Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Retrospective evaluation of transportability and uncertainty in ICU mortality prediction across MIMIC-IV and eICU

Tianyi Yu

tianyi.1.yu@kcl.ac.uk

King's College London

Research Article

Keywords: intensive care, mortality prediction, measurement harmonisation, external evaluation, calibration, conformal prediction, dataset shift, reproducibility

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11173088/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

ICU prediction models must retain clinically meaningful performance across differences in patients, measurement practices and outcome recording. Discrimination alone cannot establish this reliability. We examined model transportability and prediction-set coverage under a common cross-database design.

Methods

In this retrospective study, adults still recorded in the index ICU stay at 24 hours were evaluated for subsequent recorded death before index ICU discharge. Four model families used six vital-sign concepts, common preprocessing and 27,437 training patients per source. Sixteen configurations per family underwent source-only screening followed by three-seed confirmation of two finalists. Mean source-validation AUROC determined selection before final test evaluation. Reciprocal testing, separate probability and conformal calibration, hospital-disjoint eICU partitions, overlap weighting and measurement thinning assessed complementary aspects of reliability.

Results

Cohorts comprised 54,376 MIMIC-IV and 102,581 eICU patients. Source validation selected a Transformer in both directions. Source-test and external AUROCs were 0.776 and 0.720 for MIMIC-IV development, and 0.767 and 0.738 for eICU development. MIMIC-IV-trained gradient boosting had higher external AUROC, 0.743. The selected models overestimated external mortality in one direction and underestimated it in the other. A domain classifier distinguished the databases with AUROC 0.957. Marginal prediction sets covered 88.9% and 91.1% of external outcomes, but only 1.4% and 0.0% of deaths. Class-conditional death coverage was 91.6% and 84.9%, with both labels included for 54.2% and 49.3% of patients. Separate target updating yielded death coverage of 88.9% and 91.4%, with two-label frequencies of 56.1% and 54.8%.

Conclusions

Common inputs and training budgets improved the interpretability of cross-database comparisons but did not establish measurement equivalence or clinical reliability. Source model selection, external probability accuracy and event-specific coverage require separate evaluation. Prediction-set coverage must be interpreted alongside informativeness and the target data required for updating.

Background

An ICU mortality model estimates risk within a particular clinical and data-generating environment. Its usefulness elsewhere depends on more than identifying patients with relatively high scores. Recorded physiology, treatment practices and discharge decisions can differ between hospitals, changing both the available predictors and the outcome being predicted. A model may therefore retain useful ranking while providing misleading numerical risks. Understanding these distinctions is necessary before retrospective performance can inform a clinical evaluation.

MIMIC-IV and the eICU Collaborative Research Database provide complementary resources for examining this problem [1, 2, 3]. MIMIC-IV contains data from one academic medical centre, whereas eICU includes multiple hospitals participating in a critical-care telehealth programme. Their differences provide a demanding setting for cross-database evaluation, but also complicate comparisons. Harmonisation tools such as ricu explicitly map recorded variables to shared clinical concepts [4]. This semantic alignment is necessary, yet a shared variable name does not establish equivalent measurement. An hourly heart-rate value can originate from intermittent charting or a denser monitor feed. Cuff and invasive blood pressures also represent different acquisition methods. Luijken and colleagues used analytical and simulation approaches to show how predictor measurement heterogeneity can alter discrimination, calibration and overall accuracy [5]. This motivates our measurement audits; it does not establish that measurement differences caused the performance changes observed here.

Clinical time-series benchmarks provide reproducible prediction tasks [6], comparative ICU studies assess alternative learning approaches [7], and large electronic-health-record studies demonstrate the predictive capacity of flexible models [8]. Work on circulatory failure [9] and clinical deterioration [10, 11] also illustrates why the prediction time and outcome must match the intended clinical question. External performance is a separate empirical issue: an ICU review evaluates the external-validation evidence [12], Wong and colleagues found poor discrimination and calibration for a widely implemented sepsis model in an external cohort [13], and Valan and colleagues evaluated Sepsis Watch across additional hospitals [14]. Tranchellini and colleagues compared generalisation and adaptation across harmonised ICU sepsis cohorts [15]. These studies establish relevant precedents, but their disease definitions, predictor sets and deployment settings differ from the index-ICU mortality question examined here. Their numerical performance is therefore not a direct benchmark for this study.

Discrimination and probability accuracy address distinct parts of that process. The area under the receiver operating characteristic curve (AUROC) measures how often an event case receives a higher score than a non-event case. It assigns half credit to ties and depends on the predictor distributions within both outcome groups. Consequently, differences in case mix can change AUROC even when model validity changes little [16]. External evaluation must identify the population in which performance is estimated and distinguish patient hold-outs from evaluations in new hospitals [17, 18]. Calibration asks whether predicted risks agree with observed outcomes; overall predictive accuracy and decision analysis add further information [19, 20]. A higher external AUROC is therefore not inherently contradictory, and a lower one does not uniquely identify the cause of reduced performance.

Uncertainty reporting introduces a further distinction between statistical coverage and useful information. Conformal prediction constructs sets of possible labels with coverage properties under specified sampling assumptions [21]. Such sets can communicate limitations of a prediction, but their clinical interpretation requires care [22]. Marginal coverage averages over the outcome distribution. When deaths are uncommon, high aggregate coverage can conceal poor coverage of patients who die. Class-conditional construction instead targets coverage within each outcome class, potentially producing larger, less informative sets [21]. Ovadia and colleagues examined uncertainty under dataset shift [23], while the weighted conformal framework of Tibshirani and colleagues addresses a specified covariate-shift setting [24]. Neither result supplies an automatic guarantee for the unweighted, cross-database procedure evaluated here. Reporting only the aggregate percentage leaves the most consequential failure mode unresolved.

The purpose of this study was to determine what reliability claims can be supported when ICU mortality models are selected in one database and evaluated in another under a common input and training budget. The primary comparison was reciprocal discrimination of four model families selected solely from source development data. Complementary analyses asked whether their probabilities remained accurate and whether prediction sets covered deaths without routinely returning both labels. Measurement and population diagnostics examined possible explanations for performance differences. The intended readers are clinical-informatics researchers and teams evaluating models before prospective use; the study does not establish an action threshold or bedside treatment policy.

The contribution is the joint evaluation of model selection, measurement differences, probability calibration and outcome-specific coverage in the same patients. This connects a model comparison to the information a transported prediction can support, while allowing the different reliability assessments to disagree. Reporting is informed by TRIPOD + AI [25], and limitations are considered using the domains addressed by PROBAST + AI [26]. These update the earlier TRIPOD and PROBAST frameworks [27, 28]; citing them is not a claim of independent certification or a low risk-of-bias rating.

Methods

Study design and endpoint rationale

This retrospective redevelopment used local MIMIC-IV and eICU extracts identified in project provenance as versions 3.1 and 2.0. Input hashes and executable extraction code accompany the report. Both databases had been examined in preceding development work; this is not a preregistered study or a previously untouched external validation. Only results from the revised mortality study are included in the present article. Reporting follows guidance for development and external evaluation of clinical prediction models [29, 17, 18].

The analysis distinguishes three evaluation settings. Source testing estimates performance within the source database under its specified hold-out design. Source-only external testing evaluates the frozen system in the other database. Target updating evaluates a separate system that additionally uses

labelled target development data. Comparisons across these settings describe different information resources and populations; they are not interchangeable estimates of transportability.

An audit of label ascertainment motivated two changes before final model fitting. First, a vasopressor-or-death composite was retired from primary inference because absence of a medication-interface record cannot establish absence of treatment. Second, short-horizon mortality was removed as the primary endpoint because MIMIC-IV records death time while eICU identifies ICU death at the discharge offset. The timestamp amendment was recorded after initial candidate fitting began, but before review of those candidates' test performance; interrupted models were isolated and were not reused. The primary outcome was changed to death before index ICU discharge, which avoids dependence on the exact interval between death and recorded exit. Native-time and exit-offset short-horizon labels remain secondary diagnostics. Changing the endpoint changes the research question; it is not evidence that a previous prediction system improved.

Participants, outcome and observation window

Eligible stays were from adults aged at least 18 years with recorded ICU length of stay greater than 24 hours and no documented ICU death by that landmark. Age used MIMIC-IV anchor-age and admission-year information and was top-coded at 89 in both databases. eICU stays required a usable age, patient identifier and known alive/expired ICU discharge status. MIMIC-IV death was identified from hospital death time falling within the ICU stay; eICU death used expired ICU discharge status. The primary binary outcome was subsequent recorded death before departure from the index ICU stay. Live discharge was a non-event. Follow-up therefore continued to ICU exit rather than ending at a common future horizon. Events after transfer or discharge were outside this endpoint. The estimand is thus the probability of recorded death before index ICU exit, conditional on landmark eligibility and the available predictors. It is not mortality within a fixed future duration or the risk that would occur under a specified treatment strategy.

