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.

A 13-Variable Hour-24 Model for 7-Day ICU Mortality in Neurocritical Care: Strong Cross-Database Discrimination with Unit-Level Calibration Differences

Shuaijia Zeng
Affiliated Hospital of North Sichuan Medical College
Xiaodong Zhang
Affiliated Hospital of North Sichuan Medical College
Xing Cao
Affiliated Hospital of North Sichuan Medical College
Quansheng Liu
Affiliated Hospital of North Sichuan Medical College
Xue Chen
Affiliated Hospital of North Sichuan Medical College
Aili Jia
Affiliated Hospital of North Sichuan Medical College
Xing Chen
Affiliated Hospital of North Sichuan Medical College
Jinhui Wen
Affiliated Hospital of Chengdu University
Junhao Chu
chujunhao2013@163.com

Affiliated Hospital of North Sichuan Medical College

Research Article

Keywords: Calibration, Critical Care, Machine Learning, Mortality, Neurocritical Care

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11330992/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

Short-horizon neurocritical risk estimates must remain accurate across care settings. We developed a 13-variable model at hour 24 for intensive care unit (ICU) mortality by day 7 and evaluated its transport to a second database.

Methods

In this retrospective cohort study, adults alive in the ICU at hour 24 were identified in the eICU Collaborative Research Database (eICU-CRD) version 2.0 for development and Medical Information Mart for Intensive Care IV (MIMIC-IV) version 3.1 for cross-database evaluation. Earlier analyses, including MIMIC-IV target data, informed the cohort definitions; the final fitted model was applied unchanged. We harmonized 13 predictors and compared six model families with hospital- and patient-grouped repeated nested validation. Log loss and a one-standard-error rule governed model selection.

Results

eICU-CRD contributed 20,810 admissions, 885 deaths, and 192 hospitals; MIMIC-IV contributed 6,821 admissions and 203 deaths. The selection rule chose cubic B-spline ridge logistic regression. The nested selection procedure yielded an internal area under the receiver operating characteristic curve (AUROC) was 0.8765 (95% confidence interval [CI], 0.8626 to 0.8872), with a Brier score of 0.03407 and calibration slope of 0.9861. In MIMIC-IV, AUROC was 0.9037 (95% CI, 0.8873 to 0.9204), Brier score 0.02456, and calibration slope 0.9750. The pooled observed-to-expected (O/E) ratio was 0.984; By initial care setting, O/E indicated underprediction in the neurologic surgical ICU (Neuro SICU; 1.588) and overprediction in intermediate/stepdown care (0.371).

Conclusions

A transparent 13-variable hour-24 model retained strong discrimination after frozen cross-database transport. Care-setting calibration exposed offsetting probability errors hidden by pooled estimates, supporting unit-specific prospective evaluation of neurocritical early-warning workflows.

Background

After the first ICU day, neurocritical teams decide who needs closer surveillance. Risk estimates are useful only if they rank patients reliably and return credible probabilities in the receiving unit. Neurocritical services vary in organization and case mix.[1–2] Previous studies have evaluated mortality prediction in stroke and mixed ICU populations.[3–4]

External transport often preserves discrimination more readily than calibration; multicenter development does not eliminate dataset shift.[5–7] We compared six model families for a 13-variable hour-24 model of ICU death by day 7, applied the selected pipeline unchanged to MIMIC-IV, and examined calibration within neurologic care settings.

Methods

Design and Data Sources

This retrospective prediction-model study used eICU-CRD v2.0 (United States ICUs, 2014–2015) for multicenter development and MIMIC-IV v3.1 (one United States academic center, 2008–2022) for cross-database transport. Database publications describe content and provenance.[8–9] Release identifiers and access conditions are given under Availability of data and materials. We used TRIPOD + AI and TRIPOD-Cluster for reporting and PROBAST + AI for design review.[10–12] Additional file 1 contains supplementary methods, results, and the fitted equation; Additional file 2 contains the reporting checklist. We finalized protocol version 5.0.0 and modeling extension 5.2.0 before final fitting and scoring. Earlier analyses informed the expanded cohort definitions and the planned care-setting and era analyses. MIMIC-IV data structure also informed shared preprocessing, eligibility, and outcome rules. These rules were set before the final eICU fit, and MIMIC-IV outcomes played no part in fitting, tuning, or model selection in the final development run. Here, frozen transport refers to the unchanged fitted pipeline. Earlier target-data analyses and their implications are described in the Discussion.

Participants, Landmark, and Outcome

The analysis unit was the first ICU stay in each hospital admission; later admissions by the same patient could remain eligible. Adults had to be alive and still in ICU at hour 24. eICU neurologic eligibility required an exact Neuro ICU label or qualifying APACHE admission-diagnosis text or predefined timestamped neurologic evidence by the landmark. MIMIC eligibility required an initial Neuro Surgical ICU (Neuro SICU), Neuro Intermediate, or Neuro Stepdown label. Later transfers and discharge diagnoses could not establish eligibility. Diagnosis composition was summarized descriptively by database. MIMIC-IV diagnostic summaries could incorporate full-admission codes; we did not use these summaries for eligibility, prediction, or model selection.

