ over the symmetric gate ($k = 8$) were detected in 70–100% of injections, and signals of about $0\text{--}2 \times 10^{-3}$ ($k = 16$) in at most 50%, so gains of the size of the symmetric gate's own within-bin advantage could go undetected. Second, the learned model is able to find error information when it exists: given a planted feature carrying such information, L gained $25.7\text{--}30.4 \times 10^{-3}$ over the symmetric gate, nearly as much as the feature itself. Third, a likelihood-ratio test using all patients found outcome information beyond a flexible function of $\tilde{p}$ for 6 of the 60 score–setting combinations within databases and at held-out hospitals (Holm-adjusted $p < 0.05$). The scores therefore carried some information that the risk model had not used, but too little to improve deferral by a detectable amount.

The first, random-referenced contrast gave the same picture. None of D2–D5 as defined in Section 2.4 exceeded random deferral within bins in any of the 12 pairs after correction. The composite exceeded it in three pairs and a ridge composite in five, but by no more than the symmetric gate's own within-bin gain, which is why the plan made Δ_{adj} primary. With $K = 20$ bins, no primary score had a CI above zero either; the only contrast whose CI excluded zero was the unadjusted shift contrast for mortality ($+2.3 \times 10^{-3}$; 95% CI $+0.3$ to $+4.5$; Holm-adjusted $p = 0.07$; Table S6c).

3.6. Under Site and Database Shift

Reliability signals are expected to matter most when a model is deployed where it was not trained. At held-out eICU hospitals, calibration held (M-ECE 0.5–0.9%), and L gained -0.1 , $+0.6$, and -0.6×10^{-3} over the symmetric gate for sepsis, mortality, and vasopressor, with all CIs spanning zero. Transfer from MIMIC-IV to eICU degraded calibration (M-ECE 1.2%, 19.7%, and 2.6%, versus 0.4–0.7% for models trained on eICU), yet L gained -0.5 , $+0.2$, and -0.7×10^{-3} , again with CIs spanning zero; for mortality, the 95% CI was -1.1 to $+1.4 \times 10^{-3}$. The gain was not larger after transfer than within eICU (difference in Δ_{adj} , $K = 10$: adjusted $p = 0.77$ for each task). Without risk matching, the ridge composite C^m retained a lower Brier score than the symmetric gate after transfer for mortality (0.0965 vs. 0.0998 at 80% coverage), the setting in which $\tilde{p}$ was most miscalibrated, whereas L did not (0.1023; Table S6e). At matched predicted risk, C^m did not exceed the symmetric gate in that setting ($+0.3 \times 10^{-3}$; CI -0.7 to $+1.4 \times 10^{-3}$), so its whole-cohort advantage reflected which patients it retained rather than errors it identified among patients at the same predicted risk.

The transfer was dominated by differences in how the two databases record data rather than in whom they treat. eICU charts vital signs about ten times as often (median heart-rate observations in 24 h: 284 vs. 26), lacks the Fahrenheit temperature signal and records Celsius temperature in only 11% of stays, and measures lactate less often (27.7% vs. 59.2% of stays with at least one value). These differences apply to every eICU patient. They shift the predicted risk of whole groups of patients together, which degrades calibration; the null result for L is consistent with their not making some patients at a given predicted risk more error-prone than others, which is what deferral could exploit. The transfer used the features as extracted, without harmonizing units or charting frequency, and therefore represents an unharmonized deployment. Post hoc, the conclusions held with a hospital-cluster bootstrap (L's CI spanned zero in all six cluster-resampled settings; Table S14b). The likelihood-ratio test found outcome information beyond any monotone function of $\tilde{p}$ in 7 of 15 transfer tests (Holm-adjusted $p < 0.05$), most strongly for L in mortality, so the miscalibration after transfer was not uniform across patients. Yet recalibrating the transferred model on eICU validation patients reduced M-ECE for mortality from 19.7% to 0.5% and the symmetric gate's retained Brier score from 0.0998 to 0.0494, far more than C^m achieved without recalibration (0.0965; difference from the uncorrected symmetric gate -0.0033 , 95% CI -0.0042 to -0.0025 ; Table S14d). Removing the features whose meaning differs between the databases (observation counts and temperature) reduced the mortality calibration error after transfer from 19.7% to 5.0%, and with harmonized features no score improved deferral at matched risk (0 of 30 tests; Table S15c). Across held-out eICU hospitals, the difference between the symmetric gate and the composite showed no heterogeneity ($I^2 = 0$ for all tasks) and favored the symmetric gate in 36 of 46 hospital-task comparisons, although calibration in the large varied between hospitals (sepsis: -0.09 to $+0.15$; Table S15a).