Exactly one eligible stay per patient was selected by the lowest deterministic SHA-256 stay rank, independent of outcomes. This is a sampled eligible stay, not necessarily a first chronological admission. MIMIC-IV subject_id and eICU uniquepid were the within-database patient keys. Different deidentified databases cannot be linked reliably at the individual level, so absence of cross-database patient overlap cannot be verified. The 24-hour landmark restricts eligibility to patients still recorded in ICU without a documented death at that time. The model is not intended for admission-time triage. eICU cannot resolve deaths occurring before the recorded exit offset, leaving a residual uncertainty in landmark eligibility that is stated rather than treated as solved.

Inputs comprised the six hourly bins ending at hours 18 through 23 after ICU admission. Thus the interval (23,24] was withheld before prediction. Six common concepts were used: heart rate, respiratory rate, peripheral oxygen saturation measured by pulse oximetry (SpO2), and noninvasive systolic, diastolic and mean blood pressure. MIMIC-IV item identifiers were 220045, 220210, 220277, 220179, 220180 and 220181. eICU used heartrate, respiration and sao2 from vitalPeriodic and noninvasivesystolic,

noninvasivediastolic and noninvasivemean from vitalAperiodic. Invasive blood pressure was not substituted for absent cuff measurements.

Plausibility ranges were 20–250 beats/min, 4–80 breaths/min, 40–100%, 40–260 mmHg, 20–180 mmHg and 20–200 mmHg, respectively. These are data-cleaning rules, not clinical decision thresholds. For MIMIC-IV, availability was the later of chart and storage time; records entered more than six hours after measurement were excluded. eICU offered observation offsets without an equivalent storage timestamp. The last available valid value in each bin was retained, with deterministic handling of equal-time conflicts. Neither forward nor backward filling was performed. An outcome-independent sample of 2,000 patients per database was also audited directly in the raw measurement files for last-reading recency within each bin and, in MIMIC-IV, recording lag. This procedure aligns clinical concepts and input windows while retaining unresolved differences in observation frequency and documentation [5, 30]. Missingness was handled with a fixed median-and-mask representation, not multiple imputation. The distinction matters because the missing-data literature treats recovery of missing information and its uncertainty as a separate inferential problem [31]. The withheld final hour provides a temporal gap before prediction, but cannot establish real-time availability where storage timestamps are absent.

Partitions and reciprocal comparisons

Nominal proportions were 50% training, 10% tuning, 10% probability calibration, 10% conformal calibration and 20% testing. MIMIC-IV roles were assigned by patient hash. eICU roles were assigned by hospital hash, and patients whose eligible stays crossed hospital roles were excluded before index-stay selection. All eICU test hospitals were consequently absent from eICU training, tuning and calibration. Realized counts, including exclusions, are reported. The MIMIC-IV source test is a patient hold-out from one centre; the eICU source test is a hospital hold-out within that database. These are deliberately distinguished throughout the article.

Two independent development experiments trained on MIMIC-IV and eICU, respectively. Each used the smaller of the available source-training patient counts, with patients chosen by an outcome-independent hash. Every model in both directions received exactly 27,437 training patients. Event counts and source calibration sample sizes were not artificially equalised and are disclosed. The fixed MIMIC-IV and eICU test populations were identical across model families and directions. Target test outcomes were never used to choose a model, epoch, seed or probability-calibration family. Target development partitions were used only in explicitly labelled diagnostic weighting or local updating analyses.

Preprocessing and model development

Missing values were replaced with feature medians estimated on the actual source training sample. Imputed values were standardised using that sample and clipped to [− 10,10]. Each model received 36 values, 36 missing indicators, age, male and unknown-sex indicators, and a constant landmark-time field. The constant field carries no discriminatory information but preserves a common 76-input implementation. Tabular models received the flattened history; the Transformer received the same

information as six sequence tokens plus the static fields. Missingness can encode the recording process, so mask removal is not interpreted as eliminating every trace of that process [30].

The four families were L2-regularised logistic regression, histogram gradient boosting, extra trees and a compact Transformer [32, 33, 34, 35]. Sixteen configurations per family were specified before the final search. The grid varied regularisation and class weighting for logistic regression; learning rate, leaf size, iteration count and regularisation for boosting; leaf size, feature fraction, tree count and depth for extra trees; and embedding width, layer count, learning rate, dropout, weight decay and duration for the Transformer. The complete grid and runtimes accompany the report. The Transformer used two attention heads, feed-forward width twice its embedding dimension, AdamW, batches of 1,024 and gradient clipping at one. Every duration was fixed by its candidate configuration, and its final checkpoint was evaluated.

Model selection had two stages. Within each matched training sample, an outcome-independent hash threshold of 0.20 assigned MIMIC-IV patients and eICU hospitals to an internal screening partition. The remaining inner training patients supplied both preprocessing estimates and model fitting. Thus, the nominal screening fraction applies to hospitals in eICU, not to its patient count. Seed-42 screening AUROC ranked the 16 settings, with average precision and then configuration identifier breaking ties. The two leading settings per family were then refitted on all 27,437 source training patients with seeds 42, 117 and 2026; preprocessing was re-estimated on that full training sample. Mean AUROC on the separate source-tuning partition selected each family's configuration and then the source model family, with mean average precision breaking ties. The source-tuning partition was therefore a model-selection resource rather than an unbiased performance test.

Seed 42 was fixed as the primary fit; the other seeds assessed fitting variability without ensembling or test-based seed selection. Training AUROC was reported to assess fit and the training–validation gap, but did not determine selection. Both sources' configuration and family choices were locked before calculating the final systems' test predictions. The search was a retrospective design revision after earlier analyses of these databases, so the reported hold-outs are reused benchmarks rather than newly untouched validation data. Its protocol, timing and preceding analyses are retained in the audit trail. No further tuning followed the final test evaluation.

Equal candidate counts and training rows control specific sources of unfairness; they do not equalise computation, event information or each family's distance from its best possible configuration. The two-stage design limits full-budget comparisons but can screen out a setting that would perform better with more training data. Finite validation samples also remain susceptible to selection overfitting; the implications are addressed in the Discussion. These limitations constrain claims of optimality. The benchmark concerns the evaluated implementations, rather than all sequence models or replacements for interpretable clinical assessment [36].

Probability calibration and statistical evaluation

Post hoc probability calibration has been studied for conventional classifiers [37] and neural networks [38]. Here, each system used an affine logistic transformation of its raw probability logit, fitted on a separate source probability-calibration partition. We evaluated its resulting probabilities using calibration concepts described by Van Calster and colleagues [19]. The calibration family was fixed. Evaluation included AUROC, average precision, Brier score, log loss, mean predicted risk and observed-to-expected risk. Average precision was accompanied by prevalence; an additional diagnostic standardised precision–recall calculations to a 10% reference prevalence while retaining the empirical true- and false-positive rates [39, 40]. This removes a mathematical prevalence difference, not case-mix or label differences.

Evaluation calibration slope and free-slope intercept were estimated jointly. Calibration-in-the-large was estimated separately with the logit slope fixed at one. Those intercepts answer different questions and are not interchanged. Calibration curves used ten prediction-quantile groups, and interval estimates for regression coefficients used patient-cluster sandwich covariance. Probability updating on the target calibration partition was reported separately on the unchanged target test population; eICU updating hospitals were disjoint from eICU test hospitals. It changes the data resources available to the system and is not source-only transportability.

Main metric intervals used 1,000 patient bootstrap samples, shared across model families within each evaluation population. Because the primary sample has one stay per patient, each patient contributes one prediction. Paired differences used percentile intervals and exploratory centred-bootstrap p values, with Holm adjustment over the six model pairs separately for each source, population and metric. eICU hospital bootstrap intervals assessed between-hospital sampling sensitivity. Differences between external and source-test AUROC used independent resampling of the two populations; they are not paired comparisons of the same patients. No significance test is interpreted as proving transportability. Intervals condition on fitted systems and fixed preprocessing, rather than incorporating all model-development uncertainty [41]. All eligible data were considered before outcome-independent sampling for the matched training budget. No prospective sample-size calculation or clinical power claim was made [42, 43].

Measurement and population diagnostics

A fixed domain classifier used equally sampled, unlabelled development inputs to distinguish databases. Its predictors were six within-window means, six missing fractions, age and recorded male sex. The classifier was histogram gradient boosting with 150 iterations, 15 leaves and L2 penalty 5. Its AUROC was assessed in the held-out populations. For a domain probability e = P(eICU|X), bounded to [0.05,0.95], diagnostic weights were e in MIMIC-IV and 1 − e in eICU. Li and colleagues developed overlap weights to target covariate overlap in propensity-score analyses [44]. Here, analogous weights based on database membership define a descriptive common-input population; the database label is not treated as a causal intervention. Effective sample size, extreme domain probabilities and standardised mean differences are

reported. Estimated and clipped weights do not guarantee exact balance, equal outcomes or causal adjustment. Bootstrap intervals condition on the fitted weighting model.

A symmetric thinning analysis used the lower development-sample observation frequency in the two databases, separately for each feature and hour. Observed test inputs were randomly removed to approach that frequency, using three fixed seeds. Source medians, fitted models and calibrators remained unchanged. Achieved availability was reported. Thinning is an artificial stress test that approximately aligns marginal availability; it does not align devices, charting error, dependence between measurements or clinical decisions about when to observe a patient [5, 45]. A further evaluation required at least three observed bins for each of heart rate, respiratory rate and SpO2, with its changed sample size explicitly stated.

