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.

Temporal representation of vital signs for mortality prediction after major abdominal surgery: external validation and sampling- resolution analysis

Qiuxia Huang
Fujian Provincial Hospital
Long Huang
Fujian Provincial Hospital
Sheng Lin[2]
Fujian Provincial Hospital
Fengfeng Lu
fengwind@fjmu.edu.cn

Fujian Provincial Hospital

Research Article

Keywords: electronic health records, vital signs, temporal aggregation, mortality prediction, external validation, medical informatics

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11383335/v1

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

Additional Declarations: No competing interests reported.

Temporal representation of vital signs for mortality prediction after major abdominal surgery: external validation and sampling-resolution analysis

Qiuxia Huang[1], Long Huang[2], Sheng Lin[2], Fengfeng Lu[2]*

[1] Department of Echocardiography, Fuzhou University Affiliated Provincial Hospital, Fujian Provincial Hospital, Shengli Clinical Medical College of Fujian Medical University, Fuzhou, China.

[2] Department of Respiratory and Critical Care Medicine, Fuzhou University Affiliated Provincial Hospital, Fujian Provincial Hospital, Shengli Clinical Medical College of Fujian Medical University, Fuzhou, China.

*Correspondence: Fengfeng Lu; fengwind@fjmu.edu.cn
ORCID: 0009-0004-6137-9795

Author emails: Qiuxia Huang, Syj158@fzu.edu.cn; Long Huang, huanglong1988@fjmu.edu.cn; Sheng Lin, sy0920@fzu.edu.cn.

Article type: Research article
Main text: 3147 words | Abstract: 288 words
5 figures | 3 tables | 35 references

Abstract

Background: Temporal representation determines which features of routinely recorded vital signs are available for analysis. We evaluated changes in physiological information and mortality prediction associated with different representations after major abdominal surgery.

Methods: This retrospective study included adults alive and still in the intensive care unit (ICU) 12 hours after admission. Models using first-hour snapshots, 12-hour summaries, or post hoc same-window medians of heart rate, mean arterial pressure, and oxygen saturation were developed in MIMIC-IV and evaluated in the Salzburg Intensive Care database (SICdb) without refitting. A separate within-SICdb analysis compared 5-, 15-, and 60-minute median-based representations in identical densely monitored patients using common cross-validation folds. Processed one-minute records provided the physiological reference. Death was assessed after the landmark through day 28 from ICU admission. Performance differences used patient-paired bootstrap confidence intervals.

Results: MIMIC-IV included 1740 patients (168 deaths), and SICdb included 2376 (185 deaths). Among 2061 densely monitored SICdb patients (151 deaths), hourly aggregation changed positive to negative abnormality classification in 392/1388 reference-positive patients for hypotension (28.2%), 237/835 for tachycardia (28.4%), and 468/671 for desaturation (69.7%). Paired 5-minute minus 60-minute differences in area under the receiver operating characteristic curve (AUROC) were 0.007 (95% confidence interval −0.001 to 0.016) for regression and 0.019 (−0.001 to 0.041) for XGBoost. In external evaluation, full-summary regression achieved

AUROC 0.779, with differences of 0.040 (0.019–0.063) versus snapshots and 0.013 (−0.004 to 0.032) versus same-window medians. External calibration-in-the-large was 0.243 and calibration slope was 1.293.

Conclusions: Hourly aggregation changed recorded abnormality classifications. AUROC gains from finer representations were not clearly established. Twelve-hour summaries discriminated mortality better than admission snapshots, but their increment over same-window medians was uncertain externally. These findings describe retrospective data representations and do not establish a preferred bedside monitoring frequency.

Keywords: electronic health records; vital signs; temporal aggregation; mortality prediction; external validation; medical informatics

Background

Clinical prediction models reuse monitoring data that have already undergone choices about sampling, aggregation, and feature construction. The same vital-sign record can be represented by a single hourly value, a summary of an observation window, or a sequence. Recent studies have evaluated engineered vital-sign representations and features of the electronic health record (EHR) measurement process [1, 2]. In a dynamic ICU-acquired infection model, a score derived from high-frequency vital signs added limited predictive performance beyond conventional features [3]. These findings motivate evaluation of representations for a specified task rather than assuming that additional detail will improve prediction.

Physiological information retention is a related but different consideration. Prospective postoperative monitoring has identified hypoxemia not apparent in intermittent observations [4], and analyses of high-resolution ICU records have characterized oxygen-saturation exposure and cumulative hypotension [5, 6]. Aggregation can alter these descriptions by smoothing brief excursions and extremes. Whether the same changes materially affect prediction of a later outcome is an empirical question; an abnormality-retention measure is not a measure of predictive performance.

Comparisons also depend on the observation window. A first-hour snapshot and a 12-hour summary differ in both the time observed and the features supplied to a model. A comparator using median values from the same 12-hour window helps assess the additional contribution of fuller descriptors, although it does not isolate individual features. Evaluation in another database is relevant because patient mix and recording practices may differ [7].

Postoperative intensive care after major abdominal surgery offers a defined setting for these comparisons. Early observations follow a major physiological stress, and postoperative complications contribute to subsequent mortality [8–10]. Broader postoperative prediction models have included clinical and operative information [11]; other ICU models have used repeated observations or trajectories [12–14]. The processed minute-level records in SICdb permit a within-patient comparison of temporal aggregation [15].

We examined temporal representation in two related analyses: MIMIC-IV model development with external evaluation in SICdb, and a separate within-SICdb comparison of 5-, 15-, and 60-minute representations using the same patients and validation folds. The latter assessed recorded physiological abnormalities and 28-day mortality prediction. A post hoc same-window median comparator contextualized the external prediction results.

Methods

Study design, data sources, and ethics

We conducted a retrospective prognostic-modeling study using the Medical Information Mart for Intensive Care IV (MIMIC-IV), version 3.1, and the Salzburg Intensive Care database (SICdb), version 1.0.8 [16–18]. MIMIC-IV contains records from Beth Israel Deaconess Medical Center, Boston, during 2008–2022; SICdb includes four ICUs at University Hospital Salzburg during 2013–2021 [15, 16]. MIMIC-IV provided the development cohort. SICdb provided an independent external evaluation cohort and minute-level recordings for a separate sampling-resolution analysis. Both databases are distributed through PhysioNet [19]. Reporting was guided by STROBE, RECORD, and TRIPOD+AI [20–22].

Data were accessed through PhysioNet under the applicable credentialing requirements and data-use agreements. One author (Fengfeng Lu) completed the Collaborative Institutional Training Initiative (CITI) Program “Data or Specimens Only Research” refresher course (Record ID: 68221589). MIMIC-IV data sharing was approved by the institutional review board of Beth Israel Deaconess Medical Center [16]. SICdb was approved by the Ethics Commission of the State of Salzburg (EK 1115/2021) [18]. Informed consent was waived for the source databases. No additional ethics review was required for this de-identified secondary analysis under the requirements of the authors’ institution.

Study population and outcome

We included adults aged at least 18 years admitted to ICU after qualifying upper gastrointestinal, intestinal/colorectal, or hepatobiliary/pancreatic surgery and retained the first eligible ICU stay per patient. In MIMIC-IV, mapped procedure codes were combined with postoperative care-unit evidence or procedure-date and surgical-service alignment. In SICdb, surgical-site and admission-type classifications were aligned to the ICU entry offset. Open and laparoscopic operations could qualify. Diagnostic endoscopy, drainage and catheter-based procedures, minor bedside interventions, transplantation, and vascular procedures were excluded. Database-specific definitions and code lists are provided in Additional file 1, Table S1, and the extraction code.