3.7. Retained-Subset Metrics Depend on Whom Each Gate Keeps

Whole-cohort metrics computed on retained patients answer different questions for different gates. On the retained-set Brier skill score, which scales the Brier score by the retained prevalence, the one-sided gate ranked above the symmetric gate in 11 of 12 pairs (CIs excluding zero; the exception was MIMIC-IV sepsis). The one-sided gate keeps a higher-prevalence population, against which the same absolute error is a smaller fraction of the outcome variance; the symmetric gate keeps the patients whose predictions carry the least expected error (Table S10). Which is preferable depends on what deferral is for. When its purpose is to send the clinician the predictions most likely to be wrong, the relevant quantity is the error left with the model, the retained Brier score, which the symmetric rule minimizes by Equation (2).

Net benefit on retained patients showed the same dependence (Figure 5). At $p_t = 0.10$ on MIMIC-IV, the one-sided gate's net benefit (0.151, 0.079, and 0.011) equaled the no-deferral net benefit divided by 0.8 (0.121, 0.064, and 0.009), because the patients it deferred lay below the threshold (maximum deferred $\tilde{p}$ 0.011, 0.005, and 0.024) and contribute neither true nor false positives. The symmetric and reliability gates (0.055 and 0.056 for sepsis) remove the higher-risk patients on whom threshold decisions turn. A valid comparison of net benefit across gates requires a model of what happens to deferred patients, which requires data on clinician decisions that we do not have. For decisions at a fixed threshold t between 0.05 and 0.50, the relevant gate defers the predictions with the highest expected decision loss at t ; the composite's retained decision loss was higher than that gate's in 8–11 of 12 pairs at each threshold and lower in none (Table S15b).

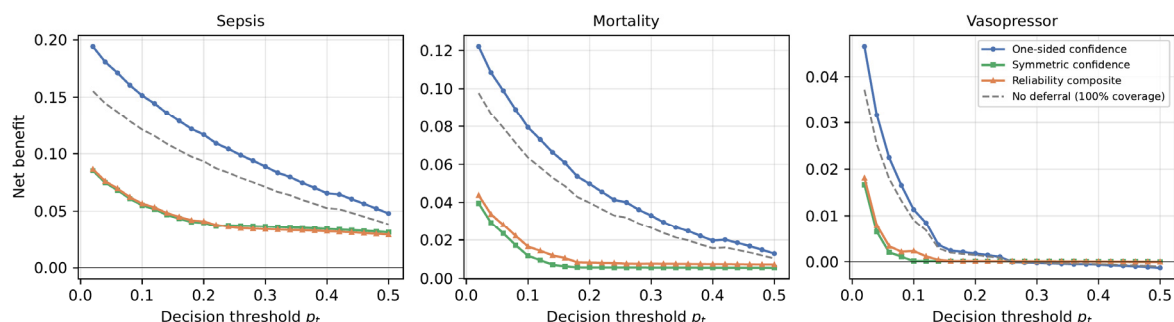


Figure 5. Net benefit on retained patients at 80% coverage, MIMIC-IV. The dashed line is the calibrated model applied to all patients without deferral. The one-sided curve equals the no-deferral curve divided by 0.8, because the deferred patients lie below every threshold.

3.8. Alternative Methods, Base Models, Labels, and Stress Tests

Every score that approximates distance from 0.5 reached the lowest retained Brier scores on MIMIC-IV (Table 6): symmetric confidence, temperature-scaled entropy, and, within 0.005, the composite and ensemble variance. Monte Carlo dropout variance and Bayesian logistic regression variance, which do not scale with $\tilde{p}(1 - \tilde{p})$, were worse. With the MLP base model, the symmetric gate again matched or exceeded the composite (Brier difference CI, sepsis -0.0020 to 0.0005 ; mortality -0.0042 to -0.0008 ; vasopressor -0.0019 to 0.0029), and SelectiveNet was within 0.006 of the symmetric gate, lower for vasopressor (0.033 vs. 0.035) and higher for sepsis and mortality, without CIs.

With Sepsis-3 labels (MIMIC-IV prevalence 49.8%), the symmetric gate was slightly better than the composite (retained Brier 0.156 vs. 0.157; difference CI -0.0019 to -0.0001) and the one-sided gate was worse (0.194; Table S8). Sepsis-3 onset occurred within the first 24 h in 47% of MIMIC-IV stays; restricting the task to onset after the prediction time (prevalence 4.9%), no score improved deferral at matched risk, and the symmetric gate again matched the composite (retained Brier 0.0273 vs. 0.0285; difference CI -0.0036 to

+0.0014; Table S15d). First-day SOFA discriminated mortality less well than the XGBoost model (AUROC 0.728 vs. 0.858), and the symmetric gate lowered its retained Brier score from 0.090 to 0.067. In eICU, APACHE IVa and the XGBoost model performed similarly (AUROC 0.828 vs. 0.822; symmetric retained Brier 0.041 vs. 0.040). Under the perturbation stress tests, the symmetric gate’s retained Brier score was equal to or lower than the composite’s at every magnitude in all three tasks (Table S3).