For secondary ablations, boosting was retuned under a separate eight-candidate budget after removing blood pressure or explicit masks. It was also refitted for two 24-hour future labels: native recorded death time and death-associated ICU exit by hour 48. These change the task and are not improvements over the primary outcome. An age/sex-only model provided a restricted comparator, and boosting models trained on 20 fixed permutations of source labels provided negative controls. Age and sex subgroup estimates were descriptive, with no assertion of equitable clinical benefit or equivalence [46, 47].

Prediction sets and the rare-event failure baseline

For predicted death risk p and observed label y, the nonconformity score was 1 − p when y = 1 and p when y = 0. The finite-sample threshold was the ceil((n + 1)×0.9)-th ordered score; if that index exceeded the sample size, the conservative maximum score of one was used [21]. The conventional marginal construction used one threshold for both labels. It was designated a failure baseline, not the primary evidence of reliable uncertainty. If its threshold q is below 0.5, it includes the non-event label for $p \le q$, the event label for $p \ge 1 - q$, and neither label for $q < p < 1 - q$. Empty sets are therefore possible and must not be recoded as negative predictions.

The principal uncertainty analysis used separate class-conditional calibration thresholds. It reported overall, event and non-event coverage, empty sets, two-label sets and set size. Event denominators and Wilson intervals accompany coverage proportions, with separate hospital-bootstrap intervals for source-calibrated eICU coverage; these intervals do not remove within-hospital dependence or distribution shift. A trivial always-non-event set was also tabulated, showing why high overall coverage alone is an inadequate clinical criterion. Target-updated class-conditional results used target probability calibration and a distinct target conformal partition. Source, target and hospital exchangeability is not assumed to follow from calibration, and no distribution-free external or clinical safety guarantee is claimed [24, 48, 22].

Reproducibility and tool assistance

Checks included patient and hospital separation, exact event boundaries, independent reconstruction of sampled feature tensors, comparison with raw eICU discharge status, numerical agreement with established AUROC/AP implementations, and independent enumeration of conformal sets. Synthetic tests deliberately reproduce high marginal coverage with zero event coverage. Passing them verifies computation, not clinical label validity. Protocols, timing amendments, complete candidate and seed results, aggregate evidence, source hashes and code accompany this report. OpenAI Codex assisted with code debugging and manuscript grammar checking. Analyses were executed locally; responsibility for scientific interpretation and the submitted work rests with the human author.

Results

Tables 1–4 and Figs. 1–6 report the final models selected by source-validation AUROC and evaluated under the same locked protocol.

Cohort composition and residual measurement differences

The cohorts contained 54,376 MIMIC-IV and 102,581 eICU patients, each contributing one stay and prediction. The source training budget was 27,437 patients per model and direction. All 42 eICU test hospitals were absent from the other eICU roles. Table 1 reports realized partitions; the accompanying flow table includes exclusions before the index-stay sample. The training-partition eICU count exceeds the matched budget because selection within that partition was outcome-independent. The matched training samples contained 2,172 MIMIC-IV and 1,356 eICU deaths; equal patient budgets therefore did not equalise event information.

Table 1
Cohorts and validation roles.

Database / role	Patients	ICU deaths	Mortality	Hospitals
MIMIC-IV / probability calibration	5332	421	7.9%	1
MIMIC-IV / conformal calibration	5451	465	8.5%	1
MIMIC-IV / testing	10721	813	7.6%	1
MIMIC-IV / training	27437	2172	7.9%	1
MIMIC-IV / tuning	5435	441	8.1%	1
eICU / probability calibration	10740	640	6.0%	20
eICU / conformal calibration	13606	883	6.5%	24
eICU / testing	18766	1245	6.6%	42
eICU / training	46922	2327	5.0%	105
eICU / tuning	12547	767	6.1%	17

Counts refer to unique patients with one index stay, before the matched training subsample.

Among 6,519 MIMIC-IV stays with recorded ICU death in the extraction audit, the median interval from recorded death to ICU exit was 2.77 hours, and its 95th percentile was 10.23 hours. There were 169 deaths recorded during hours (24,48] whose ICU exit occurred after hour 48. This supplied a concrete reason to remove exact future death timing from the primary endpoint. It did not establish that eICU had the same delay distribution.

Figure 1 shows input missingness and the original observation density. A raw-record audit of 2,000 outcome-independently sampled patients per source found median last-reading recency of 26.8–27.0 minutes before bin end across MIMIC-IV features, versus 2 minutes for eICU core vital signs and 15–16 minutes for cuff pressures. MIMIC-IV mean chart-to-availability lag was 27.6–30.4 minutes; equivalent eICU storage lag was unavailable. These timing differences remain a limitation of the common hourly representation. A domain classifier fitted to development inputs achieved held-out AUROC 0.957, demonstrating residual source separability after common concept and window definitions. Under diagnostic overlap weighting, effective sample sizes were 3,121 MIMIC-IV and 7,825 eICU test patients. Balance diagnostics accompany the weighting analysis and are supplied numerically.

Reciprocal discrimination and probability reliability

The search completed 128 screening fits and 48 full-budget confirmation fits. MIMIC-IV screening used 22,033 inner training patients and 5,404 screening patients; eICU used 18,314 and 9,123, respectively. Mean source-validation AUROC selected a Transformer in both development directions. The selected MIMIC-IV configuration used embedding width 32, one layer, learning rate 0.003, 30 epochs, dropout 0.1 and weight decay 0.001. The selected eICU configuration used width 32, one layer, learning rate 0.001, 45 epochs, dropout 0.2 and weight decay 0.01. Mean validation AUROCs were 0.759 and 0.713, with standard deviations across fitting seeds of 0.0019 and 0.0055. These seeds share a validation population and are not independent validation cohorts.

Tables 2 and 3 report all locked family selections; test results use primary seed 42. Table 2 also reports mean training and validation AUROC across three seeds, separating fit from held-out performance. Figure 2 presents test discrimination and average precision. Complete candidate results, selected hyperparameters and all fitting seeds accompany the report; neither target ranking nor the best test seed replaces source selection.

Model ordering differed between evaluation settings. Among MIMIC-IV-trained systems, the Transformer had source-test AUROC 0.776 and AP 0.282, compared with 0.770 and 0.265 for gradient boosting. In eICU, gradient boosting had AUROC 0.743 and AP 0.217, compared with 0.720 and 0.164 for the source-selected Transformer. Among eICU-trained systems, the Transformer had the highest external AUROC point estimate, 0.738. These target comparisons are retrospective descriptions; they did not determine selection.

Table 2
Training, validation and reciprocal test AUROC.

Training database	Model	Training	Validation	MIMIC-IV test (95% CI)	eICU test (95% CI)
MIMIC-IV	Logistic regression	0.723	0.698	0.728 (0.710–0.745)	0.700 (0.685–0.717)
MIMIC-IV	Gradient boosting	0.889	0.748	0.770 (0.754–0.786)	0.743 (0.729–0.758)
MIMIC-IV	Extra trees	0.962	0.743	0.758 (0.742–0.775)	0.733 (0.720–0.749)
MIMIC-IV	Transformer (selected)	0.806	0.759	0.776 (0.760–0.792)	0.720 (0.706–0.735)
eICU	Logistic regression	0.730	0.657	0.693 (0.674–0.713)	0.709 (0.694–0.725)
eICU	Gradient boosting	0.870	0.706	0.729 (0.711–0.745)	0.762 (0.748–0.777)
eICU	Extra trees	0.941	0.706	0.725 (0.707–0.743)	0.768 (0.754–0.782)
eICU	Transformer (selected)	0.806	0.713	0.738 (0.720–0.755)	0.767 (0.753–0.781)

Training and validation values average three fitting seeds for each locked configuration. Test results use seed 42 with 95% patient-bootstrap intervals conditional on the fitted model. Selected identifies the family chosen using source validation. MIMIC-IV testing is a within-centre patient hold-out; eICU testing uses held-out hospitals.

Table 3
Reciprocal average precision.

Training database	Model	MIMIC-IV test AP (95% CI)	eICU test AP (95% CI)
MIMIC-IV	Logistic regression	0.208 (0.185–0.236)	0.163 (0.149–0.182)
MIMIC-IV	Gradient boosting	0.265 (0.235–0.297)	0.217 (0.197–0.242)
MIMIC-IV	Extra trees	0.255 (0.229–0.287)	0.197 (0.180–0.222)
MIMIC-IV	Transformer	0.282 (0.254–0.314)	0.164 (0.150–0.183)
eICU	Logistic regression	0.188 (0.166–0.214)	0.200 (0.180–0.223)
eICU	Gradient boosting	0.219 (0.195–0.246)	0.240 (0.217–0.265)
eICU	Extra trees	0.213 (0.189–0.240)	0.241 (0.219–0.266)
eICU	Transformer	0.226 (0.201–0.256)	0.239 (0.217–0.263)

Test prevalence is 0.0758 in MIMIC-IV and 0.0663 in eICU. Intervals use the same patient resamples as Table 2.

For the MIMIC-IV-selected system, external minus source-test AUROC was − 0.056 (95% CI -0.078 to -0.033). External mean predicted mortality was 10.5% versus observed 6.6%; calibration-in-the-large was − 0.573 and calibration slope 0.693. Separate target updating changed external Brier score from 0.0640 to 0.0591.

