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.

Transportability, dataset shift, and local recalibration of early ICU deterioration prediction across MIMIC-IV, eICU-CRD, and NWICU: a retrospective cross-system validation study

Zheng Yao

yaozheng199@outlook.com

Jiaxiang County People's Hospital

Zhaotao Fang

Lvling Town Central Health Center, Heze High-Tech Industrial Development Zone

Shijia Zhu

Wenshang County Second People's Hospital

Research Article

Keywords: critical care, transportability, external validation, calibration, recalibration, dataset shift, machine learning, prediction model

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11054375/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

Clinical prediction models may retain discrimination after transport while producing unreliable absolute risks. We evaluated cross-system transportability of early ICU deterioration models and how local recalibration changed as local outcome-event information accumulated, together with the number of local patients required to accrue those events.

Methods

We developed penalized logistic regression and XGBoost models in MIMIC-IV using 17 predictors from ICU hours 0–24. Eligible adults were alive and still in the ICU at 24 hours and had not received invasive mechanical ventilation and did not have established tracheostomy during the first 24 hours. The primary outcome was ICU death or new invasive mechanical ventilation during hours 24–48; the 24–72-hour composite was secondary. Source models were evaluated without adaptation in eICU-CRD and NWICU. External cohorts were divided into patient-disjoint adaptation and fixed evaluation sets, and intercept-only (M1) or intercept-plus-slope (M2) recalibration was repeatedly estimated from samples targeting 20, 50, 100, or 200 local events.

Results

The MIMIC development cohort contained 31,054 ICU stays; eICU and NWICU contained 90,393 and 16,562 external stays. Primary event prevalences were 3.45%, 2.42%, and 2.75%. External discrimination did not collapse uniformly (AUROC 0.674–0.728), but calibration failures differed by environment and model. In eICU, calibration slopes were 0.893 for logistic regression and 1.003 for XGBoost; in NWICU, they were 1.234 and 1.054. Intercept-only recalibration was often effective for risk-level shift, whereas slope updating became more useful with increasing local event information when slope mismatch was present. At matched event targets, the 24–72-hour composite required 31.5%–32.6% fewer labeled patients in eICU and 33.9%–35.2% fewer in NWICU than the 24–48-hour outcome.

Conclusions

Transport failure was driven more consistently by environment-specific miscalibration than by loss of ranking performance. Recalibration benefit and stability depended on model, environment, failure mode, horizon, and local event information; the results do not support a universal event-count threshold.

Background

Clinical prediction models are commonly evaluated by discrimination, yet deployment decisions also depend on whether predicted probabilities retain their meaning outside the development setting. A model can rank patients reasonably well while systematically overestimating or underestimating absolute risk, or while producing risk estimates that are too compressed or too extreme. Such calibration failures can materially affect threshold-based decisions, resource allocation, and clinical interpretation even when the area under the receiver operating characteristic curve remains acceptable [1, 2].

Transporting an ICU prediction model across institutions is especially challenging because hospitals differ not only in patient case mix but also in measurement practices, documentation workflows, laboratory availability, treatment patterns, and the operational definitions that can be constructed from routinely collected data. Differences in observed covariate distributions and missingness patterns therefore provide evidence of deployment-environment shift, but they do not by themselves establish a change in the conditional outcome mechanism [3].

Local recalibration offers a comparatively low-complexity strategy for adapting an externally developed model without rebuilding it from scratch. An intercept-only update can correct systematic risk-level differences, whereas an intercept-plus-slope update can additionally correct mismatched risk spread [2, 4]. Prior empirical and formal work has examined how update-set size, event rate, model complexity, and the estimand being targeted influence prediction-model updating and recalibration [5, 6, 7, 8]. A deployment-focused question is how prespecified local outcome-event targets translate into out-of-sample recalibration performance and into the number of local patients needed to accrue those events across environments and prediction horizons.

We therefore studied the transportability of early ICU deterioration prediction across MIMIC-IV, eICU-CRD, and NWICU. Our objectives were to characterize raw external discrimination and calibration, quantify dataset and observation-process shift, describe hospital-level heterogeneity within eICU, and determine empirically how the performance of simple local recalibration strategies changes as increasing amounts of local event information become available. We also examined whether the relationship between event count and labeling burden differed when the prediction horizon was extended from 24–48 to 24–72 hours.

Methods

Study design and data sources

We performed a retrospective model-development, external-validation, and local-recalibration study using three ICU data systems. MIMIC-IV served as the source development environment [9]. eICU-CRD served as the primary external validation system and additionally allowed hospital-level deployment-environment analyses [10]. NWICU served as a second external system-level validation environment [11]. Major analytic decisions—including the protocol, predictor definitions, model-development workflow, external-validation plan, and recalibration design—were finalized before the corresponding downstream analyses were performed.

Cohort definition and prediction time

The prediction time, T0, was defined as 24 hours after ICU admission. Eligible patients were adults aged at least 18 years who were alive and remained in the ICU at T0. For hospitalizations containing more than one ICU stay, the earliest ICU stay satisfying the 24-hour eligibility requirement was fixed before ventilation-related exclusions were applied; a later ICU stay was not substituted if that earliest stay was subsequently excluded.

Patients with invasive mechanical ventilation or established tracheostomy during the first 24 hours of the ICU stay were excluded. Database-specific representations of invasive ventilation were harmonized as closely as possible while explicitly excluding noninvasive ventilation.

Outcomes

The prespecified primary outcome was a composite of ICU death or new invasive mechanical ventilation after T0 and within hours 24–48 of the ICU stay. Secondary analyses considered the same composite during hours 24–72 and the individual components of ICU death and new invasive mechanical ventilation at both horizons.

Secondary outcome models used the same prespecified predictor set and fixed modeling framework without outcome-specific hyperparameter retuning. Component risks were not inferred from the primary composite probability, and component-specific recalibration learning curves were not introduced.

Predictors

The primary predictor set consisted of 17 structured variables available across all three systems: age, sex, heart rate, respiratory rate, peripheral oxygen saturation, noninvasive systolic and diastolic blood pressure, temperature, white blood cell count, hemoglobin, platelet count, sodium, potassium, bicarbonate or total carbon dioxide, blood urea nitrogen, creatinine, and glucose.

For continuous physiologic and laboratory measurements, the last physiologically plausible value observed during ICU hours 0–24 and at or before T0 was used. Prespecified physiologic plausibility ranges were applied before aggregation. Values outside those ranges were treated as unavailable rather than clipped or winsorized. Measurement counts and missingness indicators were not included as predictors in the primary models.

Source preprocessing

Preprocessing parameters were estimated exclusively within the MIMIC development data and were then fixed for all validation and external analyses. Continuous variables were imputed using MIMIC-development medians. Sex was imputed using the development mode. Penalized logistic regression used development-set mean and population standard deviation for z-standardization of continuous predictors, while XGBoost used the same imputed values on their original scales. No external median, scaling parameter, or missingness model was learned.