Table 6. Brier score of retained predictions at 80% coverage, MIMIC-IV.

Base Model	Deferral Rule	Sepsis	Mortality	Vasopressor
XGBoost	One-sided confidence	0.116	0.094	0.059
	Symmetric confidence	0.059	0.043	0.032
	Temperature-scaled entropy	0.059	0.043	0.032
	Reliability composite	0.061	0.045	0.033
	Ensemble variance (conformal)	0.064	0.046	0.033
	Monte Carlo dropout variance	0.071	0.048	0.037
	Bayesian logistic regression	0.091	0.068	0.042
MLP	One-sided confidence	0.118	0.094	0.059
	Symmetric confidence	0.062	0.044	0.035
	Reliability composite	0.063	0.047	0.035
	SelectiveNet (end-to-end)	0.068	0.048	0.033

No deferral (100% coverage): 0.094, 0.076, and 0.049.

3.9. Clinical Operation and Subgroups

Deferral helps only if reviewers decide deferred cases better than the model. At 80% coverage on MIMIC-IV, the model misclassified 38%, 31%, and 14% of the patients deferred by the symmetric gate (sepsis, mortality, vasopressor), so review adds errors unless clinicians are more than 62%, 69%, and 86% accurate on these cases; under the one-sided gate, this break-even accuracy was at least 98.9% (Table S9). The symmetric gate’s retained set had higher specificity and lower sensitivity at 0.5 than no deferral (sepsis: specificity 0.990 vs. 0.946; sensitivity 0.402 vs. 0.493), because it defers the ambiguous predictions near 0.5 (Table S7).

Under the symmetric gate, the largest spread was across age bands (up to 14.4 percentage points, MIMIC-III mortality) and race/ethnicity groups (up to 13.6 points, MIMIC-III vasopressor); between sexes it was at most 4.0 points. The composite’s spreads were similar, and the one-sided gate had the largest age spread overall (32.5 points, eICU-H mortality). Error capture, retained Brier score, and inappropriate deferral by subgroup are in Table S4.

4. Discussion

4.1. Principal Findings

We asked whether a multidimensional reliability score is a better basis for clinical deferral than model confidence. Against the one-sided gate used in the earlier preprint, it appeared to be deferred sets were nearly disjoint, and retained predictions had 1.5- to 2.1-fold lower Brier scores, in three databases and at held-out hospitals. Against the symmetric confidence rule, the advantage disappeared. The composite never outperformed the symmetric rule, which was better in five of 12 database–task pairs, and the two rules deferred largely the same patients. The risk-matched test then asked the question directly: among patients with the same predicted risk, does any signal identify the predictions that are wrong well enough to improve deferral? None did detectably. Missingness-aware reliability dimensions, a ridge composite, and a gradient-boosted model of error given the entire patient record failed to exceed the symmetric rule in any of 75 planned tests (54 confirmatory), including at held-out hospitals and after transfer from MIMIC-IV to

eICU, and a continuous data-sufficiency measure failed in all 15 setting–task combinations. Positive controls were detected in every combination, and titrated controls showed that the test reliably detects gains of about $4\text{--}7 \times 10^{-3}$ in retained Brier score.

4.2. Why Confidence Is Hard to Beat

Equation (1) explains both the whole-cohort and the risk-matched results. For a calibrated model, expected squared error depends on predicted risk through $\tilde{p}(1 - \tilde{p})$, and among patients with the same $\tilde{p}$ it varies only with the true risk $\eta(\mathbf{x})$. A reliability score can therefore improve on confidence only by knowing something about the outcome that the risk model does not. When the score is computed from the same record the model was trained on, that information has largely been used already. Missingness is a clear example: the XGBoost model was trained with missing values left in place, so the informativeness of an absent measurement [41] enters $\tilde{p}$ directly, and the missingness-aware similarity D3^m and the imputation measure D2* added too little to change deferral. The learned model L is the strongest form of the argument. It was given every feature and fitted to predict error; the upper 95% confidence limit of its gain over the symmetric rule was at most 3.4×10^{-3} in every setting (at most 8.7% of the oracle's gain over random deferral), yet it recovered a planted error signal almost completely. Some scores did carry statistically detectable outcome information beyond $\tilde{p}$, but the deferral gains they produced were small (point estimates at most 1.7×10^{-3}), below the $4\text{--}7 \times 10^{-3}$ that the titrated controls show the test detects reliably and an order of magnitude below the oracle's gain ($27.9\text{--}54.4 \times 10^{-3}$).