For the eICU-selected system, external minus source-test AUROC was − 0.028 (95% CI -0.051 to -0.007). External mean predicted mortality was 4.7% versus observed 7.6%; calibration-in-the-large was 0.530 and calibration slope 1.206. Separate target updating changed external Brier score from 0.0667 to 0.0649. The cross-database AUROC contrasts involve different patient populations. Each before-and-after updating comparison uses the same target test patients.

Figure 3 shows the cross-population AUROC contrasts without presuming their sign. Figure 4 shows calibration curves. Full Brier/log-loss results, free-slope intercepts, offset intercepts, hospital bootstrap intervals and Holm-adjusted paired comparisons are included in the supplement and aggregate tables. The statistical contrast between databases does not isolate a causal effect of transport.

Prediction-set failure and class-specific coverage

Table 4 separates the marginal failure baseline, source class-conditional construction and target-updated construction for each source-selected model. The always-non-event set would cover 92.4% of MIMIC-IV and 93.4% of eICU test patients while covering none of the deaths. This makes aggregate coverage an inadequate stand-alone success criterion in these populations.

Table 4
External prediction-set performance at nominal 90% coverage.

Direction / source-selected system	Set method	Overall coverage	Death coverage (95% CI)	Both labels	Empty
MIMIC-IV → eICU	Marginal baseline	88.9%	1.4% (0.9%–2.2%)	0.0%	5.6%
MIMIC-IV → eICU	Class conditional	83.8%	91.6% (89.9%–93.0%)	54.2%	0.0%
MIMIC-IV → eICU	Target updated class conditional	90.3%	88.9% (87.1%–90.5%)	56.1%	0.0%
eICU → MIMIC-IV	Marginal baseline	91.1%	0.0% (0.0%–0.5%)	0.0%	2.1%
eICU → MIMIC-IV	Class conditional	93.7%	84.9% (82.2%–87.2%)	49.3%	0.0%
eICU → MIMIC-IV	Target updated class conditional	90.8%	91.4% (89.3%–93.1%)	54.8%	0.0%

Death-class Wilson intervals are descriptive patient-level intervals; they do not establish hospital exchangeability. A two-label set contains both survival and death. An empty set contains neither. Separate hospital-bootstrap coverage intervals for eICU are in the aggregate supplement.

Figure 5 displays death coverage, two-label frequency and empty-set frequency together. Class-conditional thresholds address a different target from marginal calibration. More two-label sets indicate less decisive set-valued predictions, not improved discrimination. Local updating uses labelled target development data and is reported as a separate resource setting.

Stress tests, controls and verification

For the MIMIC-IV-selected system evaluated in eICU, overlap-weighted AUROC was 0.723, compared with 0.720 without weighting. Symmetric thinning yielded AUROCs of 0.715–0.719 across three draws. For the eICU-selected system evaluated in MIMIC-IV, overlap-weighted AUROC was 0.728, compared with 0.738 without weighting. All three thinned AUROCs rounded to 0.738. These analyses evaluate altered population weights or artificially reduced input availability; neither estimates a clinical intervention effect.

Figure 6 reports every model under the same thinning procedure. Boosting refits without blood pressure or explicit masks and with alternate short-horizon labels are supplied in Supplementary Figure S1. The endpoint sensitivities refer to different tasks; they are not used to select the primary task or declare a performance improvement. Core-observed subsets have their own denominators and are retained alongside the complete eligible population.

Across 20 fixed label permutations per source, mean control AUROCs ranged from 0.486 to 0.503. Individual shuffled-training-label fits produced test AUROCs from 0.415 to 0.555; age/sex-only comparators are also reported. All 79 original implementation checks and 94 model-selection and evaluation checks passed. These checks include independent conformal-set enumeration and a synthetic rare-event example with zero marginal event coverage. They verify the reported computation without certifying endpoint adjudication or clinical safety. Supplementary Figure S2 provides illustrative decision curves and workload estimates; subgroup and hospital summaries remain descriptive.

Discussion

This study examined whether a common input and development design was sufficient to support reliable mortality prediction across two ICU databases. The findings show why reliability requires several distinct assessments. Source selection identified a Transformer in both directions, but the MIMIC-IV-trained gradient boosting model performed better on external discrimination point estimates. Probability errors also differed by direction, while marginal prediction sets provided little or no coverage of deaths. These findings connect model comparison to the more demanding question of what information a transported system can credibly provide.

What harmonisation makes comparable

The common design reduces ambiguity in attributing differences to the fitted systems. All models used the same vital-sign concepts, historical window and training-patient budget, with preprocessing fitted only on the source training sample. One stay per patient avoided disproportionate representation of patients with repeated admissions. Hospital-disjoint eICU partitions prevented those hospitals from contributing simultaneously to model development and testing. The resulting comparisons are more interpretable than comparisons in which architectures receive different predictors or unequal access to target outcomes.

These controls do not make the underlying information equivalent. Raw-record recency differed despite common hourly bins, and the domain classifier distinguished the databases with AUROC 0.957. Such separability establishes differences in the represented inputs, not their cause or clinical importance. It may reflect physiology, population composition, documentation or combinations of these factors. The measurement-error literature provides a plausible account of how acquisition differences can affect predictions [5]. The present diagnostics do not identify how much of the external performance difference arose through that mechanism.

Missingness deserves particular attention because the recording process can itself carry predictive information. Work on clinical time series has used missingness patterns explicitly as model inputs [30]. Here, missing indicators could encode clinical monitoring decisions as well as technical availability. Removing explicit masks would not necessarily remove this information, because imputed values and the remaining trajectories can retain related patterns. Conversely, retaining masks does not establish that the model learned an invalid association. The unresolved question is whether those associations remain useful under the target recording process, consistent with broader accounts of dataset shift [45].

The stress tests narrow this interpretation only modestly. Overlap weighting changed the selected systems' external AUROCs little, but reduced the effective sample sizes to 3,121 and 7,825. This analysis concerns a reweighted region of measured covariate overlap rather than the complete target population. It cannot adjust for unmeasured treatment practices or outcome ascertainment, and it does not establish that case mix is unimportant. Thinning similarly assessed sensitivity to artificial loss of measurements. Stability under those perturbations does not establish robustness to altered values, devices, recording delays or clinically driven observation. These diagnostics test particular explanations; they do not eliminate all alternatives.

Source selection and the meaning of reciprocal performance

The model-ranking result is consequential because source selection approximates the choice available before target outcomes are examined. In MIMIC-IV development, the Transformer led the reported source-test AUROC and AP estimates, yet gradient boosting led both externally. This reversal shows that the preferred source system and the retrospectively strongest target system can differ under the same input budget. It does not justify selecting gradient boosting using the reported eICU test outcomes. Such a choice would require a new evaluation population to assess the resulting selection procedure.

The reverse direction also prevents a simple claim that one model family inherently transports better. The eICU-trained Transformer had the highest external AUROC and AP point estimates among the evaluated systems. Equal patient counts did not equalise event information: training samples contained 2,172 MIMIC-IV deaths and 1,356 eICU deaths. They also differed in hospital diversity, and equal configuration counts did not equalise computational effort or optimisation quality. The evidence therefore concerns these implementations and this restricted information budget. It neither establishes universal architectural superiority nor excludes benefits from richer longitudinal representations [6, 7, 8, 36].

Reciprocal transport is also not a symmetric experiment. The MIMIC-IV source test is a patient hold-out within one centre, whereas the eICU source test already evaluates new hospitals. Their respective external AUROC reductions of 0.056 and 0.028 cannot be interpreted as directly comparable measures of institutional robustness. Differences in population heterogeneity, outcome recording and predictor–outcome relationships remain possible explanations [16, 17, 18]. The contribution extends previous external-validation and ICU adaptation work by reporting these distinctions alongside measurement diagnostics and outcome-specific coverage [12, 13, 14, 15].

Probability calibration identifies errors that ranking cannot resolve

The source-selected systems made opposite errors in average external risk. MIMIC-IV development produced a mean prediction of 10.5% in eICU, where observed mortality was 6.6%. eICU development produced a mean prediction of 4.7% in MIMIC-IV, where observed mortality was 7.6%. These discrepancies matter because a decision threshold acts on the numerical probability, not solely on rank [19, 20]. Without a specified clinical action, however, the study cannot translate them into observed overtreatment or missed treatment.

The calibration slopes suggest that the errors were not identical in form. The MIMIC-IV-trained system's external slope of 0.693 indicates risk estimates that were too dispersed on the logit scale. The reverse system's slope of 1.206 indicated a compressed risk gradient, together with underestimation of average mortality. This pattern is compatible with different updating needs, but point estimates alone do not establish equivalence or a unique cause. In particular, an external slope below one is not proof of development overfitting; measurement and population changes can also contribute [5, 19].

Separate target updating improved Brier scores, with a larger numerical change in the MIMIC-IV-to-eICU direction. This supports the practical relevance of evaluating recalibration, while making its additional data requirement explicit. Target labels were used in separate calibration partitions, and eICU updating hospitals remained distinct from test hospitals. The result therefore evaluates adaptation using target-database information rather than calibration fitted within each deployment hospital. Better average probability accuracy does not by itself establish adequate calibration at clinically relevant thresholds or within every hospital or subgroup.