Source development and internal validation

MIMIC patients were split at the patient level into a 75% development cohort and 25% internal validation cohort, preserving patient-level outcome and broad deidentified temporal strata. There was no patient overlap between the two sets.

Two model families were retained throughout the study: penalized logistic regression and XGBoost [12]. Hyperparameters were selected using grouped five-fold cross-validation within the development cohort and were fixed before the internal validation cohort was evaluated. Final primary hyperparameters were C = 0.01 for logistic regression and, for XGBoost, maximum depth 3, learning rate 0.02, 450 trees, minimum child weight 10, subsample 0.9, column subsample 0.9, L2 regularization 10, and L1 regularization 0.

No class weighting, over-sampling, under-sampling, or synthetic minority sampling was used because probability calibration was a central endpoint.

Raw external validation

The fixed source models and source-derived preprocessing parameters were applied directly to eICU and NWICU before any local adaptation. We evaluated area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), Brier score, logarithmic loss, observed-to-expected ratio, calibration-in-the-large (CITL), calibration intercept and slope from joint logistic recalibration, integrated calibration index (ICI), and smooth calibration curves [1, 2, 13]. Validation uncertainty was estimated using patient-clustered bootstrap resampling with 500 replicates and is reported as percentile 95% confidence intervals.

Dataset-shift characterization

We compared predictor distributions, missingness, and outcome prevalence across environments. Continuous distribution shift was summarized using standardized mean differences and Wasserstein distances. Missingness was characterized separately because differences in measurement availability can represent an important observation-process shift.

Patient-grouped domain classifiers were fitted to distinguish each external system from MIMIC development using three representations: source-imputed physiologic values alone, missingness indicators alone, and both combined. These classifiers were used to quantify dataset distinguishability rather than to infer a causal mechanism of transport failure.

Hospital-level heterogeneity

Within eICU, external performance and data-shift summaries were additionally characterized across hospitals. Calibration and discrimination summaries emphasized hospitals with at least 20 outcome events when stable site-level estimation was required. Associations between hospital-level shift measures and calibration error were treated as exploratory and were not used to select models or recalibration strategies. NWICU was analyzed at the system level because no stable hospital/site identifier was available.

Local recalibration

External cohorts were split at the patient level into 60% adaptation-candidate and 40% fixed evaluation sets. The split was defined before recalibration learning curves were generated and was held fixed throughout those analyses.

Three deployment states were evaluated using established recalibration formulations [2, 4]: M0, no update; M1, intercept-only recalibration,

$$\text{logit}(p_{new}) = \alpha + \text{offset}\{\text{logit}(p_{old})\},$$

and M2, intercept-plus-slope recalibration,

$$\text{logit}(p_{new}) = \alpha + \beta\ \text{logit}(p_{old}).$$

For each external database, repeated local adaptation samples targeted approximately 20, 50, 100, or 200 outcome events. Sampling was patient-based. Each event target was repeated 250 times, and all recalibrated models were evaluated on the same fixed external evaluation holdout.

The primary estimands were M1 versus M0 and M2 versus M0. M2-versus-M1 comparisons were retained only as descriptive paired analyses and were not used for post hoc method selection.

Learning-curve summaries report medians and empirical 2.5th to 97.5th percentiles across repeated local adaptation samples. These intervals quantify adaptation-sampling variability and are not population confidence intervals.

Secondary and prediction-horizon analyses

The full internal and raw external validation framework was repeated for the 24–72-hour composite and for death and new invasive mechanical ventilation at both prediction horizons using outcome-specific models with the same prespecified feature set and model frameworks.

A supplementary recalibration learning curve was prespecified only for the 24–72-hour composite because it supported the complete common 20/50/100/200-event grid in both external systems. The 24–48-hour and 24–72-hour composites were then compared at matched event-information targets to quantify the number of local patients and stays required to accrue the same number of outcome events. In this study, patient-labeling burden is therefore an operational measure of the number of local patients whose outcomes must be ascertained to accrue a target event count; it does not imply that outcome ascertainment necessarily requires manual chart review.

Reporting principles

Reporting was guided by TRIPOD + AI recommendations for prediction-model studies using regression or machine-learning methods [14]. Calibration was treated as a central transportability endpoint. Raw external performance was evaluated and recorded before any local recalibration result was examined. Both model families were retained throughout the analyses rather than selecting a post hoc winner. Dataset-distribution differences were described as covariate or observation-process shift and were not interpreted as evidence of a change in the conditional outcome mechanism without additional evidence.

Use of AI-assisted tools in manuscript preparation

OpenAI ChatGPT (GPT-5.6 Sol) was used under author supervision to assist with drafting and checking analysis code, manuscript organization, language editing, reference integration, and formatting. All analysis code was executed by the authors, and numerical results and references used in the manuscript were checked against prespecified analysis outputs and source literature. The tool did not independently access the credentialed source databases, generate or alter primary data, or make autonomous scientific or clinical decisions. The authors reviewed the AI-assisted content and take responsibility for the submitted work.

Results

Cohorts and dataset shift

The MIMIC development cohort contained 31,054 ICU stays from 25,585 patients, with 1,071 primary 24–48 h events (3.45%). The eICU and NWICU external cohorts contained 90,393 and 16,562 stays, with primary event prevalences of 2.42% and 2.75%, respectively. (Table 1; Fig. 1) Observed-value covariate differences were mostly small to moderate: the maximum absolute standardized mean difference versus MIMIC development was 0.206 in eICU and 0.229 in NWICU. (Table 1 / Table S1; Fig. 1B) Observation-process differences were more pronounced. All 17 predictors were observed in 92.1% of MIMIC development stays, 69.9% of eICU stays, and 63.5% of NWICU

stays; 26.9% of NWICU stays had no eligible vital-sign value in the 0–24 h predictor window. (Table 1; Fig. 1A) Patient-grouped domain classifiers distinguished external cohorts from MIMIC more strongly when missingness was combined with physiology. For eICU, pooled AUROC increased from 0.599 using physiology alone and 0.633 using missingness alone to 0.705 using both. For NWICU, corresponding AUROCs were 0.605, 0.670, and 0.749. (Table 1 / Table S2; Fig. 1C)

Table 1
Cohort size, primary outcome prevalence, data completeness, and dataset shift

Environment	Role	N stays	N patients	Events	Event rate (%)	All 6 vitals (%)	No eligible vital (%)	All 9 labs (%)	All 17 predictors (%)
MIMIC development	Source development	31054	25585	1071	3.45	95.3	0.2	96.7	92.1
eICU	External validation	90393	79299	2187	2.42	85.2	0.2	81.2	69.9
NWICU	External validation	16562	14938	455	2.75	72.3	26.9	88.0	63.5

Panel B. Dataset-shift summaries

Environment	Max |SMD| vs MIMIC	Domain AUROC: physiology	Domain AUROC: missingness	Domain AUROC: combined
MIMIC development	0.000			
eICU	0.206	0.599	0.633	0.705
NWICU	0.229	0.605	0.670	0.749

Primary outcome: ICU death or new invasive mechanical ventilation during ICU hours 24-48. Completeness refers to predictors measured during ICU hours 0-24. SMD, standardized mean difference; AUROC, area under the receiver operating characteristic curve.

Internal and raw external validation of the primary outcome

In MIMIC internal validation, AUROC was 0.632 (0.599–0.664) for penalized logistic regression and 0.685 (0.657–0.714) for XGBoost. Calibration slopes were 0.801 (0.615–0.958) and 0.930 (0.798–1.055), respectively. (Table 2 / Table S4; Fig. 2) In eICU, the observed 24–48 h event rate was 2.42%, while mean predicted risk was 3.52% for logistic regression and 3.69% for XGBoost. Corresponding CITL values were − 0.403 (-0.444–-0.360) and − 0.454 (-0.495–-0.411), with calibration slopes 0.893 (0.843–0.940) and 1.003 (0.960–1.049). (Table 2 / Table S4; Fig. 2) In NWICU, the observed 24–48 h event rate was 2.75%, compared with mean predicted risks of 3.26% for logistic regression and 3.01% for XGBoost. CITL was − 0.180 (-0.278–-0.100) and − 0.099 (-0.196–-0.016); calibration slopes were 1.234 (1.125–1.339) and 1.054 (0.945–1.163). (Table 2 / Table S4; Fig. 2) External discrimination did not show a uniform collapse: eICU AUROCs were 0.674 (0.662–0.687) for logistic regression and 0.719 (0.707–0.731) for XGBoost, whereas NWICU AUROCs were 0.728 (0.703–0.751) and 0.702 (0.675–0.732). The dominant

transportability failure was therefore not a uniform loss of ranking performance but environment-specific miscalibration. (Table 2 / Table S4; Fig. 2)

Table 2
Internal and raw external validation of the primary 24–48 h deterioration models

Environment	Model	Stays	Events	Observed event rate (%)	Mean predicted risk (%)	AUROC	AUPRC
MIMIC internal validation	Penalized logistic regression	10256	356	3.47	3.39	0.632 (0.599–0.664)	0.084 (0.066–0.109)
MIMIC internal validation	XGBoost	10256	356	3.47	3.32	0.685 (0.657–0.714)	0.126 (0.097–0.164)
eICU	Penalized logistic regression	90393	2187	2.42	3.52	0.674 (0.662–0.687)	0.084 (0.074–0.095)
eICU	XGBoost	90393	2187	2.42	3.69	0.719 (0.707–0.731)	0.118 (0.105–0.132)
NWICU	Penalized logistic regression	16562	455	2.75	3.26	0.728 (0.703–0.751)	0.128 (0.101–0.158)
NWICU	XGBoost	16562	455	2.75	3.01	0.702 (0.675–0.732)	0.158 (0.126–0.192)

Panel B. Calibration and scoring

Environment	Model	Brier score	Log loss	O/E	CITL	Calibration slope	ICI
MIMIC internal validation	Penalized logistic regression	0.033 (0.030–0.036)	0.146 (0.135–0.158)	1.025 (0.930–1.128)	0.027 (-0.077–0.128)	0.801 (0.615–0.958)	0.005 (0.003–0.009)
MIMIC internal validation	XGBoost	0.032 (0.029–0.035)	0.141 (0.130–0.152)	1.045 (0.946–1.147)	0.047 (-0.059–0.148)	0.930 (0.798–1.055)	0.003 (0.002–0.007)
eICU	Penalized logistic regression	0.023 (0.022–0.024)	0.109 (0.107–0.113)	0.686 (0.661–0.714)	-0.403 (-0.444–-0.360)	0.893 (0.843–0.940)	0.012 (0.011–0.013)
eICU	XGBoost	0.023 (0.022–0.024)	0.106 (0.103–0.109)	0.657 (0.632–0.684)	-0.454 (-0.495–-0.411)	1.003 (0.960–1.049)	0.013 (0.012–0.014)
NWICU	Penalized logistic regression	0.025 (0.023–0.027)	0.115 (0.107–0.122)	0.844 (0.768–0.911)	-0.180 (-0.278–-0.100)	1.234 (1.125–1.339)	0.006 (0.004–0.009)
NWICU	XGBoost	0.025 (0.023–0.027)	0.114 (0.106–0.122)	0.912 (0.832–0.985)	-0.099 (-0.196–-0.016)	1.054 (0.945–1.163)	0.005 (0.004–0.007)

Values in parentheses are patient-clustered 95% bootstrap percentile intervals. O/E, observed-to-expected ratio; CITL, calibration-in-the-large; ICI, integrated calibration index.

Hospital-level deployment heterogeneity

Among the 33 eICU hospitals with at least 20 primary outcome events, median logistic-regression AUROC was 0.652, median CITL 0.055, and median calibration slope 0.815. For XGBoost, corresponding medians were 0.684, 0.013, and 0.887. (Table S3a; Figure S1) Most prespecified exploratory site shift–calibration correlations were weak: only 2 of 24 correlations had nominal $P < 0.05$, and these correlation analyses were not used for model or recalibration selection. (Table S3b; Figure S1)

Local recalibration of the primary 24–48 h outcome

In eICU logistic regression, intercept-only recalibration was already effective with 20 local events: the empirical probabilities of lower log loss and lower ICI versus M0 were 0.960 and 0.964. Because the raw calibration slope was 0.899, M2 provided additional slope correction increasingly as local information accumulated; its probability of reducing absolute slope error rose from 0.272 at 20 events to 0.780 at 200. (Table 3 / Table S5; Fig. 3) For eICU XGBoost, the raw calibration slope was approximately one (1.005). At 20 events, M1 reduced log loss and ICI versus M0 in 0.968 and 0.968 of repeated samples, whereas M2 reduced absolute slope error in only 0.024. (Table 3 / Table S5; Fig. 3) For NWICU logistic regression, the raw calibration slope was 1.253. At 20 events, the probability that M2 reduced absolute slope error versus M0 was 0.712; at 200 events it was 1.000. This increasing stability with additional local events was also accompanied by probabilities of lower log loss of 0.556 and 1.000, respectively. (Table 3 / Table S5; Fig. 3) For NWICU XGBoost, the raw calibration slope was already close to one (1.029). At 20 events, the probability of lower log loss versus M0 was 0.388 for M1 but only 0.112 for M2. At 200 events, these probabilities increased to 0.984 and 0.904, respectively. (Table 3 / Table S5; Fig. 3)