This is consistent with work outside clinical outcome prediction. In image classification, plain softmax confidence outperformed purpose-built confidence scores once failure detection was evaluated consistently [25], and in wearable electrocardiogram quality assessment, posterior-derived uncertainty did not improve on maximum class probability [26]. Our contribution is to test the question at matched predicted risk, in clinical outcome prediction, and under site and database shift.

4.3. Shift Degraded Calibration More than It Created Rankable Unreliability

Reliability signals are usually motivated by distribution shift [19,21], so the transfer from MIMIC-IV to eICU was the setting most favorable to them. The transfer degraded calibration severely for mortality (M-ECE 19.7%), yet no score produced a statistically detectable improvement in deferral at matched risk. The shift came mainly from how the databases record data: vital signs charted ten times as often, a temperature signal absent, and lactate measured in fewer than half as many stays. These differences affect every patient. They move the predicted risk of whole groups together, and a per-patient ranking cannot find an error that every patient at a given risk shares. The miscalibration was not entirely uniform: after transfer, several scores carried outcome information beyond any monotone function of $\tilde{p}$. But that structure was too small to improve deferral detectably, whereas recalibration at the target site halved the retained Brier score for mortality, and harmonizing the features alone reduced the calibration error from 19.7% to 5.0%. In the terms of external-validation research, the transfer tested transportability rather than reproducibility: the meaning of several predictors changed with the documentation system [42]. The corresponding remedy is model updating and recalibration at the new site [43], with ongoing performance monitoring [44–46], not deferral of individual patients. Held-out eICU hospitals, which share documentation practices, showed little calibration loss (M-ECE 0.5–0.9%) and no reliability gain either.

4.4. Why the Preprint's Comparison Misled

Three features of the preprint's design combined to produce the appearance of a reliability effect. First, the comparator was one-sided. At 3–17% prevalence, it spent the deferral budget on patients who were almost certainly negative and correctly predicted (deferred prevalence 0.4–1.1% on MIMIC-IV). Second, the primary metric was computed on retained patients, whose case mix differed sharply between gates (retained sepsis prevalence 21.0% vs. 10.1% on MIMIC-IV). By Equation (2), a low-prevalence retained set has a low Brier score whether or not anything about reliability has been learned. Third, the composite was trained to predict per-patient error, a target whose expectation under calibration is $\tilde{p}(1 - \tilde{p})$; the model that predicts it best recovers confidence. The composite used here was cross-fitted to avoid in-sample leakage and kept confidence as an explicit input, and it still drew 44–85% of its split gain from confidence and predicted held-out error worse than confidence alone. None of these features is unusual. One-sided thresholds on predicted risk are natural in alerting systems, subset metrics are the default in selective prediction, and error-predicting deferral models are common.

The vocabulary of the preprint also blurred three distinct quantities: predicted risk ($\tilde{p}$), confidence (distance from 0.5), and expected error ($\tilde{p}(1 - \tilde{p})$ under calibration). The one-sided gate was labeled a confidence gate but ranks by predicted risk. Keeping the three apart removes most of the apparent paradoxes of the preprint's analysis.

4.5. Implications for Evaluating Selective Prediction

Four practices would have exposed the problem at once, and we recommend them for clinical deferral studies generally:

1. Benchmark against the symmetric confidence rule, or temperature-scaled entropy when the base model is uncalibrated. A proposed score has to beat the distance from 0.5, not a one-sided threshold. For decisions at a threshold $t \neq 0.5$ with asymmetric costs, the analogous comparator defers the highest expected decision loss at t , which the composite did not match (Table S15b).
2. Test candidate signals at matched predicted risk, against the symmetric rule, with positive controls of graded strength that show the smallest effect the test can detect. Whole-cohort comparisons cannot separate a score's information about error from the case mix it retains.
3. State what deferral is for, and choose the metric accordingly. If the purpose is to send the clinician the predictions most likely to be wrong, the error left with the model, the retained Brier score, is the relevant quantity. Retained-subset discrimination, Brier skill, and net benefit answer other questions and can rank gates in the opposite order (Section 3.7), because the Brier score depends on prevalence [23,24] and selective metrics must match the task [47]. Report retained and deferred prevalence next to every subset metric.
4. Treat overlap statistics as descriptive. A Jaccard index near zero says that two rules cut the risk distribution in different places, not that either rule is better.

Recent clinical and biomedical selective-prediction studies evaluate deferral by the error of retained cases at fixed coverage [48,49]; adding a symmetric-confidence comparator and a risk-matched analysis would show how much of the reported benefit is specific to the proposed method.