ICU admission was time zero. Predictor data were collected from admission to, but not including, 12 hours, and patients had to remain alive and in ICU at the 12-hour landmark. The outcome was all-cause death after the landmark through day 28 from ICU admission. MIMIC-IV ascertainment used hospital death timestamps and linked death dates; one record with an ambiguous date-only death at a classification boundary was excluded. SICdb death offsets, derived from sources that include governmental mortality records and out-of-hospital deaths [18], were evaluated relative to the ICU entry offset. Records without a death in the specified window were coded as non-events; source-specific rules are detailed in Additional file 1. All eligible records were included; no prospective sample-size calculation was performed. Figure 1 shows cohort selection and the observation and outcome windows.

Data extraction and predictor definitions

Data were extracted and harmonized using database-specific source mappings, common units, and source-priority rules. Model inputs were age, sex, heart rate, mean arterial pressure (MAP), and peripheral oxygen saturation (SpO_2). The clinical, laboratory, and organ-support variables in Table 1 described the cohorts but were not additional predictors. Heart-rate values outside 20–250 beats/min, MAP outside 20–200 mmHg, and SpO_2 outside 50–100% were excluded. Source mappings and signal availability are reported in Tables S1–S3.

The common representation consisted of within-hour medians without carrying observations across hours. The admission snapshot used the first hourly bin, age, sex, and missingness indicators. Full 12-hour summaries included each signal's median, extreme, standard deviation, linear slope, abnormal-value burden, valid-hour fraction, and longest missing streak. Extremes were the heart-rate maximum and MAP/SpO_2 minima. Burden was the proportion of valid bins with heart rate >100 beats/min, MAP <65 mmHg, or SpO_2 <92%. The full-

summary matrix had 23 base features and their 23 missingness indicators. The same-window median comparator used age, sex, the three 12-hour medians, and five corresponding missingness indicators. Imputation medians and scaling parameters were estimated within training partitions. Table S2 provides complete definitions.

Sampling-resolution and sensitivity analyses

The SICdb dense-core cohort required each signal to have at least 45 observed minutes in at least 8 of the first 12 hours. Nonoverlapping 5-, 15-, and 60-minute median bins required at least 3, 8, and 30 valid minutes, respectively. The three temporal-model families were fitted within SICdb using identical patients and repeated cross-validation folds, with model complexity fixed across resolutions. Slopes, longest missing streaks, and recurrent-model elapsed-time inputs used hour-based units at every resolution. The principal contrast was 5 minutes minus 60 minutes. This experiment was separate from external validation of MIMIC-IV models.

Processed one-minute records served as the measurement reference. Reference-positive patients had at least five consecutive observed abnormal minutes; missing minutes interrupted a run. Aggregated records were positive if at least one valid median bin crossed the same threshold. Positive-status loss was the percentage of reference-positive patients classified as negative after aggregation. This operational comparison incorporates the aggregation and observation-support rules; it is not episode-level detection sensitivity. Additional measures were absolute burden error, standard-deviation retention, and attenuation of extremes. A separate sparse-recording analysis used one observation per hour at fixed minutes 00, 15, 30, 45, or 59. A fixed-setting unweighted recurrent-model sensitivity analysis examined the consequences of class weighting for predicted probabilities [23].

Model development and internal validation

We evaluated L2-penalized logistic regression for admission snapshots, elastic-net-family logistic regression and XGBoost for full summaries, and a compact gated recurrent unit with decay (GRU-D) for hourly sequences [24–26]. GRU-D incorporated values, observation masks, and time since observation, with age and sex added at the output layer. The level-only comparator used the same penalized-regression family and candidate grid as the full-summary model.

Internal validation used five repeats of stratified five-fold cross-validation with shared patient-level folds. Imputation, scaling, and tuning were confined to outer training partitions. Tabular models used inner three-fold

cross-validation; GRU-D used a validation subset within the training partition for early stopping [27]. Selection prioritized log loss, followed by Brier score and lower complexity. Out-of-fold probabilities were averaged across repeats. Final settings were selected from development fits before fitting the complete MIMIC-IV cohort. The full-summary regression selected the L1 boundary (mixing parameter 1.0), whereas the level-only comparator selected 0.5. GRU-D had 1241 trainable parameters and used positive-class-weighted binary cross-entropy (Table S4).

Direct summary-versus-snapshot contrasts and the unweighted GRU-D sensitivity analysis were added after review of the initial four-model comparison. The same-window level-only comparator was added during manuscript preparation, after the earlier external results had been examined. Its fitting and tuning used MIMIC-IV only. Supplementary Methods document the analysis sequence.

External validation and performance assessment

Final MIMIC-IV models and preprocessing were applied to SICdb without refitting. Discrimination was assessed using the area under the receiver operating characteristic curve (AUROC) and average precision, the implemented area under the precision–recall curve (AUPRC); probability accuracy was assessed using Brier score and log loss [28, 29]. Calibration-in-the-large used the prediction logit as an offset, and calibration slope was estimated with a free intercept [30]. Intercept-only and intercept/slope recalibration were fitted and assessed in SICdb as diagnostic analyses. Exploratory decision curves evaluated net benefit across hypothetical risk thresholds of 0.02–0.25; no clinical action rule was tested [31].

Statistical analysis and software

Continuous characteristics are reported as medians and interquartile ranges and categorical characteristics as counts and percentages. Absolute standardized mean differences (SMDs) describe cross-cohort differences; they were not reported for structurally different surgical classifications. For predictive performance and paired contrasts, 95% confidence intervals (CIs) were calculated from 2000 patient-level bootstrap replicates, using shared resamples for paired comparisons. Models were not refitted during bootstrapping; intervals are conditional on fitted predictions and are pointwise rather than adjusted for multiple comparisons. Sampling-resolution comparisons were not designed as equivalence tests.

Analyses used Python 3.12.14; figures were prepared in R 4.5.2 and Python. ChatGPT (OpenAI) assisted study planning, literature organization, manuscript preparation, language editing, and an internal pre-submission

critique. OpenAI Codex assisted data-extraction and statistical code development. The authors reviewed the outputs and retain responsibility for the work. Package versions, seeds, and implementation details are provided in Additional file 1 and the code archive.

Results

Study population and cohort characteristics

MIMIC-IV included 1740 patients, of whom 168 died after the landmark and by day 28 from ICU admission (9.7%). SICdb included 2376 patients with 185 deaths (7.8%). The SICdb dense-core subgroup included 2061 patients with 151 deaths (7.3%) (Figure 1).

Median age was 66 years in MIMIC-IV and 70 years in SICdb; 55.9% and 57.2% of patients were male (Table 1). Intestinal/colorectal procedures accounted for 47.3% and 42.0%, and upper gastrointestinal procedures for 26.1% and 40.7%, respectively. MIMIC-IV identified multiple qualifying surgical families in 9.4%; SICdb did not have a directly corresponding category. Invasive ventilation was recorded in 52.0% versus 29.3%, and vasopressor use in 35.8% versus 55.3%. The median of patients' 12-hour heart-rate summaries was 91.5 versus 79.5 beats/min, and the corresponding SpO_2 values were 98.0% versus 96.6%; median MAP was 74 mmHg in both cohorts. Source definitions and availability are detailed in Tables S1B and S1C.

Physiological information altered by aggregation

After hourly aggregation, 392 of 1388 reference-positive patients for hypotension (28.2%), 237 of 835 for tachycardia (28.4%), and 468 of 671 for desaturation (69.7%) no longer met the aggregated-record positivity criteria (Figure 2). Median standard-deviation retention was 76.5% for MAP, 84.4% for heart rate, and 74.4% for SpO_2. Aggregation attenuated MAP minima upward by a median 17.5 mmHg, heart-rate maxima downward by 16 beats/min, and SpO_2 minima upward by 6.45 percentage points.

Sampling resolution and mortality prediction

Within the same SICdb dense-core patients, full-summary regression had AUROCs of 0.752, 0.752, and 0.745 at 5-, 15-, and 60-minute resolution (Figure 3; Table 2). Paired 5-minute minus 60-minute differences were 0.007 (95% CI −0.001 to 0.016) for regression, 0.019 (−0.001 to 0.041) for XGBoost, and 0.002 (−0.020 to 0.024) for weighted GRU-D. All three AUROC-difference CIs included zero. Log-loss differences were −0.00183 (−0.00362 to −0.00016) for regression and −0.00472 (−0.00911 to −0.00047) for XGBoost. The

XGBoost Brier-score difference was −0.00098 (−0.00190 to −0.00007). These probability-score changes were small in absolute terms; all metrics and model comparisons are reported in Table 2.

External evaluation of observation-window representations

Internally, full-summary penalized regression achieved an AUROC of 0.763 (95% CI 0.723–0.801), compared with 0.676 (0.634–0.717) for the admission snapshot (Figure 4; Table 3). The paired increase was 0.087 (0.054–0.120). Average precision increased from 0.194 to 0.271, a difference of 0.077 (0.027–0.134). Brier score decreased from 0.084 to 0.080 and log loss from 0.299 to 0.279 (Table S5).

XGBoost had an internal AUROC of 0.776 (0.738–0.813), with a paired increase of 0.100 (0.063–0.137) over the snapshot. Weighted GRU-D had an AUROC of 0.739 (0.699–0.781) and a Brier score of 0.176, compared with 0.079 for XGBoost and 0.080 for full-summary regression.

In SICdb, full-summary regression achieved an AUROC of 0.779 (95% CI 0.741–0.812), versus 0.738 (0.698–0.774) for the snapshot. The paired increase was 0.040 (0.019–0.063). Average precision increased from 0.236 to 0.281, a difference of 0.045 (0.008–0.088), and Brier score and log loss were lower (Table 3; Table S5). External AUROCs were 0.750 (0.709–0.787) for XGBoost and 0.721 (0.680–0.759) for weighted GRU-D. The XGBoost–snapshot difference was 0.012 (−0.014 to 0.038).

The 12-hour level-only comparator had AUROCs of 0.713 (0.669–0.754) internally and 0.765 (0.728–0.802) externally. The full-summary minus level-only differences were 0.050 (0.022–0.078) internally and 0.013 (−0.004 to 0.032) externally (Table S11). External CIs for differences in average precision, Brier score, and log loss also included zero.

For full-summary regression in SICdb, calibration-in-the-large was 0.243 (0.095–0.390) and calibration slope was 1.293 (1.083–1.520), indicating overall underprediction and a narrower predicted-risk spread than observed (Figure 5). Recalibration diagnostics and exploratory decision curves are provided in Table S6 and Figure S1/Table S9.

Recording-process sensitivity and model interpretation

Across five fixed sparse-hourly phases, full-summary regression had external AUROCs of 0.755–0.761, compared with 0.779 for hourly medians (Table S7). Recurrent-model phase results are shown in Figure S2. The fixed-setting unweighted GRU-D sensitivity had an external AUROC of 0.764 (95% CI 0.725–0.800), Brier score of 0.066, calibration-in-the-large of 0.039, and calibration slope of 1.503 (Figure S3; Table S8).

Age had the largest absolute standardized coefficient in the full-summary regression. SpO_2 valid-hour fraction, tachycardia burden, hypotension burden, and the heart-rate extreme were also prominent (Figure S4). These coefficients describe contributions within the fitted model.

Discussion

In this postoperative population, hourly aggregation changed abnormality classifications, variability, and extremes. In the same dense-core patients, the estimated AUROC differences between 5- and 60-minute representations were small, with confidence intervals including zero. Separately, MIMIC-trained 12-hour summaries showed higher external discrimination than admission snapshots, while their increment over same-window medians was uncertain. Together, these comparisons describe the recorded physiological features and mortality-prediction performance of the selected representations.

Physiological information capture and prediction

The observed classification changes arose from applying different representations to the same underlying records, rather than from comparing different patient groups. The percentage losing reference-positive status ranged from 28.2% for hypotension to 69.7% for desaturation. These values depend on the threshold, persistence definition, and observation-support requirements. The one-minute reference was processed data, and reference positivity required five consecutive abnormal minutes, whereas aggregated positivity required an abnormal bin median. The comparison therefore does not measure clinically adjudicated events missed by routine care.

The physiological metrics and prediction metrics describe different properties and are not on comparable scales. We did not link a particular lost excursion to an individual prediction error or establish why discrimination changed little. Broad physiological levels may retain information for later mortality even when brief excursions are smoothed, but that explanation remains a possibility rather than a demonstrated mechanism.

Resolution under a common learning setup

The resolution analysis held patient membership, validation folds, and model complexity constant while fitting models within SICdb at each resolution. It estimates performance under common development-derived settings, rather than the best performance attainable after separate optimization at each resolution.

Some log-loss and Brier-score intervals excluded zero, but the absolute differences were small and arose within several model–metric comparisons. They do not demonstrate clinically important superiority of finer resolution. Conversely, AUROC intervals containing zero do not demonstrate equivalence. The study evaluated a later mortality outcome, not immediate deterioration, alarm detection, or treatment response.

Observation window and external evaluation

The larger external difference against the admission snapshot became smaller when the comparator summarized the same 12-hour window. The snapshot comparison changes both the observation window and the feature set. The median comparator was added after examination of earlier results, although its fitting and tuning remained confined to MIMIC-IV; this comparison should be interpreted as exploratory.

Studies of engineered representations and high-frequency infection prediction have reported differing incremental gains [1, 3]. The present findings concern three vital signs, a 12-hour window, and a postoperative mortality task; they should not be interpreted as a general comparison of simple and complex models. In particular, the weighted GRU-D benchmark had poor probability accuracy, and the unweighted sensitivity produced different results. This illustrates the importance of the training objective in interpreting that benchmark [23], rather than establishing inferiority of recurrent models.

Implications for reuse of monitoring data

Recording completeness was itself part of the full-summary representation. This is consistent with work examining measurement-process features in EHR prediction [2], but our analysis did not separate physiological descriptors from recording-process features by ablation. Coefficients for valid-hour fraction or missingness therefore cannot be interpreted as physiological effects.

The sparse-phase analysis further distinguishes hourly aggregation from one observation per hour. An hourly median uses the observations available within that hour; it is not a simulation of reducing bedside acquisition to hourly measurements. Reporting binning rules, valid-data requirements, and source priorities alongside variables and units helps define the evaluated input. External calibration also matters: full-summary regression underpredicted mortality in SICdb despite its higher AUROC than the snapshot [30].

Limitations

This retrospective study involved two hospital systems and one postoperative population. Patients who died or left ICU before 12 hours were excluded, so the results do not apply to all surgical admissions. Coding and

ascertainment differed between databases. Dense-core selection limits extension to sparsely recorded encounters; the four Salzburg ICUs do not constitute four independent hospital validations. The minute-level reference was not a waveform-level gold standard, and pulse oximetry has additional measurement limitations [32].

Development included 168 deaths, limiting flexible models [33]. Bootstrap intervals were conditional on fitted predictions and excluded full model-development uncertainty. Several comparisons were added after initial results review. Snapshot and summary regressions used different penalty families, so their contrast changes both the information window and modeling choices. No comprehensive clinical score or last-observation-at-landmark comparator was included. The fixed-complexity resolution comparison and weighted recurrent benchmark do not support broad algorithm rankings. Operational costs, alarm burden, and effects on clinical decisions or patient outcomes require different evaluations [34].

Conclusions

Hourly aggregation altered recorded physiological abnormalities in postoperative ICU patients. Under the evaluated learning setup, AUROC gains from finer representations were not clearly established. Twelve-hour summaries discriminated mortality better than admission snapshots in external evaluation, but their added value over same-window medians remained uncertain. These findings characterize the evaluated data representations; they do not establish an optimal sampling interval or a benefit from changing clinical monitoring.

List of abbreviations

AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; EHR, electronic health record; GRU-D, gated recurrent unit with decay; ICU, intensive care unit; MAP, mean arterial pressure; MIMIC-IV, Medical Information Mart for Intensive Care IV; SICdb, Salzburg Intensive Care database; SMD, standardized mean difference; SpO_2, peripheral oxygen saturation.

Declarations

Ethics approval and consent to participate

This study used de-identified MIMIC-IV and SICdb data under their respective credentialing and data-use agreements. MIMIC-IV data sharing was approved by the institutional review board of Beth Israel Deaconess Medical Center, with a waiver of informed consent [16]. SICdb was approved by the Ethics Commission of the

State of Salzburg (EK 1115/2021), with a waiver of individual consent [18]. Under the requirements of the authors' institution, no additional institutional ethics review was required for this secondary analysis.

Consent for publication

Not applicable. No identifiable individual participant information is presented.

Availability of data and materials

MIMIC-IV version 3.1 and SICdb version 1.0.8 are available through PhysioNet under credentialed-access conditions and their respective data-use agreements [17, 18]. Researchers must obtain access directly from the data custodians. Aggregate numerical results and fixed-grid curve-display data are provided in Additional file 2. Patient-level source or derived records, individual predictions, and fitted model objects are not redistributed.

Code availability

Analysis code for the postoperative vital-sign summaries and sampling-resolution study is available in the public GitHub repository as release v1.0.1 (commit be5a802) [35]. The Python and R source is distributed under the MIT license; environment records, required packages, a synthetic test, and Windows execution instructions accompany the release. Full reproduction requires authorized access to both databases and configuration of the documented local workflow. Patient-level data, individual predictions, and fitted model objects are excluded from the public release.

Protocol, registration, and patient/public involvement

The study was not prospectively registered. Supplementary Methods describe the original analyses and subsequent additions; the versioned public code release documents their implementation. Patients and members of the public were not involved in the design, conduct, reporting, or dissemination planning.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Joint Funds for the Innovation of Science and Technology, Fujian Province, China (Grant Nos. 2024Y9069 and 2025Y9010). The funders had no role in study design, data extraction or processing, statistical analysis, interpretation, manuscript preparation, or the decision to submit the manuscript for publication.

Authors' contributions

QH contributed to conceptualization, methodology, data curation, formal analysis, visualization, and drafting. LH contributed to methodology, validation, interpretation, and critical revision. SL contributed to data curation, validation, visualization, and critical revision. FL contributed to conceptualization, methodology, supervision, project administration, funding acquisition, interpretation, and critical revision. All authors read and approved the final manuscript and agree to be accountable for the work.

Acknowledgements

The authors thank the investigators, clinical teams, data contributors, and maintainers of MIMIC-IV and SICdb for making these resources available to qualified researchers.

Additional files

Additional file 1: Additional_file_1_Supplementary_Material.docx. Supplementary methods, Tables S1–S11, and Figures S1–S4. Cohort and variable definitions, missingness, model specifications, analysis chronology, and supporting comparisons.

Additional file 2: Additional_file_2_Aggregate_Sources.zip. Frozen aggregate numerical sources for tables and figures, with a crosswalk from original labels to the current manuscript. Patient-level records, individual predictions, and fitted model objects are not included.

Author-prepared STROBE/RECORD and TRIPOD+AI reporting checklists are supplied as reporting documentation.

References

1. Mamandipoor B, Shen I, Hsu CN, Gabriel RA. Engineering biomarker representations of vital signs data enhances deep learning mortality prediction. J Am Med Inform Assoc. 2026;33:1381–1386. https://doi.org/10.1093/jamia/ocag066

2. Zou J, Wang H, Xin Z, Zhou Q, Chen J, Huang J. From predictive signal to deployable value: EHR measurement-process features in ICU AKI prediction. Int J Med Inform. 2026;221:106668. https://doi.org/10.1016/j.ijmedinf.2026.106668

3. Varkila MRJ, Lancia G, van Smeden M, Bonten MJM, Spitoni C, Cremer OL. Early detection of ICU-acquired infections using high-frequency electronic health record data. BMC Med Inform Decis Mak. 2025;25:273. https://doi.org/10.1186/s12911-025-03031-6

4. Sun Z, Sessler DI, Dalton JE, Devereaux PJ, Shahinyan A, Naylor AJ, et al. Postoperative hypoxemia is common and persistent: a prospective blinded observational study. Anesth Analg. 2015;121:709–715. https://doi.org/10.1213/ANE.0000000000000836

5. Ding XY, Xu HP, Zhang JR, Chen H. Temporal dimensions of oxygen saturation exposure and clinical outcomes in critically ill patients: a retrospective multi-database cohort study. Crit Care. 2025;29:402. https://doi.org/10.1186/s13054-025-05652-9

6. Ding XY, Xu HP, Zhang JR, Chen H. Development and validation of the Hypotensive Exposure Duration Index for mortality risk prediction in critically ill patients. J Intensive Care. 2025;13:65. https://doi.org/10.1186/s40560-025-00834-7

7. Allyn J, Oliver M, Cerveau R, Allou N, Barennes T, Ferdynus C. Multi-source data integration through pooling and transfer learning improves generalizability and specialization of deep learning models for ICU mortality and length of stay prediction: a four-database external validation study. Crit Care. 2026;30:364. https://doi.org/10.1186/s13054-026-06034-5

8. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, et al. Mortality after surgery in Europe: a 7 day cohort study. Lancet. 2012;380:1059–1065. https://doi.org/10.1016/S0140-6736(12)61148-9

9. International Surgical Outcomes Study group. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. Br J Anaesth. 2016;117:601–609. https://doi.org/10.1093/bja/aew316

10. Spence J, LeManach Y, Chan MTV, Wang CY, Xavier D, Pearse RM, et al. Association between complications and death within 30 days after noncardiac surgery. CMAJ. 2019;191:E830–E837. https://doi.org/10.1503/cmaj.190221

11. Park DJ, Baik SM, Hong KS, Yi H, Lee JG, Lee JM. Development and external validation of an artificial intelligence model for predicting mortality and prolonged ICU stay in postoperative critically ill patients: a retrospective study. World J Emerg Surg. 2025;20:79. https://doi.org/10.1186/s13017-025-00650-2

12. Oh W, Veshtaj M, Sawant A, Agrawal P, Gomez H, Suarez-Farinas M, et al. ORAKLE: Optimal Risk prediction for mAke30 in patients with sepsis associated AKI using deep LEarning. Crit Care. 2025;29:212. https://doi.org/10.1186/s13054-025-05457-w

13. Liu Z, Li L, Zhang Y, Xue S, Zhang L, Qian Z, et al. Stratifying mortality risk in hypotension using outcome-oriented clustering of vital sign trajectories. iScience. 2026;29:114821. https://doi.org/10.1016/j.isci.2026.114821

14. Shen Y, Zhang W, Lin K, Zheng P, Liu X, Weng J, et al. Blood pressure response index trajectories identify distinct hemodynamic phenotypes and predict mortality in septic shock: a two-database retrospective cohort study. Crit Care. 2026;30:243. https://doi.org/10.1186/s13054-026-06059-w

15. Rodemund N, Wernly B, Jung C, Cozowicz C, Koköfer A. Harnessing Big Data in Critical Care: Exploring a new European Dataset. Sci Data. 2024;11:320. https://doi.org/10.1038/s41597-024-03164-9

16. 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. https://doi.org/10.1038/s41597-022-01899-x

17. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi LA, Mark R. MIMIC-IV (version 3.1) [dataset]. PhysioNet. 2024. https://doi.org/10.13026/kpb9-mt58

18. Rodemund N, Kokoefer A, Wernly B, Cozowicz C. Salzburg Intensive Care database (SICdb), a freely accessible intensive care database (version 1.0.8) [dataset]. PhysioNet. 2024. https://doi.org/10.13026/8m72-6j83

19. 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;1:792–795. https://doi.org/10.1038/s44360-026-00096-z

20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4:e296. https://doi.org/10.1371/journal.pmed.0040296

21. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12:e1001885. https://doi.org/10.1371/journal.pmed.1001885

22. 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. https://doi.org/10.1136/bmj-2023-078378

23. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022;29:1525–1534. https://doi.org/10.1093/jamia/ocac093

24. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

25. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–794. https://doi.org/10.1145/2939672.2939785

26. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8:6085. https://doi.org/10.1038/s41598-018-24271-9

27. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91

28. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

29. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432

30. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7

31. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565–574. https://doi.org/10.1177/0272989X06295361

32. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med. 2020;383:2477–2478. https://doi.org/10.1056/NEJMc2029240

33. 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. https://doi.org/10.1136/bmj.m441

34. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28:924–933. https://doi.org/10.1038/s41591-022-01772-9

35. Huang Q, Huang L, Lin S, Lu F. Postoperative vital-sign summaries and sampling resolution: analysis code. Version v1.0.1 [software]. GitHub; 2026. https://github.com/fengwind11/postop-vitals-resolution/releases/tag/v1.0.1. Accessed 14 Oct 2026.

Figure legends

Figure 1. Cohort construction and the 12-hour landmark

(a) MIMIC-IV and SICdb cohort selection using database-specific surgical definitions. One MIMIC-IV record was excluded after the 12-hour landmark because date-only mortality information did not permit unambiguous classification. (b) The predictor window extends from ICU admission to, but not including, 12 hours. Death is assessed after that landmark through day 28 from ICU admission. Patients must be alive and remain in ICU at the landmark. The timeline is schematic and not to scale.

Figure 2. Physiological information altered by hourly aggregation

SICdb dense-core patients (N = 2061), with decoded one-minute records as the reference. Reference positivity requires at least five consecutive observed abnormal minutes (heart rate >100 beats/min, MAP <65 mmHg, or SpO_2 <92%); aggregated positivity requires at least one abnormal valid hourly median. Positive-status loss is the percentage of reference-positive patients no longer positive after aggregation. Standard-deviation retention is the hourly/reference ratio multiplied by 100. Median extreme attenuation is expressed in beats/min for heart rate, mmHg for MAP, and percentage points for SpO_2. Burden error is the 90th percentile absolute difference in abnormal-bin proportions, in percentage points (pp). This is a patient-level classification comparison under the stated rules, not an episode-level missed-event rate.

Figure 3. Mortality prediction across sampling resolutions in the same patients

(a) AUROC and average precision for 5-, 15-, and 60-minute representations. (b) Paired 5-minute minus 60-minute differences. All resolutions use the same 2061 SICdb dense-core patients and 151 deaths. Models were fitted within SICdb using common repeated cross-validation folds and fixed complexity settings. Error bars are 95% percentile intervals from 2000 patient-level bootstrap replicates, conditional on fitted predictions. Positive

differences favor 5-minute resolution for both displayed discrimination metrics. Table 2 reports probability-accuracy measures. This is a within-SICdb comparison, not external validation of high-frequency MIMIC-trained models.

Figure 4. Mortality discrimination with admission and early-course vital-sign representations

Receiver operating characteristic (ROC) and precision–recall curves for MIMIC-IV repeated out-of-fold evaluation and SICdb external evaluation. The snapshot uses the first hourly bin; the median-level, full-summary, and recurrent models use the first 12 hours. All include age and sex. AUROC, average precision, and confidence intervals are reported in Table 3; paired contrasts are in Tables S5 and S11. Diagonal reference lines apply only to ROC panels. Horizontal precision–recall reference lines show each cohort's event proportion. Curves are displayed on a fixed grid; metrics were computed from the original predictions.

Figure 5. Calibration of locked models in the external cohort

Observed mortality versus mean predicted risk in decile-based groups for the snapshot, full-summary regression, and XGBoost models in SICdb (N = 2376; 185 deaths), without refitting or recalibration. Point size indicates group size and the dashed line indicates perfect calibration. Calibration-in-the-large and slope are reported in Table 3. The median-level comparator is also reported in Table S11, and weighted/unweighted GRU-D calibration in Figure S3.

Table 1. Clinical characteristics of the development and external validation cohorts

Characteristic	MIMIC-IV N = 1740	SICdb N = 2376	Absolute SMD
Outcome			
28-day mortality after the 12-h landmark, n (%)	168 (9.7)	185 (7.8)	—
Demographics			
Age, years	66.0 [55.0, 76.0]	70.0 [55.0, 80.0]	0.12
Male sex, n (%)	972 (55.9)	1358 (57.2)	0.03
Surgical category			
Upper gastrointestinal, n (%)	455 (26.1)	968 (40.7)	—
Intestinal/colorectal, n (%)	823 (47.3)	998 (42.0)	—
Hepatobiliary/pancreatic, n (%)	299 (17.2)	410 (17.3)	—
Multiple qualifying families, n (%)	163 (9.4)	Not separately coded	—
Recorded comorbidities			
Hypertension, n (%)	1004 (57.7)	1081 (45.5)	0.25

Characteristic	MIMIC-IV N = 1740	SICdb N = 2376	Absolute SMD
Diabetes mellitus, n (%)	427 (24.5)	401 (16.9)	0.19
Chronic lung disease, n (%)	392 (22.5)	274 (11.5)	0.30
Chronic renal disease, n (%)	242 (13.9)	292 (12.3)	0.05
Early postoperative laboratory values			
Creatinine, mg/dL	0.90 [0.70, 1.30]	0.90 [0.70, 1.14]	0.16
Hemoglobin, g/dL	10.7 [9.4, 11.9]	10.9 [9.4, 12.5]	0.10
White blood cell count, 10^9/L	11.2 [7.7, 15.9]	11.0 [7.7, 14.7]	0.13
Platelet count, 10^9/L	212 [152, 288]	231 [177, 296]	0.13
Organ support during the first 12 h			
Invasive mechanical ventilation, n (%)	905 (52.0)	695 (29.3)	0.48
Any vasopressor, n (%)	623 (35.8)	1313 (55.3)	0.40
Core physiology during the first 12 h			
Heart rate, beats/min	91.5 [79.0, 104.0]	79.5 [69.5, 90.5]	0.65
MAP, mmHg	74.0 [68.0, 82.0]	74.0 [68.0, 83.0]	0.06
SpO_2, %	98.0 [96.5, 100.0]	96.6 [95.4, 97.8]	0.59

Values are n (%) or median [interquartile range]. Deaths occurred after the 12-hour landmark through day 28 from ICU admission. Laboratory values are the first valid measurements during the initial 12 hours; physiological levels are medians of valid hourly bins. Available-case denominators are in Table S1C. Surgical classification differed structurally: SICdb did not separately identify multiple qualifying families, and surgical-category SMDs are not reported. Comorbidities were derived from index-hospital diagnoses in MIMIC-IV and processed preconditions in SICdb; these and organ-support comparisons are descriptive. Additional clinical descriptors were not model predictors. SMD, standardized mean difference; MAP, mean arterial pressure; SpO_2, peripheral oxygen saturation.

Table 2. Sampling-resolution performance and paired differences in SICdb

Model	Resolution (minutes)	AUROC	Average precision	Brier score	Log loss
A. Within-SICdb repeated cross-validation (N = 2061; 151 deaths at each resolution)					
Full 12-h summaries (penalized logistic)	5	0.752 (0.711–0.793)	0.246 (0.186–0.318)	0.062 (0.054–0.071)	0.231 (0.205–0.257)
Full summaries (XGBoost)	5	0.761 (0.719–0.801)	0.231 (0.179–0.299)	0.062 (0.054–0.071)	0.230 (0.204–0.256)
Sequence (weighted GRU-D)	5	0.750 (0.712–0.788)	0.252 (0.188–0.318)	0.182 (0.175–0.189)	0.538 (0.521–0.556)
Full 12-h summaries (penalized logistic)	15	0.752 (0.712–0.792)	0.247 (0.189–0.317)	0.062 (0.053–0.071)	0.231 (0.204–0.257)
Full summaries (XGBoost)	15	0.752 (0.709–0.794)	0.220 (0.174–0.287)	0.063 (0.054–0.072)	0.232 (0.205–0.260)
Sequence (weighted GRU-D)	15	0.752 (0.713–0.790)	0.250 (0.188–0.322)	0.179 (0.171–0.187)	0.529 (0.511–0.547)
Full 12-h summaries (penalized logistic)	60	0.745 (0.703–0.785)	0.237 (0.182–0.314)	0.062 (0.054–0.071)	0.232 (0.206–0.258)
Full summaries (XGBoost)	60	0.743 (0.697–0.783)	0.209 (0.164–0.276)	0.063 (0.055–0.072)	0.235 (0.209–0.262)

Model	Resolution (minutes)	AUROC	Average precision	Brier score	Log loss
Sequence (weighted GRU-D)	60	0.748 (0.705–0.788)	0.227 (0.172–0.295)	0.181 (0.173–0.189)	0.535 (0.516–0.554)
B. Paired 5-minute minus 60-minute differences					
Full 12-h summaries (penalized logistic)	—	0.007 (−0.001 to 0.016)	0.009 (−0.005 to 0.023)	−0.00038 (−0.00084 to 0.00005)	−0.00183 (−0.00362 to −0.00016)
Full summaries (XGBoost)	—	0.019 (−0.001 to 0.041)	0.021 (−0.005 to 0.050)	−0.00098 (−0.00190 to −0.00007)	−0.00472 (−0.00911 to −0.00047)
Sequence (weighted GRU-D)	—	0.002 (−0.020 to 0.024)	0.025 (−0.010 to 0.054)	0.00090 (−0.00341 to 0.00503)	0.00363 (−0.00651 to 0.01320)

Models were fitted within SICdb using identical patients and repeated cross-validation folds, with complexity fixed across resolutions. Positive discrimination differences and negative Brier/log-loss differences favor 5-minute sampling. Paired 95% CIs used 2000 patient-level bootstrap replicates. Additional decimal places are retained for small probability-score differences. These are within-SICdb comparisons, not MIMIC-trained external predictions or equivalence tests.

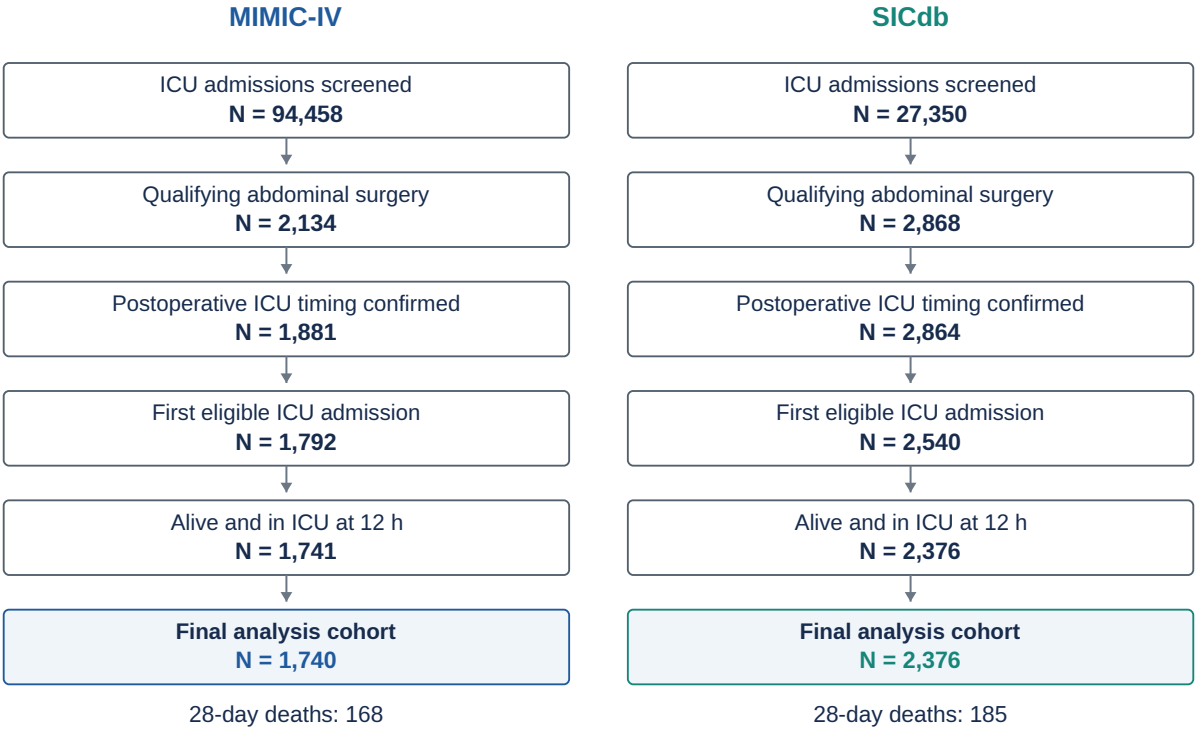
Table 3. Mortality prediction in the development and external cohorts

Model / representation	AUROC	Average precision	Brier score	Log loss	Calibration-in-the-large	Calibration slope
A. MIMIC-IV repeated out-of-fold evaluation (N = 1740; 168 deaths)						
Admission snapshot (L2 logistic)	0.676 (0.634–0.717)	0.194 (0.155–0.251)	0.084 (0.073–0.095)	0.299 (0.269–0.331)	0.006 (−0.166–0.169)	0.937 (0.711–1.186)
12-h median levels (penalized logistic)	0.713 (0.669–0.754)	0.237 (0.186–0.300)	0.082 (0.071–0.092)	0.291 (0.260–0.321)	0.001 (−0.175–0.158)	1.076 (0.838–1.320)
Full 12-h summaries (penalized logistic)	0.763 (0.723–0.801)	0.271 (0.221–0.345)	0.080 (0.070–0.090)	0.279 (0.250–0.310)	−0.011 (−0.185–0.152)	1.028 (0.831–1.255)
Full summaries (XGBoost)	0.776 (0.738–0.813)	0.270 (0.221–0.337)	0.079 (0.069–0.089)	0.275 (0.246–0.305)	0.027 (−0.144–0.192)	1.112 (0.931–1.311)
Hourly sequence (weighted GRU-D)	0.739 (0.699–0.781)	0.263 (0.211–0.334)	0.176 (0.168–0.185)	0.533 (0.513–0.553)	−2.059 (−2.235 to −1.891)	0.890 (0.726–1.071)
B. SICdb locked external evaluation (N = 2376; 185 deaths)						
Admission snapshot (L2 logistic)	0.738 (0.698–0.774)	0.236 (0.184–0.297)	0.067 (0.059–0.076)	0.248 (0.225–0.273)	0.000 (−0.144–0.144)	1.463 (1.185–1.753)
12-h median levels (penalized logistic)	0.765 (0.728–0.802)	0.242 (0.193–0.308)	0.066 (0.058–0.075)	0.241 (0.217–0.267)	0.003 (−0.165–0.152)	1.353 (1.128–1.604)
Full 12-h summaries (penalized logistic)	0.779 (0.741–0.812)	0.281 (0.226–0.351)	0.065 (0.058–0.074)	0.238 (0.214–0.265)	0.243 (0.095–0.390)	1.293 (1.083–1.520)
Full summaries (XGBoost)	0.750 (0.709–0.787)	0.231 (0.185–0.295)	0.067 (0.059–0.075)	0.245 (0.221–0.272)	0.184 (0.035–0.332)	1.184 (1.004–1.379)
Hourly sequence (weighted GRU-D)	0.721 (0.680–0.759)	0.224 (0.175–0.285)	0.090 (0.083–0.097)	0.314 (0.296–0.334)	−1.229 (−1.402 to −1.056)	0.659 (0.547–0.772)

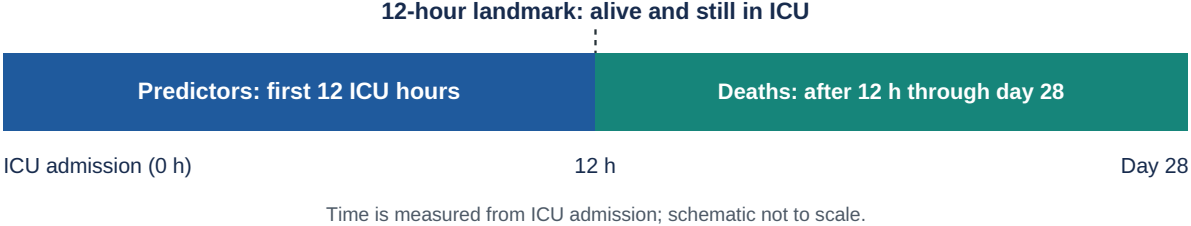
Values are estimates (95% CIs) conditional on fitted predictions. Average precision is the implemented AUPRC. Calibration-in-the-large uses an offset-only model; slope uses an estimated intercept. Penalized regression was tuned within the elastic-net family; the final full-summary and median-level fits selected mixing parameters 1.0 and 0.5, respectively. External evaluation used no SICdb refitting.

The original weighted GRU-D benchmark is shown; the unweighted sensitivity appears in Table S8. Tables S5 and S11 report paired differences.

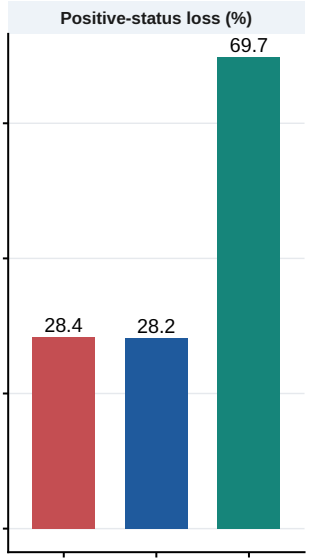
a Cohort construction

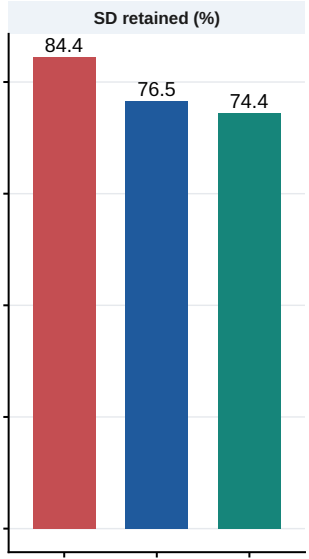


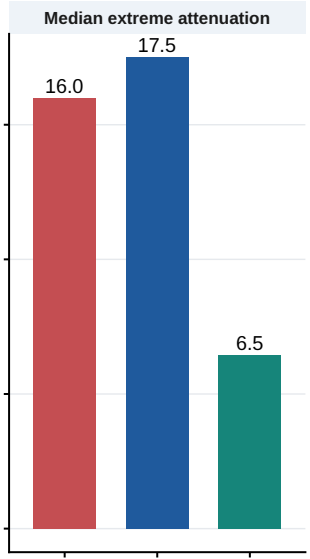
b Predictor and outcome windows

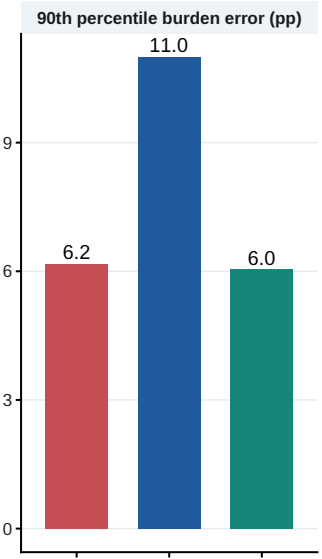


SICdb dense core (N = 2061); one-minute reference

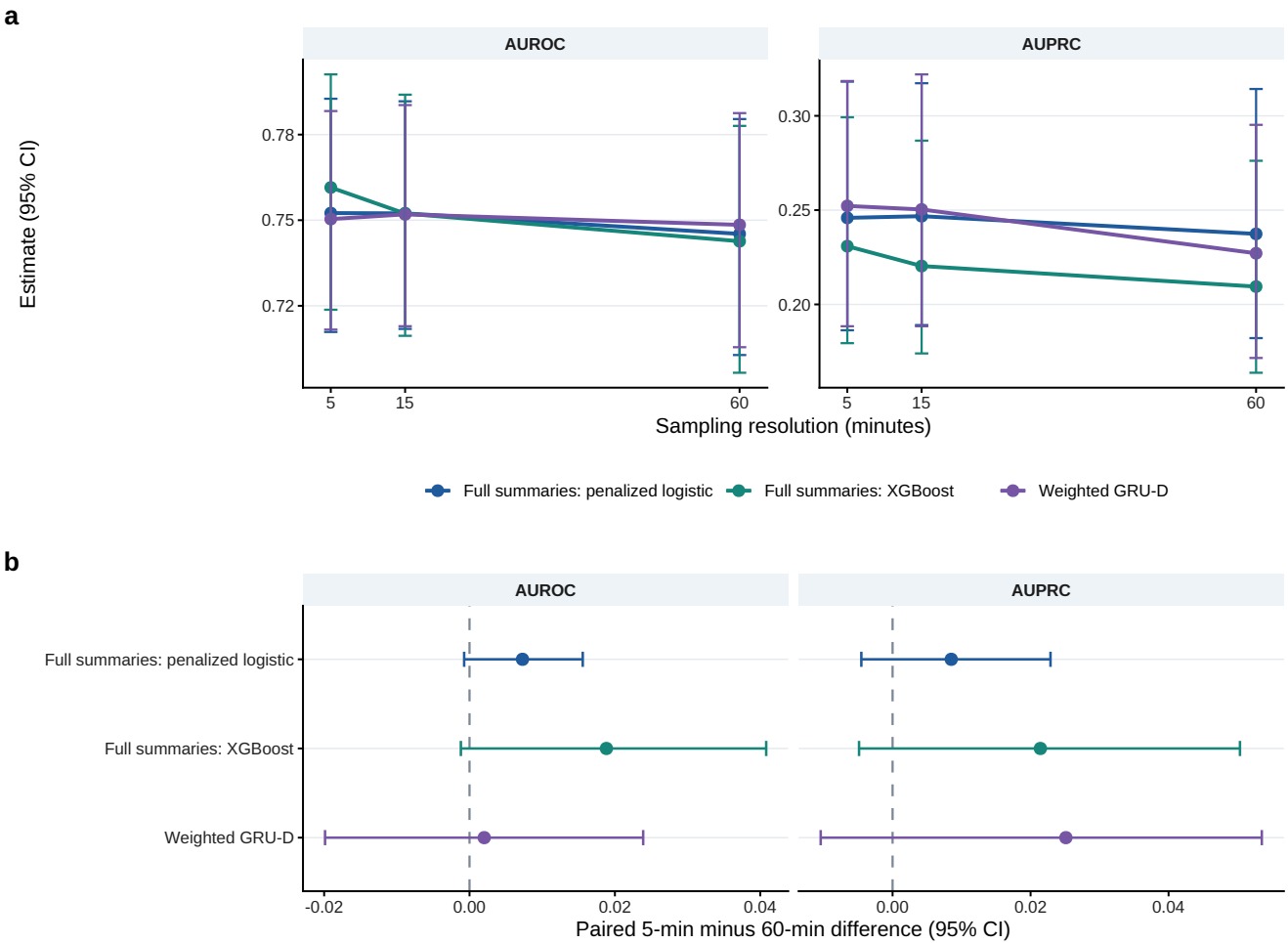


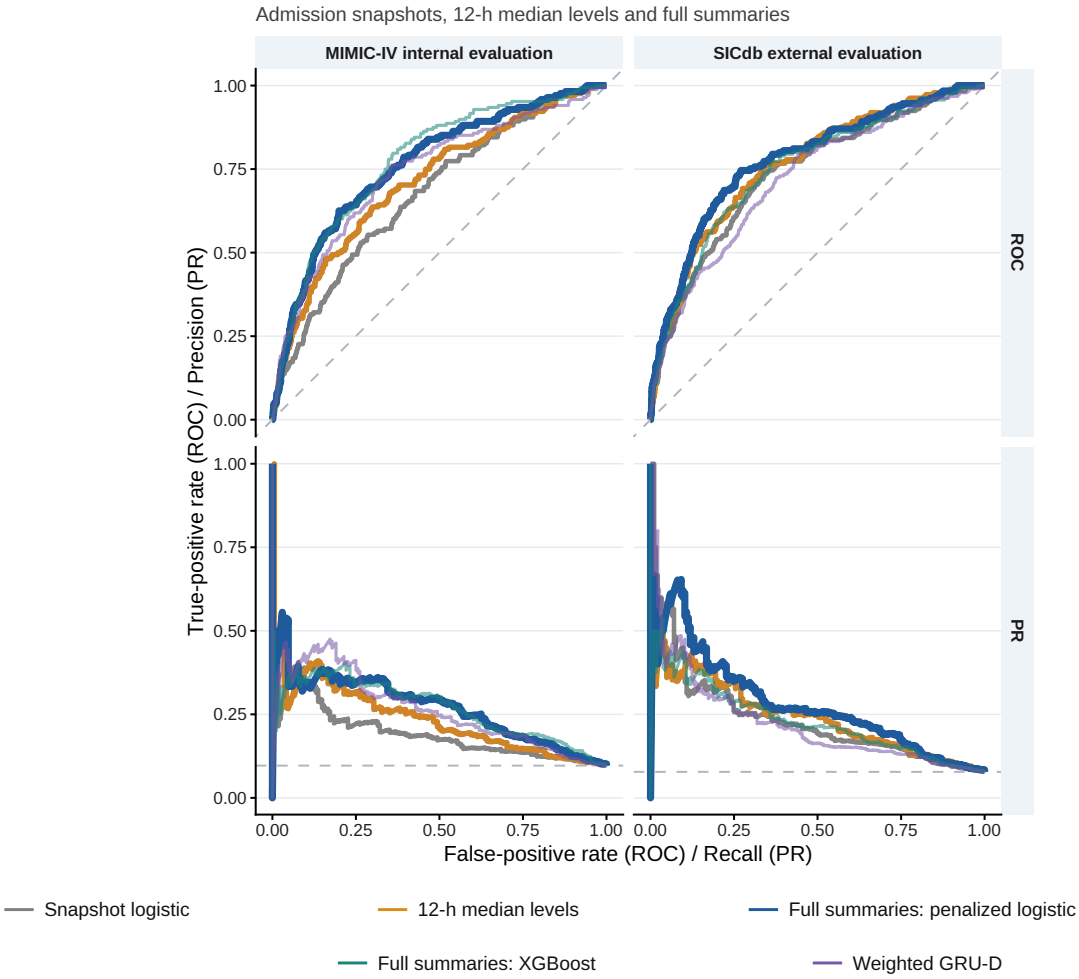


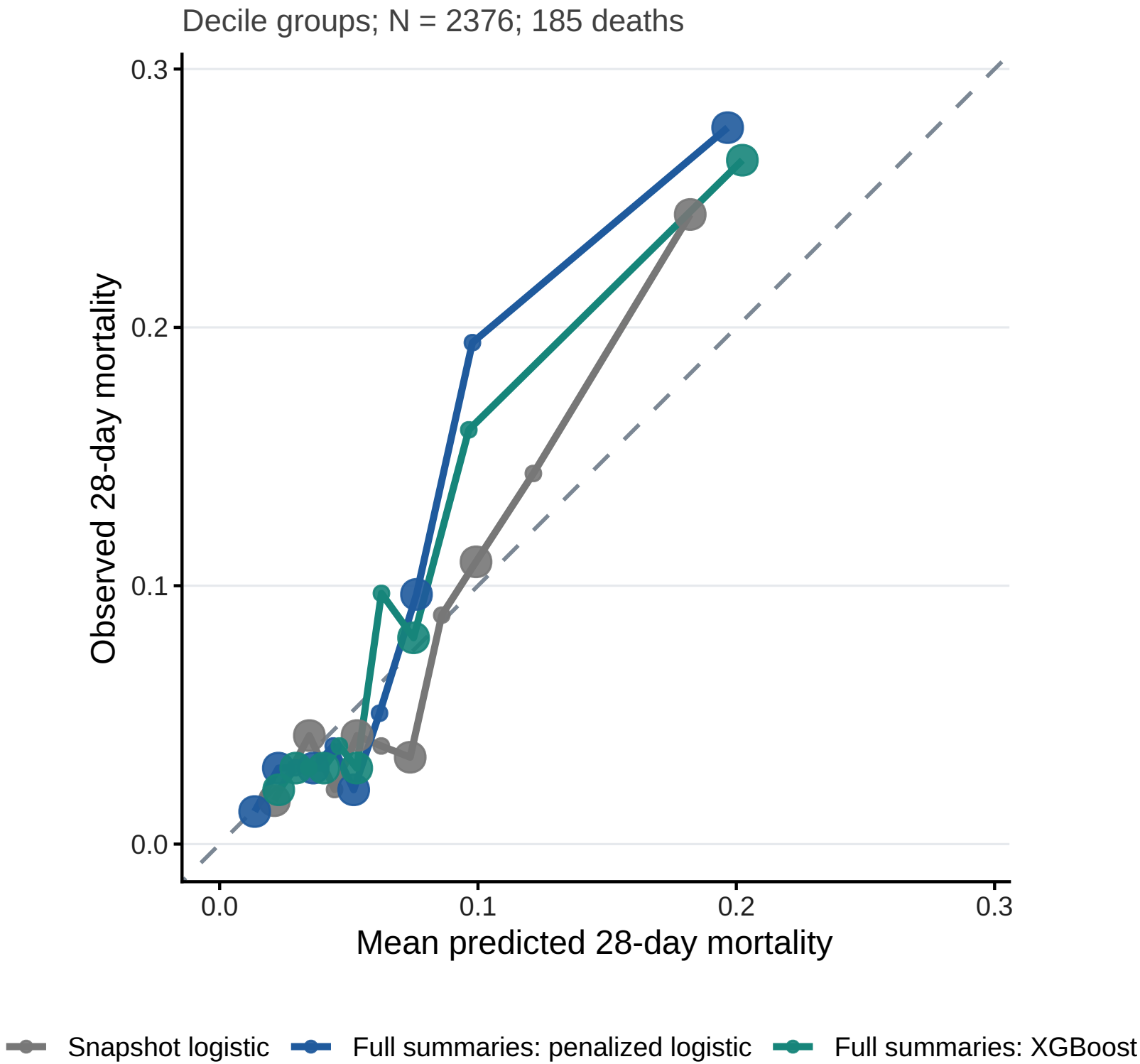




SICdb dense core: N = 2061; 151 deaths; within-cohort repeated cross-validation




Admission snapshots, 12-h median levels and full summaries
MIMIC-IV internal evaluation
SICdb external evaluation
ROC
PR
True-positive rate (ROC) / Precision (PR)
False-positive rate (ROC) / Recall (PR)
0.00
0.25
0.50
0.75
1.00
Snapshot logistic
12-h median levels
Full summaries: penalized logistic
Full summaries: XGBoost
Weighted GRU-D


Decile groups; N = 2376; 185 deaths
Observed 28-day mortality
0.0
0.1
0.2
0.3
Mean predicted 28-day mortality
0.0
0.1
0.2
0.3
Snapshot logistic
Full summaries: penalized logistic
Full summaries: XGBoost

Supplementary Files

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

- Additionalfile1SupplementaryMaterial.docx
- Additionalfile2AggregateSources.zip