We set ICU admission as time zero and derived predictors from measurements recorded during hours 0–24. The outcome was ICU death between hours 24 and 168. Alive ICU discharge before hour 168 and survival in ICU through hour 168 were non-events; this fixed-time binary endpoint did not capture deaths after ICU discharge. Outcomes were not imputed. Outcome derivation used objective time-stamped disposition fields and required no subjective adjudication. Treatments were neither eligibility criteria nor predictors; post-landmark care was not modeled.

Predictors and Missing Data

Before modeling, we selected 13 predictors based on clinical relevance and availability in both databases: age, sex, minimum Glasgow Coma Scale, maximum heart rate, minimum mean arterial pressure, maximum respiratory rate, minimum oxygen saturation, last creatinine, sodium range, minimum bicarbonate, maximum glucose, maximum white blood cell count, and minimum platelets. Continuous variables remained continuous. Implausible values were removed and same-time duplicates were median-collapsed. Median imputation and, where applicable, standardization were learned within training partitions; we added neither missingness indicators nor measurement-frequency variables. The Glasgow Coma Scale summarizes eye, verbal, and motor responses.[13] Values came from routine documentation; assessor identities were unavailable. In the spline family, only the 12 continuous predictors were expanded; sex remained one linear binary term.[14–15]

Model Development and Selection

We compared six model families: elastic-net logistic regression, cubic B-spline ridge logistic regression, random forest, LightGBM, XGBoost, and CatBoost. A prevalence-only model and a fixed six-variable clinical logistic model were nonselectable benchmarks. To preserve split independence, hospitals and repeated patients were joined into 125 connected components, each kept within one split.[16–17] Five outer folds were repeated five times, each with three inner folds. Inner log loss selected hyperparameters; one-standard-error rules then chose the first eligible family in the predefined simplicity order. We refitted the selected pipeline on all eICU observations and applied it to MIMIC-IV without refitting or recalibration. We used no synthetic data, outcome resampling, or class weighting. Model complexity can affect prediction stability.[18] We included regression and tree-based candidates because greater algorithmic complexity has not consistently improved clinical prediction.[19]

For sample-size assessment, we specified 73 candidate degrees of freedom, including the largest spline basis, rather than only the 13 named predictors. The available sample exceeded the largest of Riley's three applicable criteria, which address regression rather than the effective complexity of tree ensembles.[20] The MIMIC-IV sample was fixed by database availability and eligibility; its confidence intervals describe achieved transport precision.[21]

Performance, Heterogeneity, and Secondary Analyses

We assessed discrimination with AUROC and average precision (AP); overall accuracy with Brier score and log loss; and calibration with calibration-in-the-large, slope, integrated calibration index (ICI), and observed/expected (O/E) ratio.[22–25] Calibration-in-the-large was the intercept estimated with logit predicted risk as an offset (slope fixed at 1). ICI was the mean absolute difference between predicted risk and LOWESS-smoothed observed risk. Calibration slope came from a separate logistic regression on logit predicted risk with a free intercept; the prognostic model was not refitted. Internal estimates

averaged metrics across complete outer repeats. Uncertainty used 1,000 whole connected-component bootstrap draws in eICU and patient bootstrap draws in MIMIC; all eight primary performance metrics had 1,000 valid replicates in each database. Resamples in which a metric could not be estimated were omitted for that metric, not assigned a value of zero. The internal bootstrap conditioned on the realized repeated out-of-fold predictions rather than rerunning the full nested selection process.

Before the final rerun, initial-care-setting strata compared Neuro SICU with combined Neuro Intermediate and Neuro Stepdown care. Paired patient bootstrap draws retained all admissions for each sampled patient and estimated differences in observed and predicted risk, calibration-in-the-large, slope, and log O/E. Illustrative thresholds of 2%, 5%, 10%, 15%, and 20% were used for retrospective decision curves for additional clinical assessment, not treatment recommendations, in the supplement.[26] Model-specific feature attribution and outer-resample rank stability described the fitted model and were not interpreted causally. Age and care-setting analyses were descriptive. Race and ethnicity could not be harmonized, and small-cell rules limited some sex-stratified outputs.

We included the eICU-provided APACHE IVa predicted ICU-mortality probability post hoc as a contextual benchmark.[8, 27] Its native endpoint differs from day-7 ICU mortality. Supplementary Methods and Table S10 report this analysis; it did not inform model selection.

We also explored patient-grouped target recalibration and an approximate pre-2020 setting-offset update evaluated in 2020-or-later admissions.[28]

Reproducibility

The archive records protocol, code, configuration, input, split, and release hashes. The version-locked source snapshot and computational environments are available from the corresponding author on request. Analyses used Python with scikit-learn 1.9.0, statsmodels 0.14.6, LightGBM 4.6.0, XGBoost 3.2.0, and CatBoost 1.2.10; the archived environment records the remaining dependencies. Additional file 3 gives the clinical extraction rules, finite model grids, selection order, and fold-local preprocessing.

Results

Cohorts and Data Completeness

The final eICU cohort included 20,810 admissions from 19,587 patients and 192 hospitals, with 885 deaths (4.25%). MIMIC-IV included 6,821 admissions from 6,201 patients and 203 deaths (2.98%). Among retained landmark risk-set admissions, every primary outcome was classifiable; Fig. 1 reports the preceding exclusions and sequential cohort flow.