4.6. Deferral in ICU Practice

At 80% coverage, deferral sends one in five predictions to a clinician: with one prediction per ICU stay at 24 h, that is 20 reviews per 100 admissions. Our analysis treated review as an independent decision whose accuracy does not depend on the case. In practice,

deferred predictions under the symmetric rule are the ambiguous ones, on which clinicians may also be least accurate, and the break-even accuracies we found (62–86% on MIMIC-IV) are therefore not trivially met. Deferral should also not delay care: a deferred prediction is better presented as “model uncertain, use clinical judgment” alongside routine assessment than withheld pending review, so that a late or missed review leaves the patient no worse off than without the model. Whether a deferral rule improves decisions, rather than retained-subset metrics, can only be established prospectively, with the reviewers’ accuracy and workload measured.

4.7. What Remains for Multidimensional Reliability

These results do not show that data sufficiency, cohort similarity, or temporal stability are clinically irrelevant. They show that, computed from the record the model already uses, these signals did not identify erroneous predictions beyond confidence. Where the composite departed from the symmetric rule, it deferred patients with less error and nearly identical dimension values (Table 4). Three roles remain open. The first is information from outside the model’s inputs, such as documentation delays, device artifacts, or a clinician’s concern, which the risk model cannot have absorbed. The second is modeling the reviewer: learning-to-defer methods gain their advantage from the human expert’s accuracy [15–17], which confidence cannot capture, and our break-even analysis shows that the value of any deferral rule depends on it. The third is explanation: telling a clinician why a deferred prediction is uncertain, a role that needs prospective evaluation with clinicians rather than retrospective Brier scores.

4.8. Relation to Prior Work

Our findings agree with the classical result that, for a calibrated binary classifier, confidence-based rejection minimizes expected loss [12], and they extend it empirically to the setting in which richer signals are most often proposed: distribution shift. The literature on confidence scoring has already found that new scores rarely beat softmax confidence under consistent evaluation [25,26], and that selective-classification metrics can mislead when they do not match the task [47]. The clinical-epidemiology literature on external validation makes the parallel point that apparent performance differences between samples often reflect case mix rather than model quality [42]. The retained sets of different deferral rules are different samples in exactly this sense.

4.9. Limitations

All three databases are from the United States, and two are from one institution, with overlapping periods (2008–2012). Because MIMIC dates are shifted per patient, no split was temporal. Sepsis labels were derived from discharge codes or admission diagnosis strings, which may record sepsis already present at the time of prediction. Because every gate ranks the same predictions, such label limitations affect the gates equally, but they make the sepsis task partly one of recognizing existing sepsis. In MIMIC-IV, Sepsis-3 labels and Sepsis-3 onset after the prediction time gave the same result; we did not repeat these analyses in MIMIC-III or eICU. The vasopressor task had low prevalence and moderate discrimination, so few of its predictions exceeded 0.5. The MIMIC-IV→eICU setting combines measurement and label differences; harmonizing the features removed most of the calibration loss for mortality, but label differences remain. The held-out-hospital split produced little shift, so other kinds of shift, such as a new device or assay affecting a subset of patients, could create patient-level unreliability that our settings did not contain. The risk-matched analysis plan was amended after the first, random-referenced results were seen; the learned scores are exploratory in the three in-distribution settings, and confirmatory only in the two shift settings, and D5^m is exploratory in MIMIC-IV. The

negative result is bounded rather than absolute: gains larger than 3.4×10^{-3} in squared error per retained patient were excluded for L, but smaller gains were not; titrated controls were detected reliably only above about $4\text{--}7 \times 10^{-3}$, and no equivalence margin was specified in advance. Recalibration used target-site labels, which a deployment may not have at first. Inference resampled patients; a hospital-cluster bootstrap for the eICU settings is reported as a sensitivity analysis (Table S14). The primary base model was XGBoost; the MLP replication covered MIMIC-IV only. Finally, no clinician evaluated any deferral, and the review analysis assumed that reviewer accuracy does not depend on the case.