Coverage can conceal failure for the outcome of interest

The marginal prediction-set results expose a failure that AUROC and aggregate coverage cannot detect. For eICU-to-MIMIC-IV transport, overall coverage was 91.1%, yet death coverage was 0.0%. This is possible because overall coverage is the prevalence-weighted average of event and non-event coverage. At the observed mortality proportions, always returning the non-event label would cover more than 90% of patients while missing every death. The problem is therefore substantive even when a marginal construction is implemented correctly.

Three assessments must be kept distinct. Bootstrap confidence intervals describe sampling uncertainty in estimated performance, conditional on the fitted systems. Probability calibration assesses agreement between risks and outcomes. Conformal prediction sets concern inclusion of the realised label under specified sampling assumptions [21, 41]. A prediction set is not a confidence interval around an individual patient's probability of death. Nor does its nominal coverage quantify the chance that a particular clinical decision is safe [22]. Median imputation remains a fixed preprocessing operation; the prediction sets do not reconstruct missing physiological values. The analysis does not separately quantify uncertainty from measurement error, missing data and model estimation.

Class-conditional construction addressed the event-coverage failure, but it did not uniformly retain nominal external coverage. Death coverage reached 91.6% in eICU and 84.9% in MIMIC-IV, with both labels returned for 54.2% and 49.3% of patients. The MIMIC-IV-to-eICU overall coverage of 83.8% further shows that adequate event coverage does not ensure adequate coverage across the full population. Outcomes must be examined separately, alongside empty sets and two-label sets. The findings concern the evaluated score and calibration procedure, rather than all possible conformal methods.

Target updating brought empirical death coverage near 90% in both directions while returning both labels for more than half of patients. Such sets may indicate cases requiring additional information, but that interpretation remains a proposed use rather than a validated referral policy. An empty set likewise cannot be treated as a low-risk prediction. Class-conditional coverage is conditional on the outcome class; it is not coverage conditional on each hospital, demographic group or individual risk profile. Distribution-shift methods provide relevant methodological directions [24, 48], but no source or target construction here establishes a distribution-free guarantee for future hospitals.

What rigorous model selection can establish

The separation of training, screening, confirmation and testing clarifies what the selected configurations establish. Extra trees reached mean training AUROCs of 0.962 and 0.941, whereas validation AUROCs were 0.743 and 0.706. The selected Transformers had lower training AUROC, approximately 0.806 in both sources, but higher source-validation AUROC. Maximising apparent training performance would therefore favour a different model without demonstrating better generalisation. Training–validation gaps reflect both fitting and differences between partitions; the hospital separation in eICU makes a purely overfitting-based interpretation especially incomplete.

Source validation answers which candidate performs best under the development selection rule. It cannot identify the highest external AUROC without access to target outcomes, and the final model-ranking reversal illustrates this distinction. The selected configuration is the best among the confirmed candidates under the disclosed search, not a demonstrated global optimum. Cawley and Talbot showed that optimising a finite-sample selection criterion can itself produce overfitting and biased performance assessment [49]. In this study, screening can discard a setting that would perform better after full-budget training, and validation rankings remain sample-dependent. The selection lock prevents the final test scores from choosing among these settings; it cannot remove the influence of earlier benchmark experience. Three fitting seeds help reveal optimisation variability but cannot quantify uncertainty from selecting a new validation sample.

The audit trail also permits comparison with preceding development choices. Paired external AUROC differences for the final source-selected models were − 0.0002 (95% CI − 0.0057 to 0.0059) after MIMIC-IV development and − 0.0065 (− 0.0146 to 0.0022) after eICU development. These exploratory pointwise intervals do not establish either improvement or equivalence. Other families, including logistic regression in both directions, obtained higher external AUROC under the final settings. Thus, tuning can improve an individual family without making source selection identify the strongest external model. The comparisons combine changes in search space, selection metric and confirmation procedure; they cannot isolate the effect of a single hyperparameter. They also reuse previously examined benchmarks. Independent data are required to assess the complete development procedure without this history.

Clinical scope and the evidence still required

The endpoint amendment improves comparability at a clear cost to the clinical question. Death before index ICU exit avoids requiring equivalent exact death timestamps, but follow-up duration varies with the remaining stay. Live discharge terminates the episode as a non-event, and subsequent deaths are unobserved by this endpoint. Discharge, transfer and treatment-limitation practices can consequently influence the recorded risk. The model estimates an observed episode outcome under existing care, not preventable mortality or benefit from an intervention. A fixed-horizon deterioration question would require separately justified labels and evaluation.

The 24-hour landmark further limits applicability to patients who remain in the recorded ICU episode. Early deaths and shorter stays are outside the target population, so these results do not support admission triage. The eICU exit-offset convention leaves uncertainty about deaths occurring before recorded departure. The six-vital-sign representation also omits laboratory data, diagnoses and treatment trajectories. Its restricted scope supports controlled comparisons, but limits claims about the best achievable mortality prediction or added value over established clinical assessment.

Retrospective redevelopment imposes another boundary. Both resources had been examined previously, and the endpoint amendment followed the start of candidate fitting, as disclosed in Methods. Hold-outs prevent direct fitting to their outcomes but do not remove the influence of earlier database experience on study design. The analysis is consequently not an untouched external validation. Fixed-model intervals

also omit uncertainty from the broader development process, and descriptive subgroup analyses cannot establish equitable clinical benefit [41, 43, 46, 47]. Independent contemporary cohorts remain necessary to assess whether the reported findings persist.

The implementation checks and shuffled-label controls strengthen confidence in the computation. They do not substitute for endpoint adjudication or prove that every leakage pathway has been excluded. Decision-curve analysis combines true- and false-positive classifications using a specified threshold trade-off [50]. Our illustrative curves apply that framework; they do not measure a treatment effect. A clinically informative next study would specify the intended action and time horizon before evaluating the locked system. It would assess label validity, real-time predictor availability, calibration and death coverage across sites, alongside the workload created by uncertain outputs. The implementation literature identifies workflow and evaluation barriers [51, 52], model facts labels provide a way to communicate intended use and limitations [53], and DECIDE-AI addresses reporting of early clinical evaluation [54]. These sources motivate the next stage of work; none establishes clinical benefit for the present models. Any subsequent intervention trial would require its own protocol and outcome evaluation [55, 56, 57].

Conclusions

Reciprocal evaluation under common inputs and training budgets made model comparisons more interpretable, while exposing limits of source selection and uncertainty reporting. A source-selected model did not necessarily lead externally, and near-nominal aggregate coverage could coexist with complete failure to cover deaths. Class-conditional construction and target updating improved event coverage at the cost of frequent two-label sets. These findings support evaluating measurement comparability, discrimination, probability calibration and prediction-set informativeness together. They provide evidence about research systems under the disclosed design, with clinical usefulness still requiring evaluation in a defined prospective workflow.

Abbreviations

AP
average precision
AUROC
area under the receiver operating characteristic curve
BP
blood pressure
CI
confidence interval
eICU
eICU Collaborative Research Database
ICU

intensive care unit
MIMIC-IV
Medical Information Mart for Intensive Care IV
PROBAST
Prediction model Risk Of Bias ASsessment Tool
SpO2
peripheral oxygen saturation
TRIPOD
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Declarations

Ethics approval and consent to participate

This secondary analysis used deidentified research data without new recruitment, intervention or patient contact. MIMIC-IV was established with Beth Israel Deaconess Medical Center institutional review board approval and waiver of informed consent; the eICU resource was released under the deidentification and governance described in its data descriptor [1, 2]. No additional study-specific ethics approval was obtained. Access and reuse remain subject to the applicable PhysioNet credentials and data-use agreements. No new approval number or institutional exemption determination is claimed.

Consent for publication

Not applicable. No identifiable individual information is reported.

Competing interests

The author declares no competing interests.

Funding

This research received no specific funding.

Author Contribution

Tianyi Yu is the sole author and is responsible for study conception, methodology, analysis, interpretation and the manuscript. The author retains responsibility for the final submitted work.

Acknowledgements

The author thanks the patients, clinical teams and database contributors who made the research resources available.

Data Availability

MIMIC-IV and eICU are available to appropriately credentialed researchers through PhysioNet [1, 2, 3]. Patient-level data cannot be redistributed with this article. Additional files 1–3 contain supplementary methods and results, extraction, modelling and evaluation code, aggregate tables, source fingerprints and software requirements. Figures are supplied separately. These files exclude patient records, patient predictions and trained patient-derived objects. Authorized investigators can reproduce the analysis using the specified source releases.

References

1. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. 10.1038/s41597-022-01899-x.
2. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. 10.1038/sdata.2018.178.
3. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation. 2000;101(23):e215–20. 10.1161/01.CIR.101.23.e215.
4. Bennett N, Plečko D, Ukor IF, Meinshausen N, Bühlmann P. ricu: R's interface to intensive care data. GigaScience. 2023;12:giad041. 10.1093/gigascience/giad041.
5. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. Stat Med. 2019;38:3444–59. 10.1002/sim.8183.
6. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data. 2019;6:96. 10.1038/s41597-019-0103-9.
7. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform. 2018;83:112–34. 10.1016/j.jbi.2018.04.007.
8. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digit Med. 2018;1:18. 10.1038/s41746-018-0029-1.
9. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. 2020;26:364–73. 10.1038/s41591-020-0789-4.
10. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the

Wards. Crit Care Med. 2016;44:368–74. 10.1097/CCM.0000000000001571.

11. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun. 2020;11:3852. 10.1038/s41467-020-17431-x.
12. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. BMC Med Inf Decis Mak. 2025;25:5. 10.1186/s12911-024-02830-7.
13. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021;181:1065. 10.1001/jamainternmed.2021.2626.
14. Valan B, Prakash A, Ratliff W, Gao M, Muthya S, Thomas A, et al. Evaluating sepsis watch generalizability through multisite external validation of a sepsis machine learning model. npj Digit Med. 2025;8:350. 10.1038/s41746-025-01664-5.
15. Tranchellini F, Farag Y, Jutzeler C, Meegahapola L. Evaluating deep learning sepsis prediction models in ICUs under distribution shift: a multi-centre retrospective cohort study. npj Digit Med. 2026;9:306. 10.1038/s41746-026-02364-4.
16. Nieboer D, van der Ploeg T, Steyerberg EW. Assessing Discriminative Performance at External Validation of Clinical Prediction Models. PLoS ONE. 2016;11:e0148820. 10.1371/journal.pone.0148820.
17. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. 10.1136/bmj-2023-074819.
18. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. 10.1136/bmj-2023-074820.
19. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. On behalf of Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. 10.1186/s12916-019-1466-7.
20. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models. Epidemiology. 2010;21:128–38. 10.1097/EDE.0b013e3181c30fb2.
21. Angelopoulos AN, Bates S. Conformal Prediction: A Gentle Introduction. Foundations and Trends® in Machine Learning. 2023;16:494–591. 10.1561/2200000101
22. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. npj Digit Med. 2021;4:4. 10.1038/s41746-020-00367-3.
23. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S et al. Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems. 2019;32.

https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html. Accessed 19 September 2026.

24. Tibshirani RJ, Foygel Barber R, Candès EJ, Ramdas A. Conformal prediction under covariate shift. Advances in Neural Information Processing Systems. 2019;32. https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html. Accessed 19 September 2026.
25. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
26. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. 10.1136/bmj-2024-082505.
27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162:55–63. 10.7326/M14-0697.
28. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170:51–8. 10.7326/M18-1376.
29. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ. 2024;386:e078276. 10.1136/bmj-2023-078276.
30. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci Rep. 2018;8:6085. 10.1038/s41598-018-24271-9.
31. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. 10.1136/bmj.b2393.
32. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1). 10.18637/jss.v033.i01.
33. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals Stat. 2001;29(5). 10.1214/aos/1013203451.
34. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63:3–42. 10.1007/s10994-006-6226-1.
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed 19 September 2026.
36. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15. 10.1038/s42256-019-0048-x.

37. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning - ICML '05. 2005:625–632. 10.1145/1102351.1102430
38. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of Machine Learning Research. 2017;70:1321–1330. https://proceedings.mlr.press/v70/guo17a.html. Accessed 19 September 2026.
39. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML '06. 2006:233–240. 10.1145/1143844.1143874
40. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE. 2015;10:e0118432. 10.1371/journal.pone.0118432.
41. Riley RD, Collins GS, Kirton L, Snell KI, Ensor J, Whittle R, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. BMJ. 2025;e080749. 10.1136/bmj-2024-080749.
42. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. 10.1136/bmj.m441.
43. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021;375:n2281. 10.1136/bmj.n2281.
44. Li F, Morgan KL, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting. J Am Stat Assoc. 2018;113:390–400. 10.1080/01621459.2016.1260466.
45. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345–52. 10.1093/biostatistics/kxz041.
46. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–53. 10.1126/science.aax2342.
47. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018;169:866–72. 10.7326/M18-1990.
48. Gibbs I, Candès E. Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems. 2021;34. https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html. Accessed 19 September 2026.
49. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research. 2010;11:2079–2107. https://www.jmlr.org/papers/v11/cawley10a.html. Accessed 19 September 2026.
50. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Making. 2006;26:565–74. 10.1177/0272989X06295361.

51. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. 10.1186/s12916-019-1426-2.
52. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25:1337–40. 10.1038/s41591-019-0548-6.
53. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. npj Digit Med. 2020;3:41. 10.1038/s41746-020-0253-3.
54. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. 10.1136/bmj-2022-070904.
55. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, The SPIRIT-AI, CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26:1351–63. 10.1038/s41591-020-1037-7.
56. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI, CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364–74. 10.1038/s41591-020-1034-x.
57. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56. 10.1038/s41591-018-0300-7.

Figures

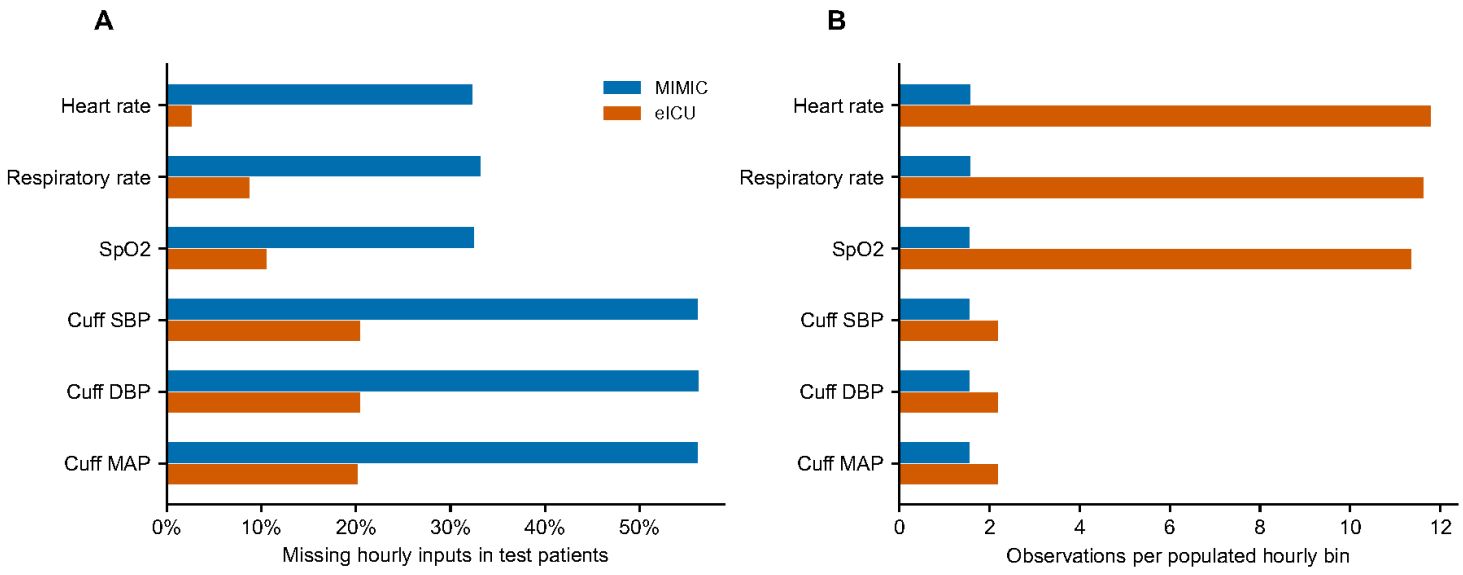


Figure 1

Measurement availability and density. Panel A reports missing hourly inputs among the fixed test patients. Panel B reports mean raw records per populated hourly bin in the eligible index-stay input

windows across roles. One last available measurement is retained per bin in the model input; original acquisition frequencies still differ.