Table 3
Local recalibration at low and high local event-information levels

Database	Source model	Method	Target events	Median patients	Median stays	Median alpha	Median beta (2.5th-97.5th)
NWICU	Penalized logistic regression	M1 intercept only	20	682.0	754.0	-0.211	1.000 (1.000–1.000)
NWICU	Penalized logistic regression	M1 intercept only	200	6549.0	7261.0	-0.182	1.000 (1.000–1.000)
NWICU	Penalized logistic regression	M2 intercept slope	20	682.0	754.0	0.425	1.200 (0.676–1.773)
NWICU	Penalized logistic regression	M2 intercept slope	200	6549.0	7261.0	0.494	1.220 (1.131–1.308)
NWICU	XGBoost	M1 intercept only	20	682.0	754.0	-0.127	1.000 (1.000–1.000)
NWICU	XGBoost	M1 intercept only	200	6549.0	7261.0	-0.097	1.000 (1.000–1.000)
NWICU	XGBoost	M2 intercept slope	20	682.0	754.0	0.008	1.039 (0.565–1.601)
NWICU	XGBoost	M2 intercept slope	200	6549.0	7261.0	0.104	1.067 (0.978–1.158)
eICU	Penalized logistic regression	M1 intercept only	20	719.5	821.5	-0.380	1.000 (1.000–1.000)
eICU	Penalized logistic regression	M1 intercept only	200	7193.0	8220.5	-0.396	1.000 (1.000–1.000)
eICU	Penalized logistic regression	M2 intercept slope	20	719.5	821.5	-0.826	0.875 (0.285–1.490)
eICU	Penalized logistic regression	M2 intercept slope	200	7193.0	8220.5	-0.756	0.880 (0.751–1.027)
eICU	XGBoost	M1 intercept only	20	719.5	821.5	-0.429	1.000 (1.000–1.000)
eICU	XGBoost	M1 intercept only	200	7193.0	8220.5	-0.447	1.000 (1.000–1.000)
eICU	XGBoost	M2 intercept	20	719.5	821.5	-0.413	1.011 (0.563–

Database	Source model	Method	Target events	Median patients	Median stays	Median alpha	Median beta (2.5th-97.5th)
		slope					1.477)
eICU	XGBoost	M2 intercept slope	200	7193.0	8220.5	-0.455	0.994 (0.875–1.119)

Panel B. Fixed-evaluation calibration summaries

Database	Source model	Method	Events	M0 CITL	M0 slope	Updated CITL	Updated slope	Updated ICI
NWICU	Penalized logistic regression	M1 intercept only	20	-0.175	1.253	0.036	1.253	0.005
NWICU	Penalized logistic regression	M1 intercept only	200	-0.175	1.253	0.007	1.253	0.004
NWICU	Penalized logistic regression	M2 intercept slope	20	-0.175	1.253	0.040	1.044	0.006
NWICU	Penalized logistic regression	M2 intercept slope	200	-0.175	1.253	0.011	1.027	0.002
NWICU	XGBoost	M1 intercept only	20	-0.101	1.029	0.027	1.029	0.005
NWICU	XGBoost	M1 intercept only	200	-0.101	1.029	-0.004	1.029	0.005
NWICU	XGBoost	M2 intercept slope	20	-0.101	1.029	0.028	0.991	0.007
NWICU	XGBoost	M2 intercept slope	200	-0.101	1.029	-0.005	0.965	0.005
eICU	Penalized logistic regression	M1 intercept only	20	-0.408	0.899	-0.028	0.899	0.004
eICU	Penalized logistic regression	M1 intercept only	200	-0.408	0.899	-0.012	0.899	0.003
eICU	Penalized logistic regression	M2 intercept slope	20	-0.408	0.899	-0.021	1.027	0.006
eICU	Penalized logistic regression	M2 intercept slope	200	-0.408	0.899	-0.012	1.021	0.003
eICU	XGBoost	M1 intercept only	20	-0.462	1.005	-0.033	1.005	0.004
eICU	XGBoost	M1 intercept only	200	-0.462	1.005	-0.015	1.005	0.003
eICU	XGBoost	M2 intercept slope	20	-0.462	1.005	-0.028	0.995	0.005

Database	Source model	Method	Events	M0 CITL	M0 slope	Updated CITL	Updated slope	Updated ICI
eICU	XGBoost	M2 intercept slope	200	-0.462	1.005	-0.017	1.012	0.003

Panel C. Probability of improvement versus M0

Database	Source model	Method	Events	Lower log loss	Lower Brier	Lower |CITL|	Lower |slope error|	Lower ICI
NWICU	Penalized logistic regression	M1 intercept only	20	0.664	0.036	0.652	0.000	0.712
NWICU	Penalized logistic regression	M1 intercept only	200	1.000	0.000	1.000	0.000	1.000
NWICU	Penalized logistic regression	M2 intercept slope	20	0.556	0.364	0.632	0.712	0.620
NWICU	Penalized logistic regression	M2 intercept slope	200	1.000	1.000	1.000	1.000	1.000
NWICU	XGBoost	M1 intercept only	20	0.388	0.080	0.384	0.000	0.608
NWICU	XGBoost	M1 intercept only	200	0.984	0.088	0.980	0.000	0.984
NWICU	XGBoost	M2 intercept slope	20	0.112	0.336	0.400	0.108	0.332
NWICU	XGBoost	M2 intercept slope	200	0.904	0.916	0.980	0.348	0.880
eICU	Penalized logistic regression	M1 intercept only	20	0.960	0.964	0.960	0.000	0.964
eICU	Penalized logistic regression	M1 intercept only	200	1.000	1.000	1.000	0.000	1.000
eICU	Penalized logistic regression	M2 intercept slope	20	0.824	0.816	0.956	0.272	0.936
eICU	Penalized logistic regression	M2 intercept slope	200	1.000	1.000	1.000	0.780	1.000
eICU	XGBoost	M1 intercept only	20	0.968	0.968	0.968	0.000	0.968
eICU	XGBoost	M1 intercept only	200	1.000	1.000	1.000	0.000	1.000
eICU	XGBoost	M2 intercept slope	20	0.888	0.800	0.968	0.024	0.944

Database	Source model	Method	Events	Lower log loss	Lower Brier	Lower |CITL|	Lower |slope error|	Lower ICI
eICU	XGBoost	M2 intercept slope	200	1.000	1.000	1.000	0.064	1.000

M0, no local update; M1, intercept-only recalibration; M2, intercept-plus-slope recalibration. Beta intervals are empirical 2.5th-97.5th percentiles across 250 repeated adaptation samples and are not population confidence intervals. Improvement probabilities are the fractions of repeated adaptation samples in which the updated model improved the stated metric relative to fixed M0.

Secondary outcomes and prediction-horizon sensitivity

For the secondary 24–72 h composite, eICU event prevalence was 3.55% while mean predicted risks were 5.21% for logistic regression and 5.45% for XGBoost. Calibration slopes were 0.945 (0.903–0.990) and 1.028 (0.990–1.066). (Table S6 / Table S8; Figure S2 / Figure S3) For the 24–72 h composite in NWICU, event prevalence was 4.08% and mean predicted risks were 4.90% for logistic regression and 4.60% for XGBoost. Calibration slopes were 1.194 (1.091–1.299) and 1.043 (0.948–1.139). (Table S6 / Table S8; Figure S2 / Figure S3) Component analyses showed different calibration directions across external systems. In eICU, new invasive mechanical ventilation was overpredicted at both horizons; for the logistic model, observed versus predicted risks were 1.70% versus 2.84% at 24–48 h and 2.43% versus 4.14% at 24–72 h. (Table S6; Figure S2) In NWICU, death tended to be underpredicted while new invasive mechanical ventilation was overpredicted. For logistic regression, observed versus predicted death risks were 0.89% versus 0.67% at 24–48 h and 1.42% versus 1.15% at 24–72 h; corresponding new-IMV risks were 1.93% versus 2.73% and 2.84% versus 4.03%. (Table S6; Figure S2) At identical local event targets, the higher-prevalence 24–72 h composite required fewer labeled patients than the primary 24–48 h composite. The median patient reduction ranged from 31.5% to 32.6% in eICU and 33.9% to 35.2% in NWICU across the 20, 50, 100, and 200 event targets. (Table S9; Figure S4)

Overall empirical synthesis

Across prediction horizons, databases, and model families, the benefit and stability of local recalibration depended on the form of calibration failure and the amount of local event information. The empirical results therefore do not support a universal local event threshold for recalibration. (Tables 3, S5, S7–S10; Figs. 3, S3–S5)

Discussion

In this cross-system study of early ICU deterioration prediction, the most consistent deployment problem was not a uniform loss of discrimination but a failure of absolute risk estimates to transport reliably. The magnitude and form of that calibration failure differed across external environments and model families. eICU showed substantial overprediction for the primary composite, whereas the NWICU logistic model additionally showed a calibration slope materially above one. These differences mattered because they determined which form of local recalibration was most efficient and how much local outcome information was needed before an additional slope parameter became stable.

The raw external results illustrate why discrimination alone is insufficient for transportability assessment. AUROC remained broadly similar to internal performance in several external settings, yet calibration-in-the-large and mean predicted risk showed clinically meaningful disagreement with observed event prevalence. A deployment evaluation based only on ranking performance could therefore have missed the principal failure mode. This distinction is particularly relevant when a model is intended to produce interpretable risk estimates rather than only relative ranking [1, 2, 15].

The dataset-shift analyses also showed that the observation process was a prominent and qualitatively distinct dimension of cross-system difference. Observed-value standardized mean differences were generally small to moderate, while missingness and predictor-panel completeness differed substantially, especially in NWICU. Domain classifiers became more discriminative when missingness was combined with physiology. These findings support treating measurement availability as part of the deployment environment rather than merely as a nuisance to be imputed away. They do not establish that observation-process differences caused miscalibration or that the conditional relationship between predictors and outcome changed.

The recalibration learning curves provide a practical view of local adaptation. When the raw calibration slope was already close to one, intercept-only updating often delivered substantial improvement with relatively little local event information. This pattern was especially clear for eICU XGBoost. In contrast, when the raw slope materially differed from one, as for NWICU logistic regression, the two-parameter update became increasingly stable and useful as more local events accumulated. At low event counts, estimating the additional slope parameter could add variance without corresponding improvement, particularly when the unadapted slope was already near one.

These findings extend prior work on prediction-model updating. Empirical studies have shown that updating performance depends on the size and composition of the updating dataset and that more complex updates can become unstable when information is limited [5, 16, 17]. Formal frameworks likewise make updating sample-size requirements dependent on the parameters, estimands, and precision or stability objectives being targeted [6, 7, 8]. Brown et al. provided a close deployment precedent by examining whether individual hospitals had sufficient local data for site-specific recalibration [18], while learning-curve approaches have been proposed to assess prediction-model stability as data accumulate [7, 19]. ICU-TRANS adds a deployment-focused empirical perspective by indexing out-of-sample recalibration performance to locally observed outcome events and translating those event targets into the corresponding local patient requirement across external ICU systems and prediction horizons. These event-indexed learning curves address a different question from formal minimum sample sizes selected for prespecified precision or stability objectives [6, 7, 8]; neither approach defines a universal M1/M2 event threshold.

A recent medRxiv preprint introduced post-hoc retrieval augmentation (PRAM), which adapts a frozen ICU prediction model using outcomes from similar patients in a local reference bank and compares retrieval with other local adaptation strategies across hospitals [24]. This represents a complementary adaptation paradigm rather than a direct test of the question addressed here: PRAM can modify patient-level predictions through retrieval, whereas ICU-TRANS isolates low-dimensional probability-scale recalibration and asks how its out-of-sample behavior changes as locally observed outcome events accumulate. PRAM was not part of our prespecified comparison set, and we therefore did not add it retrospectively as an experimental benchmark.

These findings argue against treating recalibration sample size as a single universal number. The relevant requirement depends on the deployment environment, model, form of calibration error, outcome prevalence, and prediction horizon. A fixed event count can correspond to very different numbers of local patients across settings. Extending the composite outcome horizon from 24–48 to 24–72 hours increased event prevalence and reduced

the median number of local patients required to accrue the same event target by roughly one-third in both external systems. Thus, event information and the corresponding local data-ascertainment requirement should be reported separately.

The component analyses further demonstrate why composite calibration can conceal heterogeneous failure mechanisms. In eICU, new invasive mechanical ventilation was consistently overpredicted. In NWICU, death tended to be underpredicted while new invasive mechanical ventilation was overpredicted. Opposing component-level errors can therefore partially cancel in a composite outcome even when the composite itself appears less severely miscalibrated. These analyses were used to characterize component-specific failure patterns rather than to introduce post hoc component-specific adaptation.

Hospital-level results within eICU showed additional deployment heterogeneity. Even among sites with at least 20 events, discrimination and calibration varied materially. Most exploratory correlations between simple site-level shift measures and calibration error were weak. No single summary of marginal predictor or missingness shift showed a strong or consistent relationship with site-level calibration error. More detailed study of local workflows, measurement processes, and treatment practices may be needed to understand why the same model fails differently across hospitals [3, 20, 21].

This study has several strengths. The analysis separated source development, raw external validation, and local adaptation; fixed model and preprocessing choices before external analyses; examined two structurally different model families; emphasized calibration and proper scoring rules alongside discrimination; evaluated two independent external systems; and quantified recalibration performance together with the local patient requirement across prespecified event-information targets. The use of hospital-level eICU environments further allowed transportability to be studied at a level closer to real deployment than a single pooled external cohort.

Several limitations should be considered. First, the study was retrospective and did not test prospective clinical use, workflow effects, or decision consequences. Second, despite harmonized definitions, invasive ventilation and other outcomes necessarily relied on database-specific representations, which may introduce phenotype differences across systems. Third, transportability was evaluated directionally from a single MIMIC development environment; results may differ when another health system serves as the source. Fourth, NWICU lacked a stable hospital-level identifier, and the eICU hospital-level analyses were event-limited and exploratory. Fifth, the deliberately common 17-variable predictor set favored cross-database transportability over maximal source-model complexity. Sixth, secondary outcome models reused the prespecified model framework and hyperparameters rather than undergoing outcome-specific optimization. Seventh, the primary adaptation analysis focused on intercept- and slope-based recalibration rather than broader local refitting or newer post-hoc adaptation paradigms such as retrieval augmentation [24]. Eighth, the empirical recalibration percentiles quantify adaptation-sample variability conditional on the same fixed evaluation holdout; they do not incorporate evaluation-sample variability and should not be interpreted as population confidence intervals. The observed local patient requirements are likewise empirical for these cohorts, outcomes, and horizons rather than universal planning rules. Finally, the analyses identify deployment-environment differences and calibration failure patterns but do not establish causal explanations for those differences.

Based on these empirical findings, we propose an evidence-informed deployment framework rather than an automated switching algorithm. First, a site should quantify unadapted performance (M0) on local labeled data and characterize whether miscalibration is dominated by risk-level shift or also includes a material slope mismatch, with attention to the uncertainty of the slope estimate. Second, update complexity should be matched

to the available local outcome information: when information is limited or slope mismatch is imprecisely estimated, the lower-parameter M1 update is the more conservative starting point; M2 becomes more defensible as local information accumulates and slope mismatch is both material and estimated with adequate precision. Third, any updated model should be evaluated on a patient-disjoint fixed holdout or an equivalent independent local evaluation set. The 20, 50, 100, and 200 event targets used in this study are empirical learning-curve design points, not universal clinical cut-offs. This framework is intended as actionable guidance derived from the observed patterns, not as a validated new recalibration algorithm.

Conclusions

Early ICU deterioration models did not fail uniformly across health systems. Ranking performance was often retained, whereas absolute risks showed environment- and model-specific calibration error. Intercept-only updating often improved calibration when the dominant error was a baseline-risk shift, whereas slope updating required more local event information and was most useful when a material slope mismatch was present. The patient burden required to obtain that information also depended on outcome prevalence and prediction horizon. Local recalibration should therefore be evaluated empirically for the intended deployment environment rather than reduced to a universal event-count threshold.

Abbreviations

AUPRC
area under the precision-recall curve
AUROC
area under the receiver operating characteristic curve
CITL
calibration-in-the-large
eICU-CRD
eICU Collaborative Research Database
ICI
integrated calibration index
ICU
intensive care unit
MIMIC-IV
Medical Information Mart for Intensive Care IV
M0
no local update
M1
intercept-only recalibration
M2
intercept-plus-slope recalibration
NWICU
Northwestern ICU database
SMD
standardized mean difference.

Declarations

Ethics approval and consent to participate

This study was a secondary analysis of existing deidentified critical-care databases and involved no direct participant contact or intervention. The collection of patient information and creation of MIMIC-IV were reviewed by the Institutional Review Board of Beth Israel Deaconess Medical Center, which approved the data-sharing initiative and granted a waiver of informed consent [22]. The eICU-CRD is deidentified in accordance with the HIPAA Safe Harbor standard; access requires human-research training and a data use agreement [23]. The Northwestern ICU (NWICU) project was approved by the Northwestern University Institutional Review Board, which granted a waiver of informed consent, and release of data was authorized by the Northwestern Medicine Data Steward [11]. The authors used the databases in accordance with the applicable PhysioNet credentialing and data-use requirements. No direct identifiers were available to the authors.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed in this study are available through PhysioNet under the applicable credentialing, training, and data-use requirements: MIMIC-IV version 3.1 [22], eICU-CRD version 2.0 [23], and NWICU version 0.1.0 [11]. No new participant-level dataset was generated or redistributed by the authors. The manuscript tables and supplementary materials report aggregate study results.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study.

Authors' contributions

ZY conceived the study, developed the protocol, curated the analytic datasets, performed the statistical analyses and model development, interpreted the results, prepared the tables and figures, and drafted the manuscript. ZF contributed to clinical interpretation, review of study definitions, and critical revision of the manuscript. SZ contributed to clinical interpretation, quality review, and critical revision of the manuscript. ZY supervised the work, serves as guarantor, and is the corresponding author. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, Bonnett LJ, Collins GS. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. 10.1136/bmj-2023-074820.
2. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. 2022;22:316. 10.1186/s12874-022-01801-8.
3. 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.
4. Schnellinger EM, Yang W, Kimmel SE. Comparison of dynamic updating strategies for clinical prediction models. Diagn Progn Res. 2021;5:20. 10.1186/s41512-021-00110-w.
5. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004;23(16):2567–86. 10.1002/sim.1844.
6. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins GS, van Smeden M. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. 10.1136/bmj.m441.
7. Riley RD, Collins GS, Whittle R, Archer L, Snell KIE, Dhiman P, Kirton L, Legha A, Liu X, Denniston AK, Harrell FE Jr, Wynants L, Martin GP, Ensor J. A decomposition of Fisher's information to inform sample size for developing or updating fair and precise clinical prediction models for individual risk—part 1: binary outcomes. Diagn Progn Res. 2025;9:14. 10.1186/s41512-025-00193-9.
8. Riley RD, Whittle R, Sadatsafavi M, Martin GP, Pate A, Collins GS, Ensor J. A general sample size framework for developing or updating a predictive algorithm: with application to clinical prediction models. BMC Med Res Methodol. 2026;26:121. 10.1186/s12874-026-02856-7.
9. Johnson AEW, Bulgarelli L, Shen L, Gayles L, 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.
10. 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.
11. Moukheiber D, Temps W, Molgi B et al. Northwestern ICU (NWICU) database (version 0.1.0). PhysioNet; 2024. version 0.1.0; RRID:SCR_007345. 10.13026/s84w-1829
12. Chen T, Guestrin C, XGBoost:. A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 785–794.
13. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med. 2019;38:4051–65. 10.1002/sim.8281.
14. 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.

15. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, 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.
16. Mishra A, McClelland RL, Inoue LYT, Kerr KF. Recalibration Methods for Improved Clinical Utility of Risk Scores. Med Decis Mak. 2022;42(4):500–12. 10.1177/0272989X211044697.
17. Tanner KT, Keogh RH, Coupland CAC, Hippisley-Cox J, Diaz-Ordaz K. Dynamic updating of clinical survival prediction models in a changing environment. Diagn Progn Res. 2023;7:24. 10.1186/s41512-023-00163-z.
18. Brown KE, Malin BA, Davis SE. Community medical centers struggle to produce well-calibrated clinical prediction models: Data augmentation can help. J Biomed Inf. 2026;177:105012. 10.1016/j.jbi.2026.105012.
19. Legha A, Ensor J, Whittle R, Archer L, Van Calster B, Christodoulou E, Snell KIE, Sadatsafavi M, Collins GS, Riley RD. Sequential sample size calculations and learning curves safeguard the robust development of a clinical prediction model for individuals. J Clin Epidemiol. 2026;191:112117. 10.1016/j.jclinepi.2025.112117.
20. Rockenschaub P, Hilbert A, Kossen T, Elbers P, von Dincklage F, Madai VI, Frey D. The Impact of Multi-Institution Datasets on the Generalizability of Machine Learning Prediction Models in the ICU. Crit Care Med. 2024;52(11):1710–21. 10.1097/CCM.0000000000006359.
21. Eickelberg G, Sanchez-Pinto LN, Kline AS, Luo Y. Transportability of bacterial infection prediction models for critically ill patients. J Am Med Inf Assoc. 2024;31(1):98–108. 10.1093/jamia/ocad174.
22. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet. 2024. RRID:SCR_007345. 10.13026/kpb9-mt58
23. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. RRID:SCR_007345. 10.13026/C2WM1R
24. Jeong I, Lee T, Kim B, Park JH, Kim Y, Lee H. medRxiv. 2026. 10.64898/2026.04.03.26350132. . Version 2 posted August 19, 2026. Preprint. PRAM: post-hoc retrieval augmentation adapts intensive care prediction models without retraining and can be shared across hospitals.

Figures

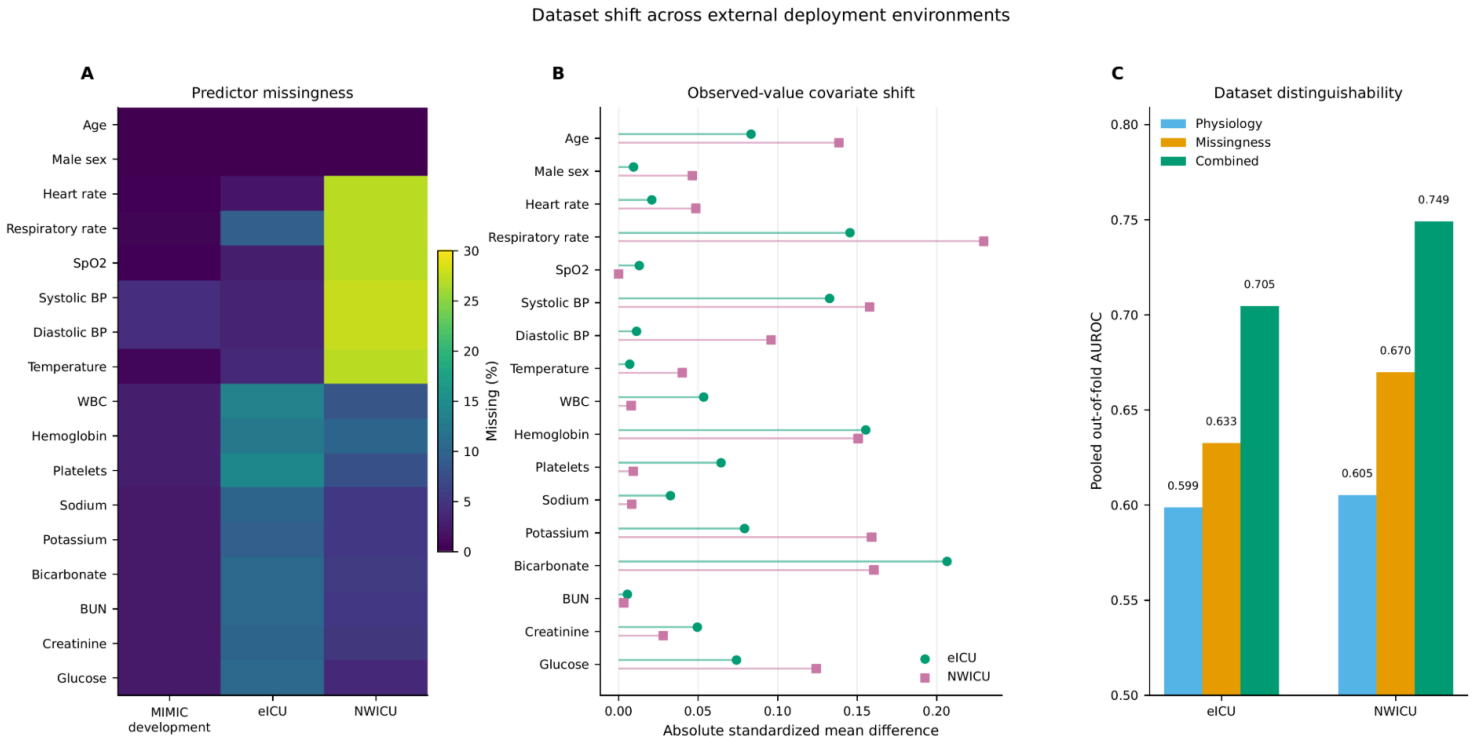


Figure 1

Dataset shift across external deployment environments. Dataset shift across external deployment environments. Panel A shows predictor missingness in the MIMIC development, eICU, and NWICU cohorts. Panel B shows absolute standardized mean differences for observed predictor values in eICU and NWICU relative to MIMIC development. Panel C shows pooled out-of-fold AUROC from patient-grouped domain classifiers using source-imputed physiology, missingness patterns alone, or both. These analyses characterize covariate and observation-process shift.