5. Conclusions

A reliability composite for ICU prediction appeared to outperform confidence-based deferral by a wide margin in three databases. The margin came from comparing against a one-sided rule that defers the lowest predicted risks. Against the symmetric confidence rule, which among rules based on predicted risk minimizes the expected Brier score of retained predictions for a calibrated model, the composite offered no advantage. At matched predicted risk, no reliability signal computed from the patient record, including a model of error given the entire record, showed a statistically detectable improvement in deferral beyond confidence within databases, at held-out hospitals, or after transfer between databases; gains above a few thousandths of a Brier point were excluded, although some scores carried small, statistically detectable information beyond predicted risk. After transfer between databases, recalibration at the target site improved retained predictions far more than any reliability score. Before attributing a gain in selective prediction to a richer uncertainty signal, studies should compare against distance from 0.5 at matched predicted risk and report whom each rule retains; under shift, recalibration is the first remedy.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/bioengineering13111255/s1>, Table S1: Training diagnostics; Table S2: Paired gate comparisons for all 12 database–task pairs; Table S3 and Figure S1: Test-time perturbation; Table S4: Subgroup audit; Table S5 and Figure S2: The data-sufficiency measure D2*; Table S6: Full risk-matched results, added value of the dimensions, and the shift contrast; Table S7 and Figure S3: Class-conditional performance, confusion matrices, and calibration; Table S8: Sepsis-3 labels and clinical scores; Table S9: Imperfect clinician review; Table S10: Retained-set Brier skill score; Table S11: Analysis plans and deviations; Tables S12 and S13: Secondary results and implementation details; Tables S14 and S15: Post hoc analyses; and the completed TRIPOD + AI reporting checklist [39].

Author Contributions: Y.Y., B.G.C. and A.J.F. conceived the study, developed the reliability framework, implemented the deferral pipeline and validation, and analyzed the data. Y.Y. and B.G.C. wrote the manuscript. S.P., C.R.M. and N.G.W. critically reviewed the manuscript. A.J.F. supervised the research, provided critical review of the reliability framework and deployment model, and contributed to manuscript revision and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Mayo Clinic and the generosity of Schmidt Sciences, Gerstner Philanthropies, and the Florida Cancer Innovation Fund.

Institutional Review Board Statement: This study used existing de-identified data from the MIMIC-III, MIMIC-IV, and eICU Collaborative Research Database, accessed through PhysioNet under the applicable credentialing and data-use requirements. The investigators had no direct interaction with the patients represented in these databases and could not readily ascertain their identities. Under the U.S. Federal Policy for the Protection of Human Subjects (45 CFR §46.102(e)(1)), these secondary analyses did not constitute human-subject research requiring additional Institutional Review Board approval. Accordingly, additional IRB approval and individual patient informed consent were

not required. The study was conducted in accordance with the applicable ethical principles of the Declaration of Helsinki.

Informed Consent Statement: Individual patient informed consent was not required because the study used only existing de-identified data (see Institutional Review Board Statement). There was no patient or public involvement in the design, conduct, reporting, or dissemination of this research.

Data Availability Statement: The data presented in this study were derived from the MIMIC-IV, MIMIC-III, and eICU Collaborative Research databases, which are available through PhysioNet to credentialed users under Data Use Agreements and are not publicly redistributable in raw form. Processed feature matrices and per-patient test predictions are available from the corresponding author upon reasonable request, subject to the same credentialing requirements. Analysis code, tests, and the dated analysis plans for the reliability composite and the risk-matched test are available at <https://github.com/yuyunguo/reliability-deferral> (accessed on 17 October 2026).

Acknowledgments: The authors thank the MIT Laboratory for Computational Physiology for providing the MIMIC-IV and MIMIC-III databases, and the eICU Collaborative Research Database contributors for providing eICU access via PhysioNet. An earlier version of this work was posted as a preprint [22]. During the preparation of this manuscript, the authors used Claude Opus 5.5 (Anthropic PBC, San Francisco, CA, USA) for assistance with analysis code, manuscript editing, and LaTeX formatting. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AUROC	Area Under the Receiver Operating Characteristic Curve
APACHE	Acute Physiology and Chronic Health Evaluation
BIDMC	Beth Israel Deaconess Medical Center
CI	Confidence Interval
ICD	International Classification of Diseases
ICU	Intensive Care Unit
M-ECE	Expected Calibration Error with Equal-Mass Bins
MIMIC	Medical Information Mart for Intensive Care
MLP	Multilayer Perceptron
NNR	Number Needed to Review
SOFA	Sequential Organ Failure Assessment
SD	Standard Deviation
TRIPOD + AI	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis, Extended for Artificial Intelligence

References

1. Johnson, A.E.W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. *Sci. Data* **2023**, *10*, 1. [[CrossRef](#)] [[PubMed](#)]
2. Veldhuis, L.I.; Woittiez, N.J.C.; Nanayakkara, P.W.B.; Ludikhuizen, J. Artificial Intelligence for the Prediction of in-Hospital Clinical Deterioration: A Systematic Review. *Crit. Care Explor.* **2022**, *4*, e0744. [[CrossRef](#)] [[PubMed](#)]
3. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and Accurate Deep Learning with Electronic Health Records. *npj Digit. Med.* **2018**, *1*, 18. [[CrossRef](#)] [[PubMed](#)]
4. Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; Dennison, D. Hidden Technical Debt in Machine Learning Systems. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2015; Volume 28.

5. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLoS Med.* **2018**, *15*, e1002683. [[CrossRef](#)] [[PubMed](#)]
6. Wong, A.; Otles, E.; Donnelly, J.P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pestrue, J.; Phillips, M.; Konye, J.; Penzo, C.; et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern. Med.* **2021**, *181*, 1065. [[CrossRef](#)] [[PubMed](#)]
7. Kompa, B.; Snoek, J.; Beam, A.L. Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning. *npj Digit. Med.* **2021**, *4*, 4. [[CrossRef](#)] [[PubMed](#)]
8. Lambert, B.; Forbes, F.; Doyle, S.; Dehaene, H.; Dojat, M. Trustworthy Clinical AI Solutions: A Unified Review of Uncertainty Quantification in Deep Learning Models for Medical Image Analysis. *Artif. Intell. Med.* **2024**, *150*, 102830. [[CrossRef](#)] [[PubMed](#)]
9. Yu, Y. Confidence Is Not Probability: Verification-Based Confidence Prediction for LLM-Generated Clinical Summaries. *Preprints* **2026**. [[CrossRef](#)]
10. Yu, Y.; Gomez-Cabello, C.A.; Haider, S.A.; Genovese, A.; Prabha, S.; Trabilsy, M.; Collaco, B.G.; Wood, N.G.; Bagaria, S.; Tao, C.; et al. Enhancing Clinician Trust in AI Diagnostics: A Dynamic Framework for Confidence Calibration and Transparency. *Diagnostics* **2025**, *15*, 2204. [[CrossRef](#)] [[PubMed](#)]
11. Yu, Y. Backcasting the Trust Gap: A Strategic Road Map for Clinician Adoption of AI Diagnostics by 2040. *J. Med. Internet Res.* **2026**, *28*, e94234. [[CrossRef](#)] [[PubMed](#)]
12. Chow, C.K. On Optimum Recognition Error and Reject Tradeoff. *IEEE Trans. Inf. Theory* **1970**, *16*, 41–46. [[CrossRef](#)]
13. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
14. Geifman, Y.; El-Yaniv, R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning*; Curran Associates Inc.: Red Hook, NY, USA, 2019; Volume 97, pp. 2151–2159.
15. Madras, D.; Pitassi, T.; Zemel, R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2018; Volume 31.
16. Mozannar, H.; Sontag, D. Consistent Estimators for Learning to Defer to an Expert. In *Proceedings of the 37th International Conference On Machine Learning*; Curran Associates Inc.: Red Hook, NY, USA, 2020; Volume 119, pp. 7076–7087.
17. Dvijotham, K.; Winkens, J.; Barsbey, M.; Ghaisas, S.; Stanforth, R.; Pawlowski, N.; Strachan, P.; Ahmed, Z.; Azizi, S.; Bachrach, Y.; et al. Enhancing the Reliability and Accuracy of AI-Enabled Diagnosis via Complementarity-Driven Deferral to Clinicians. *Nat. Med.* **2023**, *29*, 1814–1820. [[CrossRef](#)] [[PubMed](#)]
18. Angelopoulos, A.N.; Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [[CrossRef](#)]
19. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019; Volume 32.
20. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
21. D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M.D.; et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *J. Mach. Learn. Res.* **2022**, *23*, 1–61.
22. Yu, Y. The Deferral Frontier: Reliability-Gated Selective Prediction for Clinical AI. *Res. Sq. Version 1* **2026**. [[CrossRef](#)] [[PubMed](#)]
23. Assel, M.; Sjöberg, D.D.; Vickers, A.J. The Brier Score Does Not Evaluate the Clinical Utility of Diagnostic Tests or Prediction Models. *Diagn. Progn. Res.* **2017**, *1*, 19. [[CrossRef](#)] [[PubMed](#)]
24. Hoessly, L. On Misconceptions about the Brier Score in Binary Prediction Models. *Glob. Epidemiol.* **2026**, *11*, 100242. [[CrossRef](#)] [[PubMed](#)]
25. Jaeger, P.F.; Lüth, C.T.; Klein, L.; Bungert, T.J. A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
26. Tran, K.D. When Can Predictive Uncertainty Be Trusted? A Methodological Evaluation in Free-Living Wearable Electrocardiogram Signal-Quality Assessment. *medRxiv Prepr.* **2026**. [[CrossRef](#)]
27. Johnson, A.; Bulgarelli, L.; Pollard, T.; Gow, B.; Moody, B.; Horng, S.; Celi, L.A.; Mark, R. MIMIC-IV (Version 3.1). *PhysioNet* **2024**. [[CrossRef](#)]
28. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Lehman, L.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a Freely Accessible Critical Care Database. *Sci. Data* **2016**, *3*, 160035. [[CrossRef](#)] [[PubMed](#)]
29. Pollard, T.J.; Johnson, A.E.W.; Raffa, J.D.; Celi, L.A.; Mark, R.G.; Badawi, O. The eICU Collaborative Research Database, a Freely Available Multi-Center Database for Critical Care Research. *Sci. Data* **2018**, *5*, 180178. [[CrossRef](#)] [[PubMed](#)]