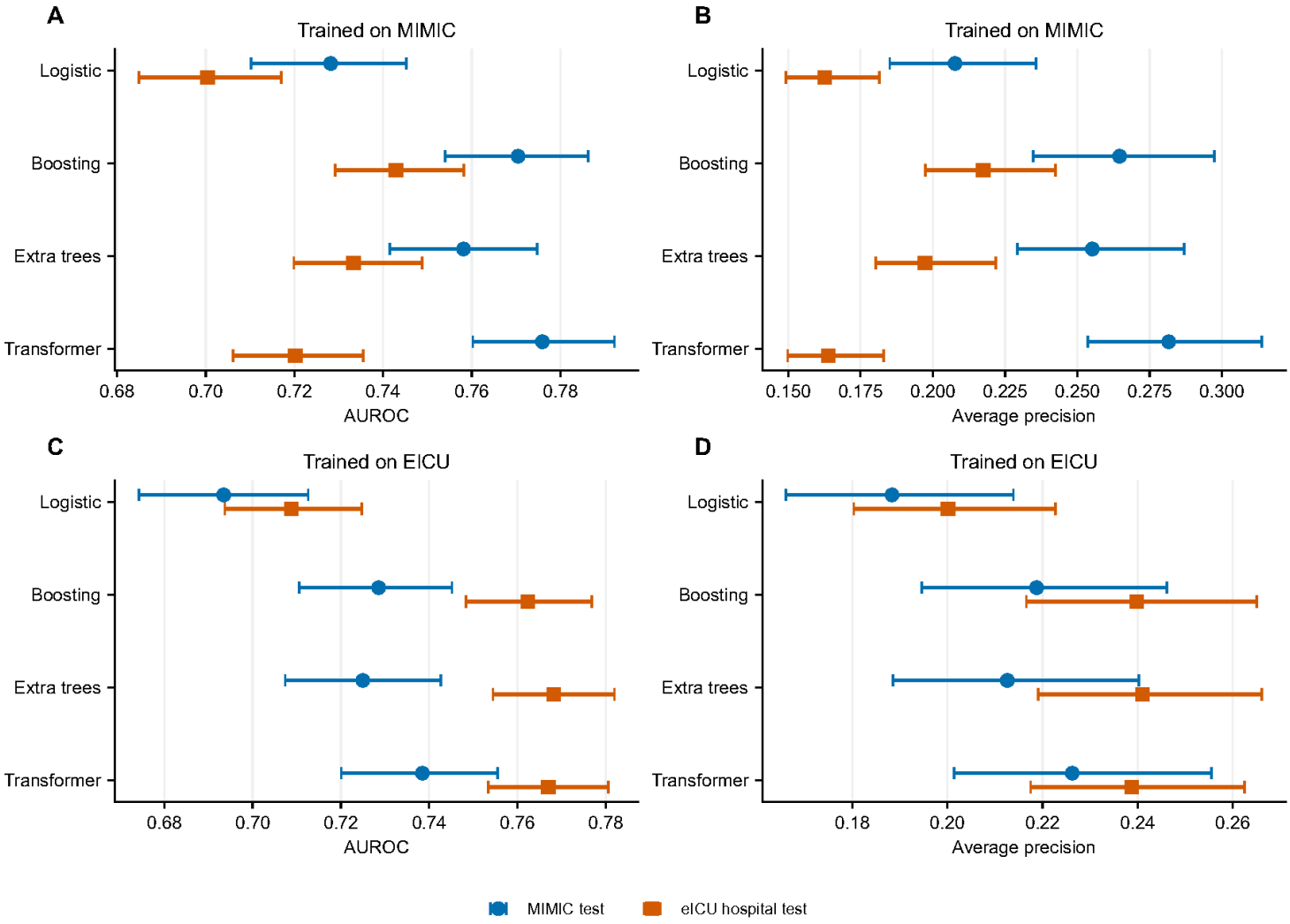


Figure 2

Reciprocal model comparison. Rows identify the training database. Columns show AUROC and average precision. Points and 95% intervals use the same 1,000 patient resamples for all models within each evaluation population. eICU tests use held-out hospitals. Population prevalence accompanies average precision in Table 3.

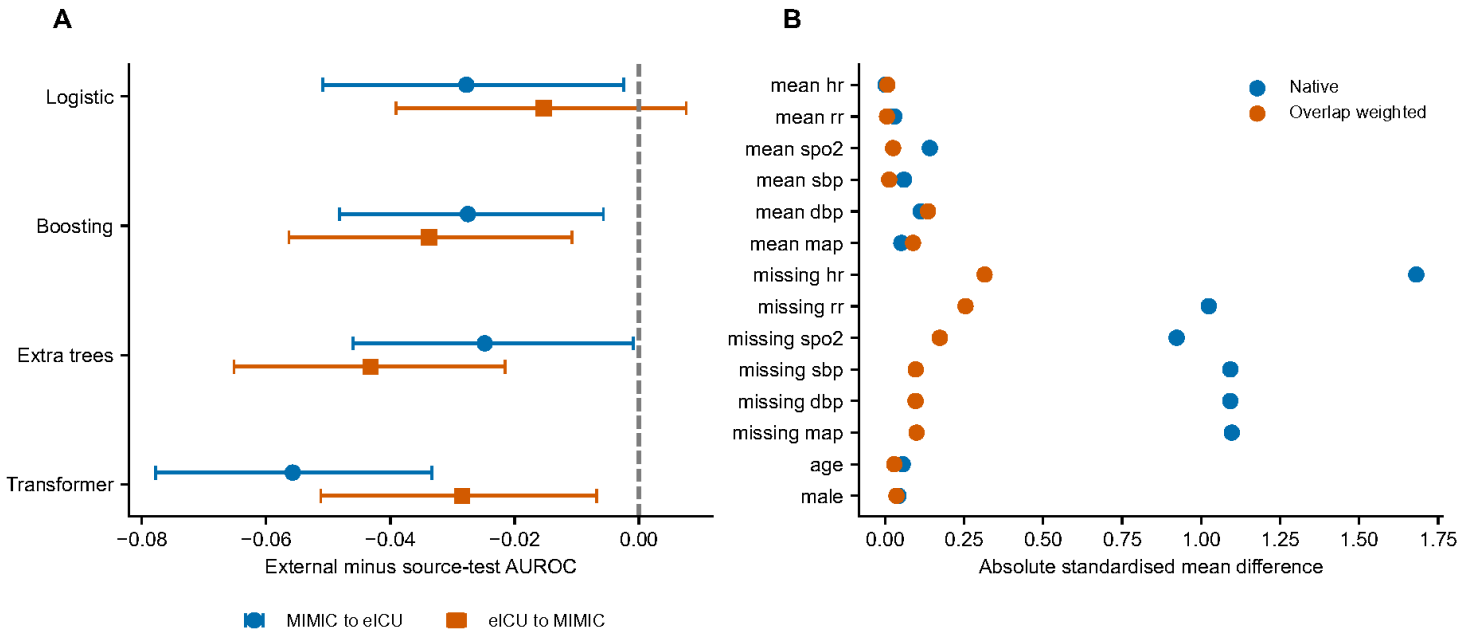


Figure 3

AUROC contrasts and population balance. Panel A estimates external minus source-test AUROC with independent source/target bootstrap samples. Zero is a reference, not an expected scientific result. Panel B compares absolute standardised mean differences before and after descriptive overlap weighting; weighting does not guarantee exact balance or causal identification.

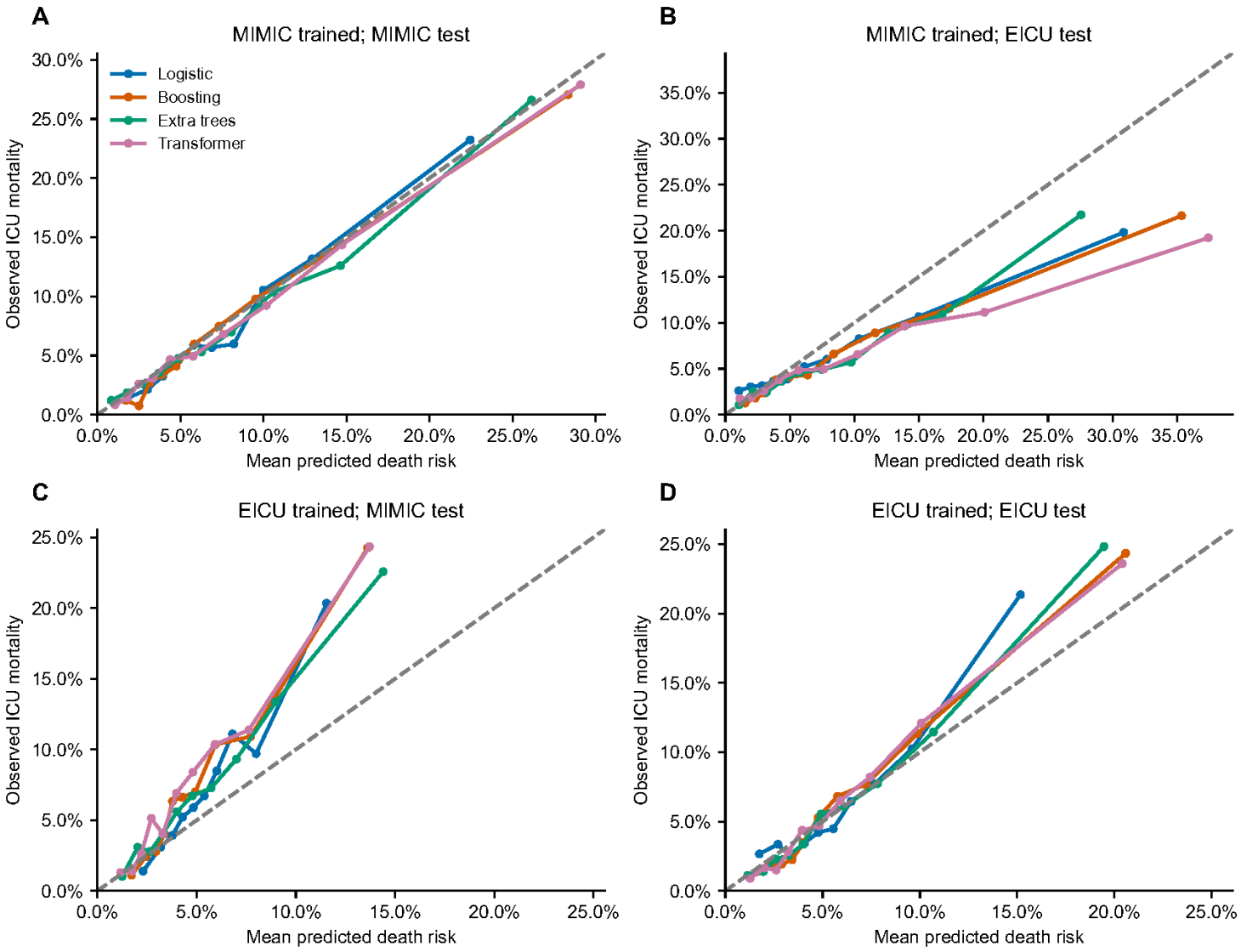


Figure 4

Probability calibration. Each panel identifies the source-trained model and test population. Ten prediction-quantile groups supply mean predicted and observed risks. The diagonal represents ideal calibration. Points are descriptive groups, not confidence intervals. Source-only probabilities are shown.