Predictor missingness was highest for minimum Glasgow Coma Scale in eICU (18.18%) and minimum platelet count in MIMIC-IV (6.98%). The largest cross-database shifts involved minimum mean arterial

pressure, minimum oxygen saturation, maximum respiratory rate, and maximum heart rate (absolute standardized mean differences, 0.43–0.57; Table 1).

Table 1
Predictor distributions, missingness, and cross-database standardized differences

Predictor	eICU distribution	Missing	MIMIC-IV distribution	Missing	|SMD|
Age, years	63 [51, 76]	0.0%	66 [54, 78]	0.0%	0.156
Male sex	53.9%	< 10 records	3425/6821 (50.2%)	0.0%	0.073
Minimum Glasgow Coma Scale	13 [8, 15]	18.2%	14 [10, 15]	< 10 records	0.158
Maximum heart rate, beats/min	104 [91, 119]	1.4%	97 [86, 108]	0.0%	0.429
Minimum mean arterial pressure, mmHg	62 [53, 72]	1.5%	70 [62, 78]	0.4%	0.569
Maximum respiratory rate, breaths/min	28 [24, 35]	6.9%	26 [22, 29]	1.2%	0.511
Minimum oxygen saturation, %	91 [86, 94]	2.1%	93 [91, 95]	< 10 records	0.551
Last creatinine, mg/dL	0.85 [0.67, 1.14]	10.8%	0.8 [0.7, 1]	6.5%	0.179
Sodium range, mmol/L	0 [0, 2]	10.5%	0 [0, 2]	6.4%	0.029
Minimum bicarbonate, mmol/L	24 [21, 26]	12.3%	23 [21, 24]	6.6%	0.335
Maximum glucose, mg/dL	140 [113, 182]	6.6%	124 [103, 151]	6.6%	0.328
Maximum white blood cell count, $\times 10^9$/L	10.6 [7.8, 14.1]	13.1%	10.1 [7.6, 13.4]	7.0%	0.108
Minimum platelets, $\times 10^9$/L	196 [152, 246]	13.9%	207 [166, 255]	7.0%	0.143

Values are median [interquartile range] or n/N (%); where complementary small cells were protected, only the aggregate percentage is shown. eICU, eICU Collaborative Research Database; MIMIC-IV, Medical Information Mart for Intensive Care IV; SMD, standardized mean difference; WBC, white blood cell count. Counts from 1 to 9 and complementary values that would reveal them are suppressed; this display rule did not affect fold-local imputation or modeling.

Model Selection and Overall Performance

All six candidate families completed nested evaluation (Table 2). XGBoost had the lowest full-development inner log loss (0.126164; SE, 0.002019), which set a one-standard-error limit of 0.128183.

Cubic B-spline ridge logistic regression also fell within that limit (0.127296) and was the first eligible model in the planned simplicity order.

The spline model's mean outer AUROC was 0.8777, compared with 0.8825 for XGBoost and 0.8826 for CatBoost. For paired descriptive contrasts, we separately averaged each admission's repeated cross-fitted predictions (Supplementary Table S15). Those bootstrap intervals included zero for spline comparisons with the four tree ensembles, but favored spline over elastic-net regression by 0.0155 (95% CI, 0.0099 to 0.0212). The primary estimate below evaluates the complete nested selection procedure.

Repeated nested validation yielded AUROC 0.8765 (0.8626 to 0.8872), AP 0.3091 (0.2711 to 0.3525), Brier score 0.03407 (0.02945 to 0.03921), and calibration slope 0.9861 (0.9355 to 1.0344). Applied unchanged to MIMIC-IV, the model yielded AUROC 0.9037 (0.8873 to 0.9204), AP 0.2949 (0.2346 to 0.3638), Brier score 0.02456 (0.02161 to 0.02764), and slope 0.9750 (0.8913 to 1.0711) (Table 2, Panel B). AUROC depends on case mix; internal and transport estimates were not compared directly.[25]

Supplementary Table S10 reports the APACHE IVa comparison. Additional file 1, Supplementary Table S16 gives the fitted equation, imputation medians, knots, scaling parameters, and coefficients; full-precision model CSV files are available on request.

Table 2
Model selection and primary performance

Panel A. Training-only nested-resampling model-family comparison

Model family	Role	Outer log loss, mean (SD)	Outer AUROC, mean (SD)	Full-development inner log loss, mean (SE)	1-SE eligible	Selected
CatBoost	selection candidate	0.12774 (0.00923)	0.8826 (0.0127)	0.12797 (0.00168)	Yes	No
LightGBM	selection candidate	0.12779 (0.00950)	0.8814 (0.0117)	0.12774 (0.00177)	Yes	No
XGBoost	selection candidate	0.12801 (0.00973)	0.8825 (0.0114)	0.12616 (0.00202)	Yes	No
cubic B-spline ridge logistic regression	selection candidate	0.12954 (0.01022)	0.8777 (0.0146)	0.12730 (0.00171)	Yes	Yes
random forest	selection candidate	0.13047 (0.01072)	0.8751 (0.0167)	0.13075 (0.00212)	No	No
elastic-net logistic regression	selection candidate	0.13344 (0.01084)	0.8651 (0.0154)	0.13301 (0.00234)	No	No

Panel A values are means (SDs) across outer resamples; the SDs summarize resample-to-resample stability rather than inferential uncertainty. Family selection used full-development inner log loss and the one-standard-error rule. Nonselectable benchmarks are reported in the online supplement. AUROC, area under the receiver operating characteristic curve; SD, standard deviation; SE, standard error.
Panel B. Primary internal and MIMIC-IV transport performance

Metric	eICU nested selection procedure (95% CI)	MIMIC-IV cubic B-spline ridge logistic regression transport (95% CI)
N (events)	20810 (885)	6821 (203)
AUROC	0.8765 (0.8626 to 0.8872)	0.9037 (0.8873 to 0.9204)
Average precision (AP)	0.3091 (0.2711 to 0.3525)	0.2949 (0.2346 to 0.3638)
Brier score	0.03407 (0.02945 to 0.03921)	0.02456 (0.02161 to 0.02764)
Log loss	0.12930 (0.11460 to 0.14704)	0.09489 (0.08469 to 0.10570)
Calibration-in-the-large	−0.0031 (− 0.1506 to 0.1387)	−0.0200 (− 0.1807 to 0.1278)
Calibration slope	0.9861 (0.9355 to 1.0344)	0.9750 (0.8913 to 1.0711)
ICI	0.00488 (0.00374 to 0.00856)	0.00591 (0.00390 to 0.00953)
O/E ratio	0.9976 (0.8882 to 1.1128)	0.9842 (0.8637 to 1.1050)

Internal estimates are means of metrics calculated separately in each complete outer repeat of the nested selection procedure. MIMIC-IV estimates evaluate one model fitted to the full eICU development cohort and applied without refitting. All intervals are percentile bootstrap 95% confidence intervals; all eight interval metrics had 1,000 valid bootstrap replicates in each database. eICU, eICU Collaborative Research Database; AP, average precision; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; ICI, integrated calibration index; MIMIC-IV, Medical Information Mart for Intensive Care IV; O/E, observed-to-expected.

Calibration Across Care Settings and Secondary Analyses

Setting-specific performance is reported in Table 3 and Fig. 2. Neuro SICU contained 165/1,338 deaths (12.33%), compared with a mean predicted risk of 7.77%; its O/E ratio was 1.5876 (1.3836 to 1.7924). The combined intermediate/stepdown setting contained 38/5,483 deaths (0.69%), compared with 1.87% predicted; its O/E ratio was 0.3714 (0.2608 to 0.4915). AUROCs were 0.8175 (0.7849 to 0.8484) and 0.8845 (0.8528 to 0.9128), respectively. In the patient-cluster bootstrap, the observed-risk difference (Neuro SICU minus intermediate/stepdown) was 11.64 percentage points (95% CI, 9.90 to 13.51); the corresponding log O/E contrast was 1.453 (95% CI, 1.130 to 1.829). The pooled O/E interval included 1,

whereas both stratum-specific intervals excluded 1 on opposite sides (Fig. 2B). In intermediate/stepdown care, the absolute overprediction was 1.17 percentage points. Model-specific explanation and rank stability are reported in Supplementary Table S9.

Table 3
Calibration across initial care settings in the cross-database evaluation

MIMIC-IV population	N (events)	Observed	Predicted	AUROC (95% CI)	Slope (95% CI)	O/E (95% CI)
All neurologic units	6821 (203)	2.98%	3.02%	0.9037 (0.8873 to 0.9204)	0.9750 (0.8913 to 1.0711)	0.9842 (0.8637 to 1.1050)
Neuro SICU	1338 (165)	12.33%	7.77%	0.8175 (0.7849 to 0.8484)	0.7828 (0.6737 to 0.9069)	1.5876 (1.3836 to 1.7924)
Neuro Intermediate/ Stepdown	5483 (38)	0.69%	1.87%	0.8845 (0.8528 to 0.9128)	0.7947 (0.6750 to 0.9282)	0.3714 (0.2608 to 0.4915)

Pooled AUROC includes separation between care settings and is therefore not directly comparable with stratum-specific AUROCs. Calibration estimates in intermediate/stepdown care are based on 38 deaths and do not establish thresholds for clinical action. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; MIMIC-IV, Medical Information Mart for Intensive Care IV; Neuro SICU, neurologic surgical intensive care unit (database label: Neuro Surgical ICU); O/E, observed-to-expected. Strata use initial unit labels and do not establish the location of care at hour 24.

At a 10% transport threshold, sensitivity was 53.7%, specificity 93.7%, positive predictive value 20.6%, and negative predictive value 98.5%. At the fixed 5% threshold, retrospective net benefit was 0.0153 relative to an assessment-none strategy. Decision curves for all thresholds are shown in the supplement.

Approximate era stratification, excluding 780 admission intervals that spanned 2020, yielded AUROC 0.8964 for intervals ending by 2019 and 0.9136 for intervals starting in 2020 or later; O/E ratios were 1.112 and 0.768, respectively. The descriptive era contrast could not separate temporal from care-setting change. In MIMIC-IV, AUROC was 0.9098 (95% CI, 0.8861 to 0.9312) in women and 0.8979 (95% CI, 0.8706 to 0.9211) in men. These descriptive estimates were not a formal fairness assessment.

Discussion

Applied unchanged to MIMIC-IV, the model retained strong discrimination (AUROC, 0.9037). Predicted and observed mortality were nearly identical overall (pooled O/E, 0.984), concealing underestimation in Neuro SICU and overestimation in intermediate/stepdown care. Both O/E intervals excluded 1, and the paired bootstrap contrast in log O/E excluded 0. Calibration slopes below 1 in both settings suggest that an update should address their spread as well as baseline risk.

These errors could affect which patients receive additional assessment. At a fixed threshold, underprediction in Neuro SICU may suppress reviews among high-risk patients, whereas overprediction in intermediate/stepdown care may add low-yield alerts. Retrospective net benefit at the 5% threshold was 0.0153. A silent prospective run can measure alert volume and yield before clinicians see the predictions.

Care-setting labels may capture differences in severity, triage, treatment, and monitoring absent from the 13 harmonized predictors. In Neuro SICU, the lower within-unit AUROC and underprediction may partly reflect differences in case mix and unmeasured perioperative factors, including decompressive craniectomy or hematoma evacuation before hour 24. Intermediate/stepdown care represents a lower-acuity population with different monitoring practices. Discrimination and calibration also differed across harmonized ICU datasets in a previous multicenter study.[6] The opposite error directions identify care setting as a target for local updating. All 13 predictors are routinely recorded; the model uses no imaging-derived features and can run in an open-source software environment. These strata describe initial unit assignment, not necessarily the patient's location at hour 24. In the exploratory later-era analysis, setting-offset updating increased AUROC but did not improve all calibration measures (Supplementary Table S11).

The one-standard-error rule favored spline regression under the planned trade-off between fit and transparency. XGBoost and CatBoost had slightly higher mean outer AUROCs, but the paired comparisons in Supplementary Table S15 were inconclusive. APACHE IVa provided context for a different endpoint. Comparing disease-specific scores across this mixed cohort would also require compatible populations, inputs, and prediction horizons.

The development cohort met all three applicable Riley criteria for 73 candidate degrees of freedom: 20,810 admissions and 885 events were available, compared with a maximum requirement of 12,772 admissions (approximately 543 events). Ridge penalization controlled model complexity, and repeated grouped nested validation estimated performance after tuning and family selection. Supplementary Table S5 reports the conventional parameter-count stress test. Internal confidence intervals condition on the realized repeated out-of-fold predictions and do not capture variation from rerunning the complete selection procedure.

Earlier results informed the expanded cohorts and heterogeneity analyses. MIMIC-IV outcomes were not used for tuning or model selection in the final development run. This prior knowledge of target-data performance and structure may have favored transport estimates and distinguishes the study from an untouched external confirmation. The databases also differ in institution, case mix, measurement, and calendar time. Because the 2020 cutoff coincided with the COVID-19 pandemic, the era comparison cannot isolate temporal drift. Both datasets come from the United States; admission thresholds, staffing, monitoring, and treatment may differ in other health systems and should be represented in a local calibration sample.

Only 38 deaths occurred in intermediate/stepdown care, limiting the precision of its calibration estimates.[21] Its relative overprediction was large; the absolute difference was 1.17 percentage points. The effect on alert burden remains unknown. Minimum GCS was missing in 18.2% of eICU admissions; recording may depend on sedation, intubation, or monitoring. Median imputation without missingness indicators may bias risk estimates, but we did not establish the missingness mechanism or its contribution to miscalibration. Alive ICU discharge before hour 168 counted as a non-event. This ICU-specific endpoint depends on discharge practice and excludes post-ICU deaths, including deaths before day 7. The 24-hour landmark excludes patients who died or left earlier, and ICU mortality does not describe neurologic function. Race and ethnicity could not be harmonized, leaving performance across these demographic groups unassessed. We did not estimate hospital-specific calibration random effects. Decision curves and updating analyses remain retrospective. Measurements were selected by chart timestamps; delayed-entry availability at the landmark was not assessed. The APACHE admission-diagnosis eligibility field had no entry-time cutoff. A prospective implementation must verify that both eligibility information and predictor values are available at hour 24.

We propose a local evaluation in two stages: first collect silent predictions and estimate calibration intercepts and slopes; then lock any update and test it in a later cohort before clinicians see the output. [28–30]

Conclusions

A transparent 13-variable hour-24 model retained strong discrimination after frozen cross-database transport. Care-setting calibration exposed offsetting probability errors hidden by pooled estimates, supporting unit-specific prospective evaluation of neurocritical early-warning workflows.

Abbreviations

AP
average precision
APACHE
Acute Physiology and Chronic Health Evaluation
AUROC
area under the receiver operating characteristic curve
CI
confidence interval
eICU-CRD
eICU Collaborative Research Database
GCS
Glasgow Coma Scale
ICI
integrated calibration index

ICU
intensive care unit
LOWESS
locally weighted scatterplot smoothing
MIMIC-IV
Medical Information Mart for Intensive Care IV
Neuro SICU
neurologic surgical intensive care unit (database label: Neuro Surgical ICU)
O/E
observed-to-expected
PROBAST + AI
Prediction model Risk Of Bias ASsessment Tool for regression and artificial intelligence models
SD
standard deviation
SE
standard error
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence.

Declarations

Ethics approval and consent to participate

We analyzed deidentified eICU-CRD and MIMIC-IV records under PhysioNet data-use agreements. The eICU release meets US Health Insurance Portability and Accountability Act safe-harbor standards. The Beth Israel Deaconess Medical Center institutional review board approved MIMIC-IV data sharing and waived informed consent. This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and applicable guidelines for research using deidentified data. We did not contact patients, intervene in their care, or access direct identifiers.

Consent for publication

Not applicable. The manuscript contains aggregate results and no identifiable individual information.

Availability of data and materials

eICU-CRD v2.0 and MIMIC-IV v3.1 are available to credentialed PhysioNet users under their respective data-use agreements.[31–33] The dataset records are https://doi.org/10.13026/C2WM1R and https://doi.org/10.13026/kpb9-mt58. The authors cannot redistribute patient-level source data.

Additional file 1 contains the reported aggregate results and fitted equation. A version-locked archive of small-cell-protected aggregate results and machine-readable figure source data is available from the corresponding author on reasonable request.

Code availability

The full-precision model specification and version-locked replication code, including computational environments, tests, and a checksum manifest, are available from the corresponding author on reasonable request. The fitted equation is reported in Supplementary Table S16.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Scientific Research Start-Up Fund for High-Level Talents of the Affiliated Hospital of North Sichuan Medical College (grant number 2023GC004). The funder had no role in study design, data collection, analysis or interpretation, manuscript preparation, or the decision to submit.

Authors' contributions

SZ: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing - original draft. XZ: Data curation, Formal analysis, Validation, Writing - review and editing. XCa: Methodology, Validation, Visualization, Writing - review and editing. QL: Data curation, Investigation, Validation, Writing - review and editing. XC: Data curation, Methodology, Validation, Writing - review and editing. AJ: Investigation, Resources, Validation, Writing - review and editing. XCh: Investigation, Software, Validation, Writing - review and editing. JW: Methodology, Resources, Validation, Writing - review and editing. JC: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review and editing. Initials XCa and XCh distinguish Xing Cao and Xing Chen, respectively. All authors read and approved the final manuscript.

Patient and public involvement

Patients and the public were not involved in the design, conduct, reporting, interpretation, or dissemination planning of this secondary analysis of deidentified data.

Study registration

This retrospective prediction-model study was not prospectively registered; the versioned protocol and modeling extension are available to editors and reviewers on request.

Acknowledgements

The authors thank PhysioNet, the MIMIC-IV and eICU-CRD investigators, contributing institutions, and the patients whose deidentified data made this research possible.

Declaration of generative AI and AI-assisted technologies

The authors did not use generative AI or AI-assisted technologies in the design, conduct, analysis, interpretation, drafting, language editing, or preparation of this manuscript. All work was performed by the authors, who take full responsibility for the content and integrity of the study.

References

1. Pham X, Ray J, Neto AS, Laing J, Perucca P, Kwan P, et al. Association of neurocritical care services with mortality and functional outcomes for adults with brain injury: a systematic review and meta-analysis. JAMA Neurol. 2022;79:1049–58. 10.1001/jamaneurol.2022.2456.
2. Powla PP, Fakhri F, Jankowski S, Mansour A, Polley EC. Clinical prediction models in neurocritical care: an overview of the literature and example application to prediction of hospital mortality in traumatic brain injury. Neurocrit Care. 2025;42:32–8. 10.1007/s12028-024-02083-2.
3. Huang J, Chen H, Deng J, Liu X, Shu T, Yin C, et al. Interpretable machine learning for predicting 28-day all-cause in-hospital mortality for hypertensive ischemic or hemorrhagic stroke patients in the ICU: a multi-center retrospective cohort study with internal and external cross-validation. Front Neurol. 2023;14:1185447. 10.3389/fneur.2023.1185447.
4. Lim L, Gim U, Cho K, Yoo D, Ryu HG, Lee HC. Real-time machine learning model to predict short-term mortality in critically ill patients: development and international validation. Crit Care. 2024;28:76. 10.1186/s13054-024-04866-7.
5. Cox EGM, Wiersema R, Eck RJ, Kaufmann T, Granholm A, Vaara ST, et al. External validation of mortality prediction models for critical illness reveals preserved discrimination but poor calibration. Crit Care Med. 2023;51:80–90. 10.1097/CCM.0000000000005712.
6. Rockenschaub P, Hilbert A, Kossen T, Elbers P, von Dincklage F, Madai VI, et al. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. Crit Care Med. 2024;52:1710–21. 10.1097/CCM.0000000000006359.
7. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385:283–6. 10.1056/NEJMc2104626.

8. 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.
9. 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.
10. 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.
11. Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ. 2023;380:e071018. 10.1136/bmj-2022-071018.
12. 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.
13. Teasdale G, Maas A, Lecky F, Manley G, Stocchetti N, Murray G. The Glasgow Coma Scale at 40 years: standing the test of time. Lancet Neurol. 2014;13:844–54. 10.1016/S1474-4422(14)70120-6.
14. 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.
15. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Med Res Methodol. 2019;19:46. 10.1186/s12874-019-0666-3.
16. de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. Stat Med. 2021;40:3533–59. 10.1002/sim.8981.
17. Takada T, Nijman S, Denaxas S, Snell KIE, Uijl A, Nguyen TL, et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. J Clin Epidemiol. 2021;137:83–91. 10.1016/j.jclinepi.2021.03.025.
18. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J. 2023;65:e2200302. 10.1002/bimj.202200302.
19. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. 10.1016/j.jclinepi.2019.02.004.
20. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, 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.
21. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, Van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ. 2024;384:e074821. 10.1136/bmj-2023-074821.
22. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. on behalf of the 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.

23. 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.
24. 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.
25. 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.
26. Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. 10.1186/s41512-019-0064-7.
27. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34:1297–310. 10.1097/01.CCM.0000215112.84523.F0.
28. 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.
29. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med. 2023;21:70. 10.1186/s12916-023-02779-w.
30. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res. 2022;6:24. 10.1186/s41512-022-00136-8.
31. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. https://doi.org/10.13026/C2WM1R
32. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet. 2024. https://doi.org/10.13026/kpb9-mt58
33. Pollard T, Moody BE, Lehman LH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1:792–5. 10.1038/s44360-026-00096-z.

Figures

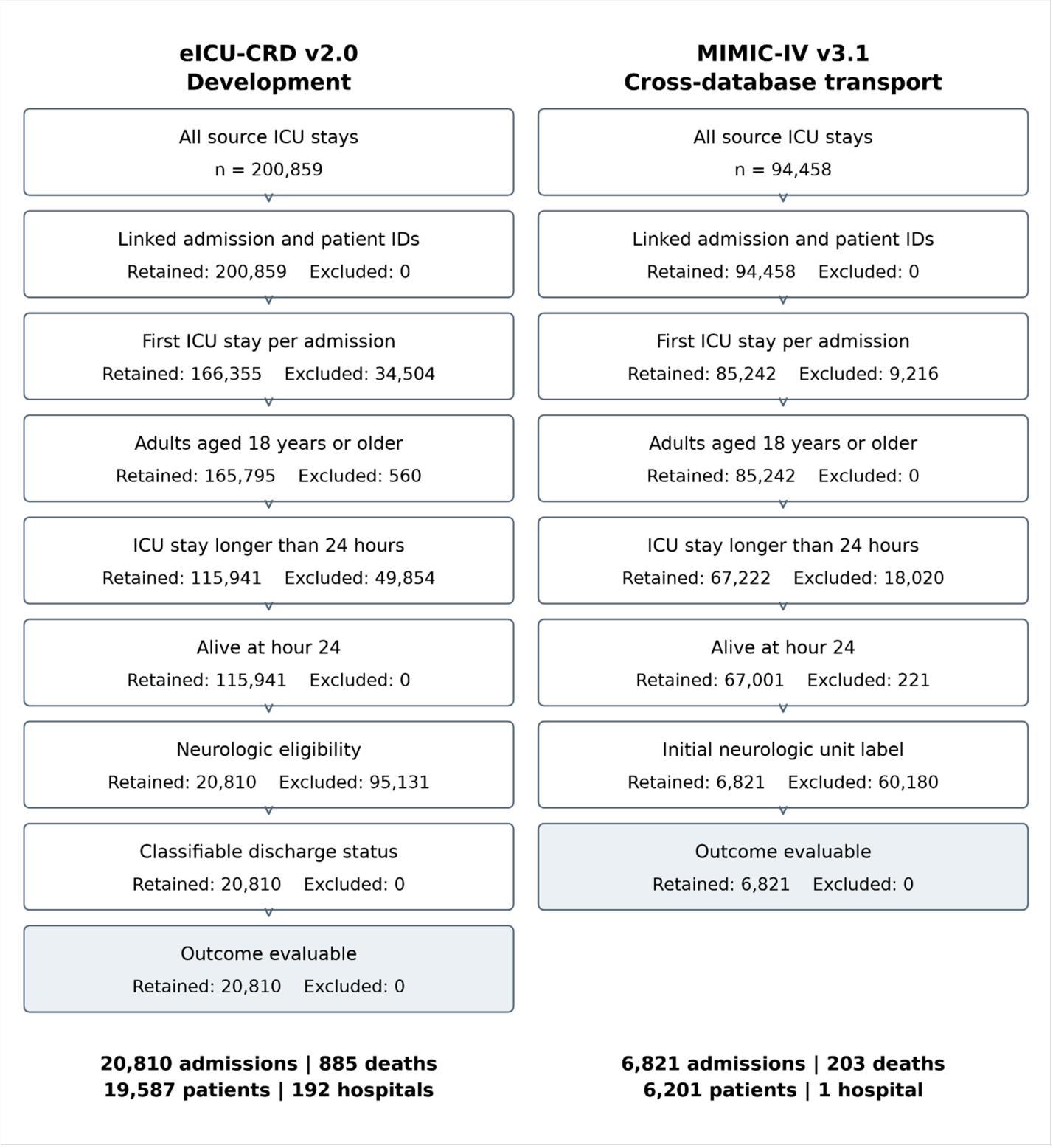


Figure 1

Cohort construction and analytic framework. Sequential eligibility and exclusion counts are shown for eICU-CRD development and MIMIC-IV transport. The flow diagram shows the hour-24 eligibility sequence; the outcome was ICU death after hour 24 and by hour 168. Abbreviations: eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; MIMIC-IV, Medical Information Mart for Intensive Care IV.

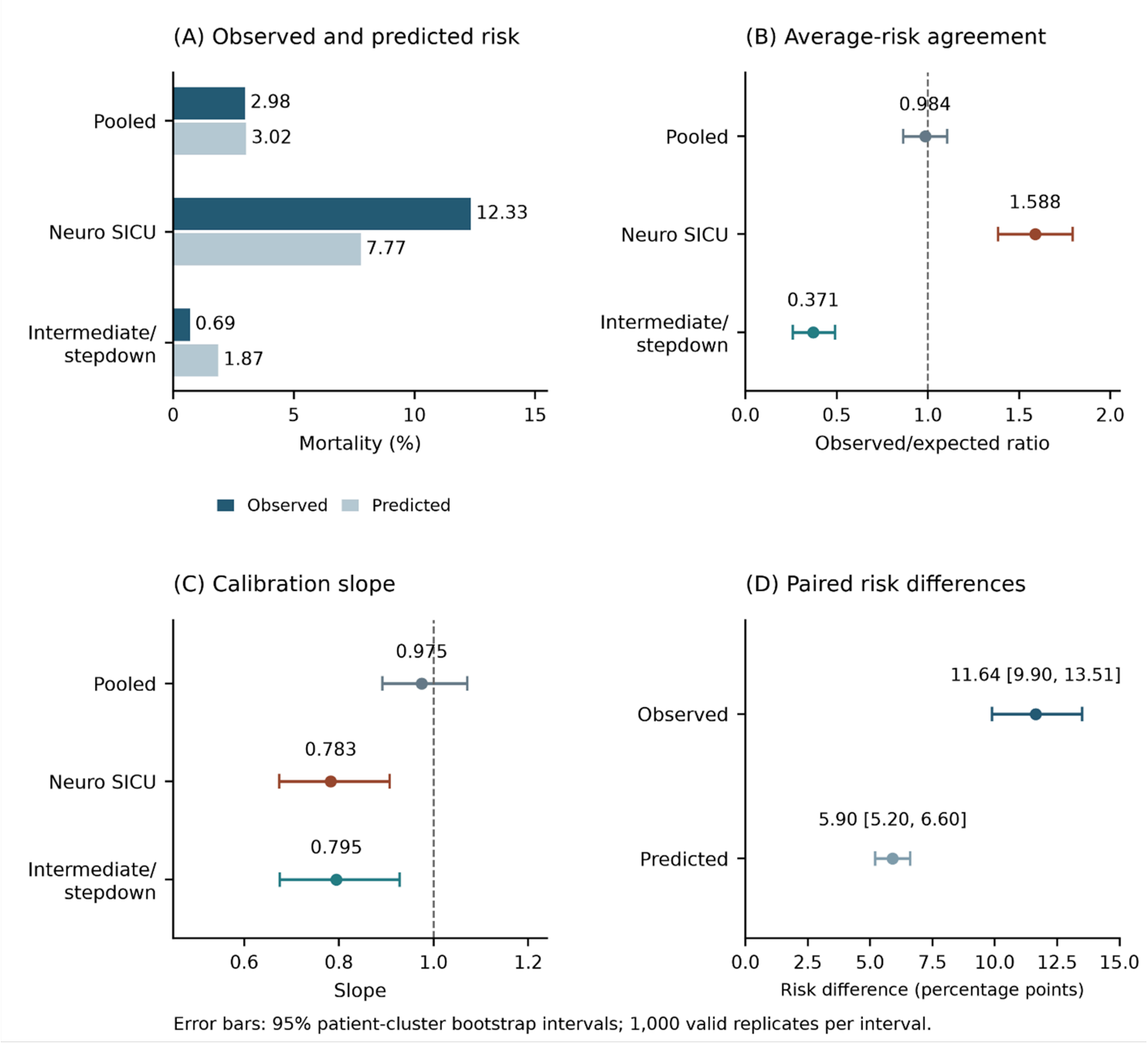


Figure 2

Calibration across initial care settings after MIMIC-IV transport. (A) Observed and mean predicted day-7 ICU mortality (point estimates; marginal CIs unavailable). (B) O/E ratios. (C) Calibration slopes. Error bars in (B) and (C) are 95% patient-cluster bootstrap CIs; each interval uses 1,000 valid replicates. For O/E, 1 denotes average-risk agreement; values above/below 1 indicate underprediction/overprediction. (D) Paired patient-cluster bootstrap differences, Neuro SICU minus intermediate/stepdown care. Abbreviations: CI, confidence interval; ICU, intensive care unit; MIMIC-IV, Medical Information Mart for Intensive Care IV; N, admissions; Neuro SICU, neurologic surgical intensive care unit (database label: Neuro Surgical ICU); O/E, observed-to-expected; pp, percentage points.

Supplementary Files

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

- Additionalfile1.docx
- Additionalfile2TRIPOD.docx
- Additionalfile3Methods.docx