30. Singer, M.; Deutschman, C.S.; Seymour, C.W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G.R.; Chiche, J.-D.; Coopersmith, C.M.; et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **2016**, *315*, 801–810. [[CrossRef](#)] [[PubMed](#)]
31. Rhee, C.; Dantes, R.; Epstein, L.; Murphy, D.J.; Seymour, C.W.; Iwashyna, T.J.; Kadri, S.S.; Angus, D.C.; Danner, R.L.; Fiore, A.E.; et al. Incidence and Trends of Sepsis in US Hospitals Using Clinical Vs Claims Data, 2009–2014. *JAMA* **2017**, *318*, 1241–1249. [[CrossRef](#)] [[PubMed](#)]
32. Ledoit, O.; Wolf, M. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *J. Multivar. Anal.* **2004**, *88*, 365–411. [[CrossRef](#)]
33. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 70, pp. 1321–1330.
34. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
35. Brier, G.W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [[CrossRef](#)]
36. Laupacis, A.; Sackett, D.L.; Roberts, R.S. An Assessment of Clinically Useful Measures of the Consequences of Treatment. *N. Engl. J. Med.* **1988**, *318*, 1728–1733. [[CrossRef](#)] [[PubMed](#)]
37. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
38. Vickers, A.J.; Elkin, E.B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med. Decis. Mak.* **2006**, *26*, 565–574. [[CrossRef](#)] [[PubMed](#)]
39. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ* **2024**, *385*, e078378. [[CrossRef](#)] [[PubMed](#)]
40. van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 67. [[CrossRef](#)]
41. Yu, Y. What the Doctor Didn't Order: Informative Missingness as a Clinical Signal in ICU Prediction Models. *Preprint* **2026**. [[CrossRef](#)]
42. Debray, T.P.A.; Vergouwe, Y.; Koffijberg, H.; Nieboer, D.; Steyerberg, E.W.; Moons, K.G.M. A New Framework to Enhance the Interpretation of External Validation Studies of Clinical Prediction Models. *J. Clin. Epidemiol.* **2015**, *68*, 279–289. [[CrossRef](#)] [[PubMed](#)]
43. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Topic Group “Evaluating diagnostic tests and prediction models” of the STRATOS initiative. Calibration: The Achilles Heel of Predictive Analytics. *BMC Med.* **2019**, *17*, 230. [[CrossRef](#)] [[PubMed](#)]
44. Yu, Y. From Monitoring to Action: A Scoping Review of Post-Deployment Clinical AI Surveillance, Risk Assessment, and Lifecycle Management. *OSF Regist.* **2026**. [[CrossRef](#)]
45. Curcin, V.; Delaney, B.; Alkhatib, A.; Cockburn, N.; Dann, O.; Kostopoulou, O.; Leightley, D.; Maddocks, M.; Modgil, S.; Nirantharakumar, K.; et al. Learning Health Systems Provide a Glide Path to Safe Landing for AI in Health. *Artif. Intell. Med.* **2026**, *173*, 103346. [[CrossRef](#)] [[PubMed](#)]
46. Chi, S.; Tian, Y.; Wang, F.; Zhou, T.; Jin, S.; Li, J. A Novel Lifelong Machine Learning-Based Method to Eliminate Calibration Drift in Clinical Prediction Models. *Artif. Intell. Med.* **2022**, *125*, 102256. [[CrossRef](#)] [[PubMed](#)]
47. Traub, J.; Bungert, T.J.; Lüth, C.T.; Baumgartner, M.; Maier-Hein, K.H.; Maier-Hein, L.; Jaeger, P.F. Overcoming Common Flaws in the Evaluation of Selective Classification Systems. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2024; Volume 37.
48. Alshamlan, H. Uncertainty-Aware Selective Prediction in Brain-Tumour MRI Classification: Selective Risk Reduction and Component-Wise Reliability Findings. *Bioengineering* **2026**, *13*, 1184. [[CrossRef](#)]
49. Kwon, H.; Kim, D.-J. Conformal Selective Prediction with Cost Aware Deferral for Safe Clinical Triage Under Distribution Shift. *Sci. Rep.* **2026**, *16*, 10016. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.