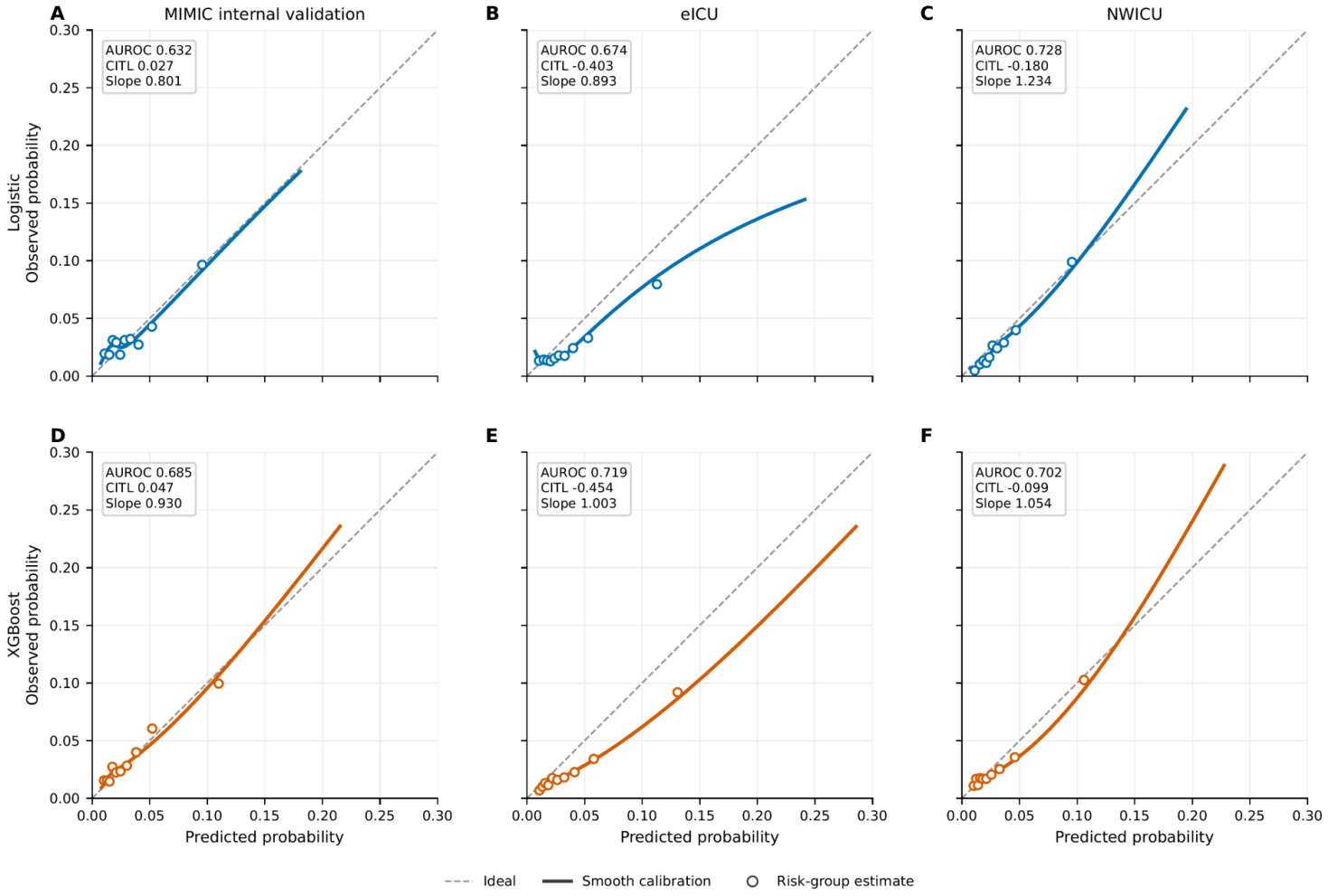


Figure 2

Primary model calibration before local adaptation. Primary 24-48 h calibration before local adaptation. Smooth calibration curves and risk-group estimates are shown for the penalized logistic regression and XGBoost models in MIMIC internal validation, eICU, and NWICU. The dashed diagonal denotes perfect calibration. AUROC, calibration-in-the-large (CITL), and calibration slope are shown within each panel.

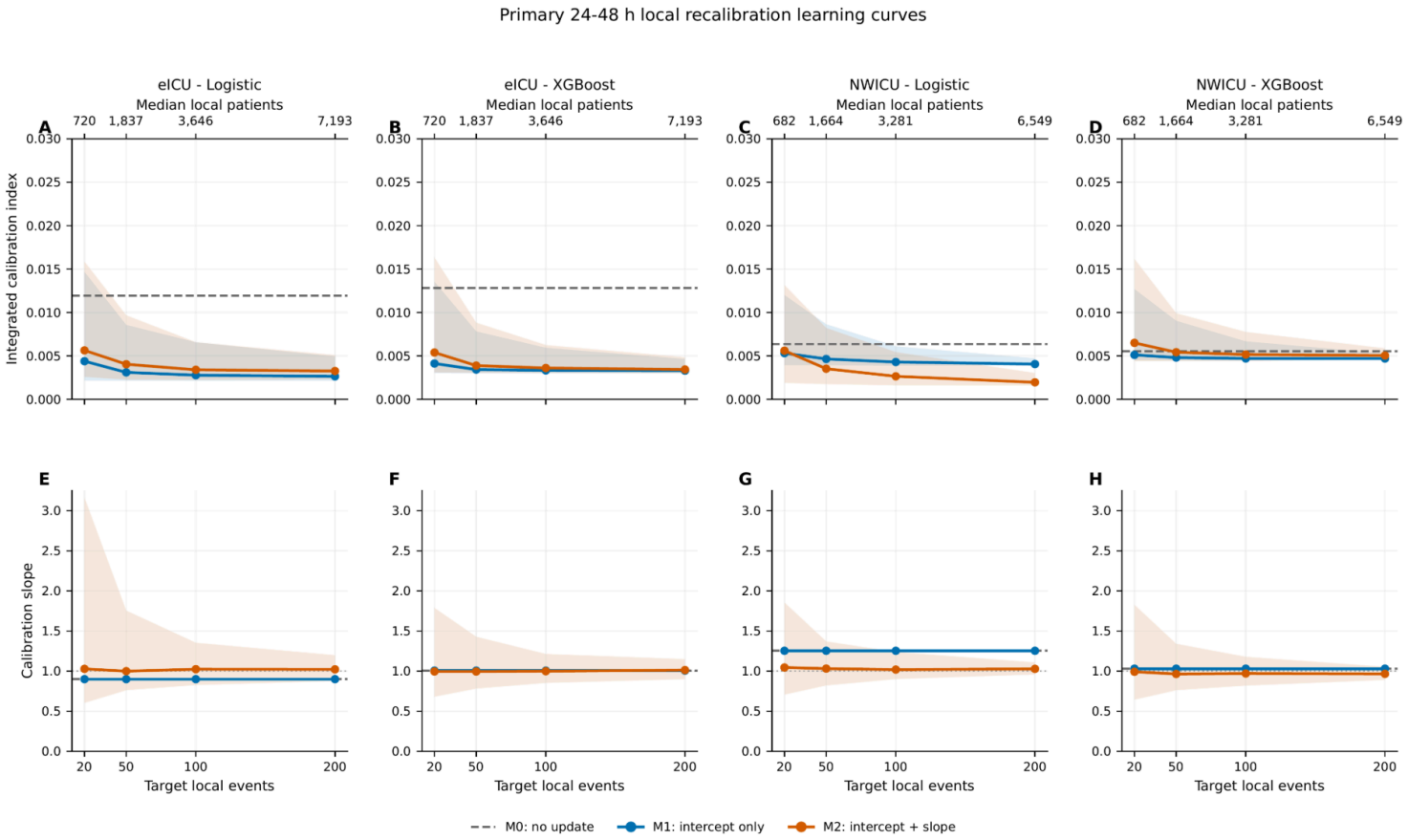


Figure 3

Primary local recalibration learning curves. Primary 24-48 h local recalibration learning curves. Integrated calibration index (Panels A-D) and calibration slope (Panels E-H) are shown across prespecified local event targets for M1 intercept-only and M2 intercept-plus-slope recalibration. Dashed horizontal lines show fixed M0 performance. Shaded regions are empirical 2.5th-97.5th percentiles across 250 repeated adaptation samples and are not population confidence intervals. Top axes show the median number of locally labeled patients required to obtain each event target.

Supplementary Files

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

- Additionalfile1.xlsx
- Additionalfile2.pdf