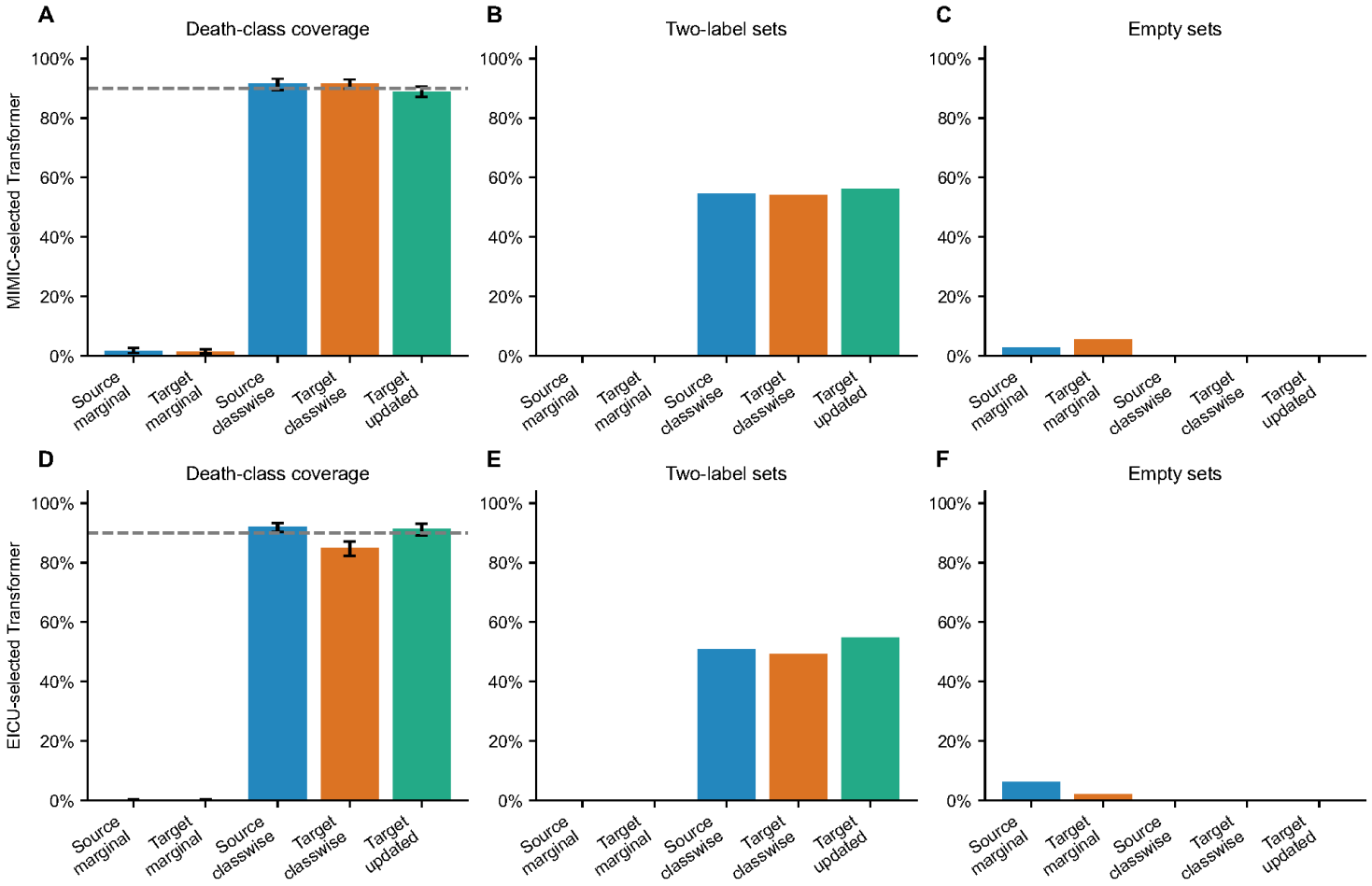


Figure 5

Coverage and informativeness of source-selected prediction sets. Rows identify the source-selected system. Columns show death coverage with Wilson intervals, two-label sets and empty sets. Marginal baseline, class-conditional and target-updated class-conditional settings are distinguished. Dashed lines mark nominal 90% coverage, which is not guaranteed after source or hospital shift. Overall coverage alone is not a success criterion.

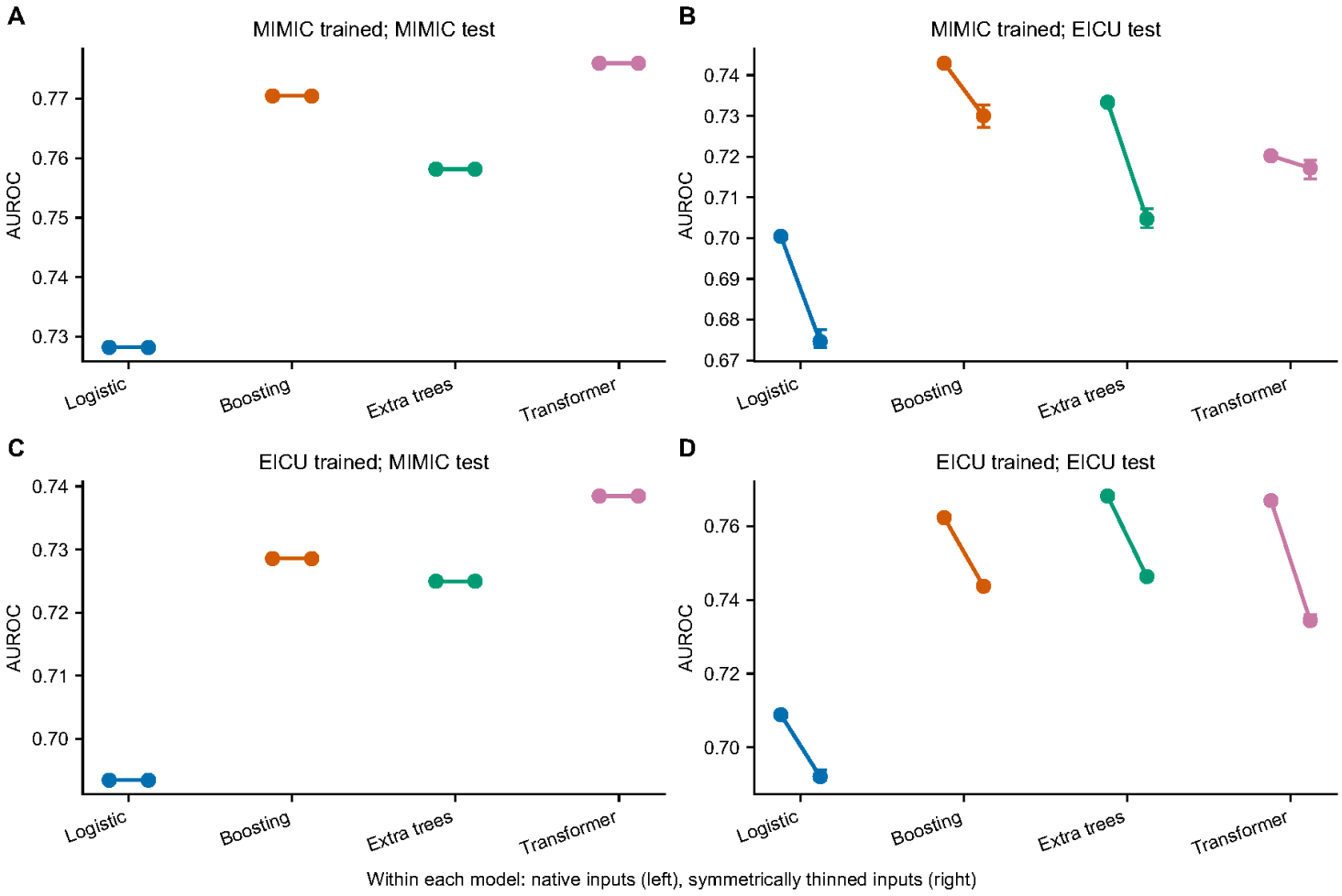


Figure 6

Symmetric measurement-thinning stress test. Each model's native AUROC is linked to its mean AUROC after thinning under the lower development-sample availability in the two databases. Error bars span the three fixed thinning draws, not a confidence interval. Models and preprocessing are frozen. The artificial perturbation is not an estimate of a clinical measurement intervention.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx