

A Rolling-Window Framework for Dynamic Prediction of Delirium in Mechanically Ventilated ICU Patients: A Study Using MIMIC-IV

Qinyan Liang

Guangdong Medical University

Weijin Huang

Guangdong Medical University

Mengyi Zhang

Guangdong Medical University

Xilin Liu

Guangdong Medical University

Yingai Cui

Guangdong Medical University

Haifen Liao

Guangdong Medical University

Wanqing Hu

Houjie School of Clinical Medicine, Guangdong Medical University (Houjie Hospital of Dongguan)

Cuiqin Chen

Houjie School of Clinical Medicine, Guangdong Medical University (Houjie Hospital of Dongguan)

Jianrong Zhang

Houjie School of Clinical Medicine, Guangdong Medical University (Houjie Hospital of Dongguan)

Hong Zhang

zhanghong1214@gdmu.edu.cn

Guangdong Medical University

Article

Keywords:

Posted Date: July 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9327142/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

A Rolling-Window Framework for Dynamic Prediction of Delirium in Mechanically Ventilated ICU Patients: A Study Using MIMIC-IV

Qinyan Liang^{1,+}, Weijin Huang^{1,+}, Mengyi Zhang^{1,+}, Xilin Liu^{1,+}, Yingai Cui¹, Haifen Liao¹, Wanqing Hu², Cuiqin Chen², Jianrong Zhang², and Hong Zhang^{1,*}

¹School of Nursing, Guangdong Medical University, Dongguan Guangdong, China

²Houjie School of Clinical Medicine, Guangdong Medical University (Houjie Hospital of Dongguan), Dongguan, Guangdong, China

*corresponding author: Hong Zhang, email: zhanghong1214@gdmu.edu.cn

⁺these authors contributed equally to this work

ABSTRACT

Delirium affects 20–80% of mechanically ventilated intensive care unit (ICU) patients and is independently associated with prolonged mechanical ventilation, increased mortality, and long-term cognitive impairment. Existing prediction models rely on static admission features and fail to capture the dynamic physiological trajectory preceding delirium onset. We developed a rolling-window prediction framework using the MIMIC-IV database, enrolling 16,011 mechanically ventilated ICU stays. A 12-hour observation window with an 8-hour prediction horizon was applied at 8-hour nursing shift intervals to generate temporally resolved risk estimates. Three models were compared: static logistic regression, dynamic logistic regression, and XGBoost. The dynamic logistic regression model achieved the highest AUROC of 0.627 (95% CI: 0.610–0.643), outperforming the static baseline (AUROC 0.570, Δ AUROC = +0.057). The XGBoost model achieved an AUROC of 0.617 but provided superior probability calibration. Sensitivity analyses across three outcome definitions and three prediction horizons confirmed consistent incremental value of dynamic features. SHAP analysis identified late ICU phase, maximum heart rate, and mean temperature as dominant predictors, with an interaction effect between fever and sedation fluctuation. This shift-aligned dynamic framework addresses key limitations of static delirium prediction models and provides temporally resolved risk trajectories that may support future early-warning systems in the ICU.

Introduction

Delirium is an acute neuropsychiatric syndrome characterized by disturbances in attention, awareness, and cognition, affecting 20–80% of mechanically ventilated intensive care unit (ICU) patients^{1–3}. It is independently associated with prolonged mechanical ventilation, increased ICU and hospital mortality, higher healthcare costs, and long-term cognitive impairment^{4,5}. Despite its clinical significance, delirium remains underrecognized in routine ICU care, with studies reporting detection rates as low as 32% without structured assessment tools⁶.

Early identification of patients at high risk for delirium is critical for timely implementation of preventive strategies, including the ABCDEF bundle (Assess, prevent and manage pain; Both spontaneous awakening and breathing trials; Choice of analgesia and sedation; Delirium assess, prevent and manage; Early mobility; Family engagement)^{7,8}. However, existing prediction models are predominantly static, relying on admission-time features such as age, illness severity scores, and comorbidities^{9,10}. These models fail to capture the dynamic physiological trajectory that precedes delirium onset, and are therefore limited in their ability to generate timely, actionable alerts during ICU care.

Recent advances in electronic health record (EHR) infrastructure have enabled the collection of high-frequency time-series data in the ICU, including continuous vital signs, sedation assessments, laboratory values, and medication administration records. Several studies have attempted to leverage this longitudinal data for dynamic delirium prediction^{11,12}, yet key methodological challenges remain unaddressed. First, most dynamic models use fixed observation windows that do not align with clinical workflows, limiting their practical utility. Second, the incremental value of dynamic features over static predictors has rarely been formally quantified. Third, the impact of outcome definition heterogeneity—a well-recognized challenge in ICU delirium research—on model performance has not been systematically evaluated.

To address these gaps, we developed a rolling-window prediction framework using the MIMIC-IV database¹³,

a large publicly available ICU database. Our framework employs a 12-hour observation window with an 8-hour prediction horizon, aligned with standard nursing shift intervals, and generates dynamic risk estimates at each step. We developed and compared three models of increasing complexity—static logistic regression, dynamic logistic regression, and XGBoost—to formally quantify the incremental predictive value of time-varying features. We further conducted systematic sensitivity analyses across multiple outcome definitions and prediction horizons to assess the robustness of our framework.

The primary contributions of this work are threefold: (1) a shift-aligned rolling-window prediction framework that enables real-time delirium risk stratification; (2) formal quantification of the incremental value of dynamic physiological and pharmacological features over static predictors; and (3) a systematic evaluation of the effect of outcome definition quality on model performance, providing methodological guidance for future ICU delirium prediction studies.

Results

Study Cohort

A total of 94,458 ICU admissions were identified in MIMIC-IV, of which 16,011 met all eligibility criteria and were included in the final analysis (Figure 1). The largest exclusion was for ICU stays shorter than 48 hours ($n = 48,116$), followed by absence of mechanical ventilation ($n = 22,210$), neurological disease or dementia ($n = 4,868$), and prevalent delirium within the first 24 hours of admission ($n = 3,253$). No patients were excluded for age.

Baseline characteristics of the final cohort are presented in Table 1. The median age was 66.0 years (IQR: 55.0–75.0), and 59.8% were male. The median SOFA score at ICU admission was 6 (IQR: 4–9). Among the 16,011 patients, 4,491 (28.0%) developed incident delirium during their ICU stay. Compared with non-delirious patients, delirious patients were older, had higher SOFA scores, and longer ICU length of stay (all $p < 0.05$). Delirium was ascertained by medication proxy in the majority of cases (3,492/4,491, 77.8%), followed by ICD diagnostic codes (903/4,491, 20.1%) and CAM-ICU assessments (96/4,491, 2.1%).

The rolling-window procedure generated a total of 259,166 prediction windows across all 16,011 patients. Under the primary label strategy (positive if delirium onset fell within the 8-hour prediction horizon), 3,493 windows (1.35%) were labeled positive, covering 3,493 of 4,491 (77.8%) delirious patients.

Table 1. Baseline characteristics of the study cohort.

Characteristic	Overall ($n = 16,011$)	Delirium ($n = 4,491$)	No Delirium ($n = 11,520$)	SMD
Age (years), median (IQR)	66.0 (55.0–75.0)	65.0 (54.0–75.0)	66.0 (56.0–75.0)	0.090
Male sex, n (%)	9,570 (59.8)	2,704 (60.2)	6,866 (59.6)	0.012
SOFA score, median (IQR)	6 (4–9)	7 (4–10)	6 (4–9)	0.130
ICU LOS (hours), median (IQR)	113.3 (72.6–214.3)	202.0 (118.7–362.0)	94.4 (65.6–160.6)	0.679
Delirium source, n (%)				
CAM-ICU	—	96 (2.1)	—	
ICD codes	—	903 (20.1)	—	
Medication proxy	—	3,492 (77.8)	—	

IQR, interquartile range; ICU, intensive care unit; LOS, length of stay; SOFA, Sequential Organ Failure Assessment; CAM-ICU, Confusion Assessment Method for the ICU; ICD, International Classification of Diseases; SMD, standardized mean difference.

Model Performance

Model performance on the held-out test set (30% of patients, chronological split) is summarized in Table 2. The static logistic regression (Model 1) achieved an AUROC of 0.570 (95% CI: 0.552–0.588), reflecting the limited predictive value of admission-time features alone. Incorporating dynamic time-varying features in the logistic regression (Model 2) improved AUROC to 0.627 (95% CI: 0.610–0.643), representing an increment of $\Delta\text{AUROC} = +0.057$ over the static baseline. The XGBoost classifier (Model 3) achieved an AUROC of 0.617 (95% CI: 0.601–0.633) and $\Delta\text{AUROC} = +0.047$ relative to Model 1. AUPRC was 0.017, 0.021, and 0.020 for Models 1, 2, and 3, respectively. While Model 2 showed marginally higher discrimination, Model 3 demonstrated superior calibration with the lowest Brier score of 0.193. ROC and precision-recall curves are shown in Figure 2.

Table 2. Model performance on the held-out test set.

Model	AUROC	95% CI	AUPRC	Brier	Δ AUROC
Model 1: Static LR	0.570	0.552–0.588	0.017	0.245	Reference
Model 2: Dynamic LR	0.627	0.610–0.643	0.021	0.236	+0.057
Model 3: XGBoost	0.617	0.601–0.633	0.020	0.193	+0.047

AUROC, area under the receiver-operating characteristic curve; AUPRC, area under the precision-recall curve; LR, logistic regression; Δ AUROC, increment over Model 1. Bootstrap 95% confidence intervals based on 500 resampling iterations.

Decision curve analysis demonstrated that net benefit across all three models was limited at clinically relevant threshold probabilities (5-25%), consistent with the low positive rate (1.35%) of the prediction windows. Model 3 showed marginally higher net benefit at thresholds above 0.30 compared with the static baseline (Fig. 4).

SHAP analysis of the XGBoost model identified prolonged ICU stay and physiological instability as the most influential predictors, with late ICU phase (`icu_phase_late`), maximum heart rate (`hr_max`), and mean temperature (`temp_mean`) as the top-ranked features (Figure 3). Neurological status (`gcs_mean`) and sedative doses (`propofol_dose`) also contributed substantially. The dependence plot demonstrated an interaction between maximum temperature and RASS standard deviation (`temp_max` $\times$ `rass_std`), suggesting a compound effect of fever and sedation fluctuation on delirium risk.

Sensitivity Analysis

Results of sensitivity analyses are presented in Supplementary Table S1. Across outcome definitions, model performance was highest under the full composite definition (XGBoost AUROC 0.636 (95% CI: 0.627–0.646)), with substantially wider confidence intervals observed when restricting to CAM-ICU alone (AUROC 0.574, 95% CI: 0.508–0.638) due to limited positive cases ($n = 72$ in the test set). Notably, the CAM-ICU-only subset yielded a higher static Model 1 AUROC (0.651) than XGBoost (0.574), suggesting that with cleaner but sparser outcome labels, simpler models may be less prone to overfitting on noisy training signals.

Across prediction horizons, AUROC increased monotonically from 0.625 (8-hour horizon) to 0.631 (16-hour horizon) to 0.636 (24-hour horizon) for the XGBoost model, while the static baseline showed a similar but attenuated trend (0.569 to 0.577 to 0.585). The Δ AUROC between Model 3 and Model 1 remained stable across all horizons (range: +0.051 to +0.056), demonstrating that the incremental value of dynamic features is robust to the choice of prediction horizon.

Competing Risk Analysis

Cumulative incidence functions accounting for ICU death as a competing event are shown in Supplementary Figure S1. Among the 16,011 patients, 1,386 (8.7%) died in the ICU without prior delirium and were classified as competing events. The cumulative incidence of delirium rose steeply during the first 72 hours of ICU admission, reaching 11.2% by day 3 and 26.7% by day 7. The cumulative incidence of ICU death without delirium was 2.2% at day 3 and 9.2% at day 7. The separation between the two CIF curves confirms that delirium and ICU death represent distinct competing outcomes, and that failure to account for competing risks would lead to overestimation of the marginal delirium probability in standard survival analyses¹⁴.

Discussion

In this study, we developed and validated a rolling-window dynamic prediction framework for incident delirium in mechanically ventilated ICU patients using the MIMIC-IV database. The framework generated temporally resolved risk estimates aligned with nursing shift intervals, providing high-resolution temporal risk trajectories that offer mechanistic insights into delirium progression and lay the groundwork for future early-warning systems.

Dynamic features provided meaningful incremental predictive value over static admission features. The dynamic logistic regression model achieved a Δ AUROC of +0.057 over the static baseline, demonstrating that time-varying physiological and pharmacological features capture delirium-relevant information not available at admission. This finding extends recent dynamic prediction studies in the ICU¹¹ by formally quantifying incremental value through a three-model comparison design. The Δ AUROC remained stable across all three prediction horizons (8h, 16h, 24h), indicating that the benefit of dynamic features reflects genuine signal in the time-series data rather than an artifact of a particular window configuration. Interestingly, the dynamic logistic regression (Model 2) marginally

outperformed the XGBoost algorithm (Model 3) in terms of AUROC (0.627 vs. 0.617), suggesting that the predictive signal in the available features is largely linear, though gradient boosting provided superior probability calibration (Brier score 0.193 vs. 0.236).

Prolonged ICU stay and physiological instability emerged as the dominant predictors of imminent delirium. SHAP analysis identified late ICU phase as the most important feature, followed by cardiovascular and temperature-related variables (maximum heart rate, mean temperature) and baseline neurological status (mean GCS). The interaction effect between fever and sedation fluctuation (RASS standard deviation) provides empirical support for the hypothesized pathway wherein systemic inflammation and fluctuating consciousness synergistically potentiate ICU delirium^{5,15}. These findings suggest that maintaining physiological stability and minimizing sedation fluctuations in febrile patients may represent high-yield intervention targets, consistent with current guidelines emphasizing early mobilization and light sedation¹⁶.

Outcome definition quality substantially influenced model performance. Restricting the outcome to CAM-ICU assessments alone yielded wider confidence intervals and counterintuitively favored the static model over XGBoost, likely due to insufficient positive cases for gradient boosting to learn stable decision boundaries. Conversely, the composite outcome definition improves statistical power but introduces label noise from medication proxy cases where antipsychotic use may not reflect confirmed delirium¹⁷. This trade-off between label quality and sample size is a fundamental methodological challenge in retrospective ICU delirium research that has been insufficiently acknowledged in prior machine learning studies^{6,10}. Future studies should explicitly report model performance stratified by outcome definition to enable meaningful cross-study comparisons.

Several prior studies have developed machine learning models for ICU delirium prediction^{9–11}. Gong et al.¹¹ achieved AUROC of 0.77 in internal and 0.72 in external validation with an XGBoost model, though generating a single prediction per admission rather than repeated dynamic estimates. Wong et al.⁹ reported AUROC of 0.855 using EHR-derived features in a non-ICU hospitalized cohort, but did not account for competing risks or evaluate outcome definition heterogeneity. Our framework differs in three key respects: predictions are generated at each nursing shift using a sliding window; the incremental value of dynamic features is explicitly quantified through a multi-model comparison; and sensitivity to outcome definition is systematically characterized. The modest absolute AUROC (0.626) relative to some prior reports likely reflects stricter outcome ascertainment and a chronological rather than random train-test split, which prevents optimistic performance inflation from temporal correlation¹⁰.

Several limitations should be acknowledged. The study was conducted in a single academic medical center, which may limit generalizability to ICU settings with different patient populations, staffing models, or sedation protocols. The medication proxy outcome definition, comprising 77.8% of delirium cases, introduces potential misclassification, as antipsychotics may be prescribed for indications other than delirium¹⁷. Nevertheless, antipsychotic administration in the ICU reflects real-world clinical responses to acute agitation or delirium-like states, and carries pragmatic validity in settings where systematic CAM-ICU screening is absent. CAM-ICU documentation in MIMIC-IV is sparse (96 positive assessments), reflecting real-world documentation patterns rather than systematic screening. Although SHAP values identify important features, causal inference cannot be drawn from observational data without additional experimental validation. Furthermore, the machine learning models employed are standard binary classifiers and do not directly incorporate competing risk penalisation into the loss function; future work may consider cause-specific hazard models or Fine–Gray regression as alternatives. Finally, the framework was not prospectively evaluated in a clinical setting, and implementation studies are needed to assess its real-world impact on delirium prevention and clinical workflows.

Methods

Data Source and Study Population

This retrospective cohort study utilized the Medical Information Mart for Intensive Care IV (MIMIC-IV, version 2.2) database, a publicly available electronic health record repository containing de-identified data from critically ill patients admitted to the Beth Israel Deaconess Medical Center between 2008 and 2019¹³. Access was obtained following completion of the required data use agreement and credentialing process (certification number: 68330408).

Adult patients (≥ 18 years) admitted to the ICU who received mechanical ventilation for ≥ 48 hours were eligible for inclusion. Patients were excluded if they had: (1) a neurological diagnosis or pre-existing dementia (ICD-9/10 codes for stroke, traumatic brain injury, dementia, or encephalopathy) that could confound delirium assessment; or (2) documented delirium within the first 24 hours of ICU admission, as the prediction framework targets incident delirium rather than prevalent delirium at admission. The final study cohort comprised 16,011 ICU stays. This study used publicly available, de-identified data from the MIMIC-IV database. As no identifiable patient information was

accessed, institutional review board approval and informed consent were not required. All methods were performed in accordance with the relevant guidelines and regulations (e.g., Declaration of Helsinki).

Outcome Definition

Given the well-recognized challenge of delirium ascertainment in retrospective EHR data⁵, we constructed a composite outcome using three complementary sources. The primary outcome was incident delirium, defined as the first occurrence of any of the following after 24 hours of ICU admission: (1) a positive CAM-ICU assessment (itemid 228300 in MIMIC-IV chartevents); (2) an ICD-9/10 diagnostic code for delirium (F05.x, 293.x); or (3) first administration of an antipsychotic medication used as a delirium proxy (haloperidol [NDC 221385] or quetiapine [NDC 225152]) after 24 hours of ICU admission^{11,16,17}. The delirium onset time was defined as the earliest documented event across these three sources. Sensitivity analyses were conducted using three progressively restrictive outcome definitions: (i) the full composite (CAM-ICU + ICD + medication proxy); (ii) CAM-ICU plus ICD codes only; and (iii) CAM-ICU alone.

Rolling-Window Prediction Framework

To generate dynamic, temporally resolved predictions, we implemented a rolling-window framework that slides across each patient's ICU stay at regular intervals (Fig. 2). For each patient, a 12-hour observation window was used to extract time-varying features, followed by an 8-hour prediction horizon. Windows were advanced in 8-hour steps, aligning with standard ICU nursing shift intervals to facilitate clinical deployment. The prediction time point was defined as the end of each observation window.

Windows were excluded if: (i) the documented delirium onset occurred prior to the prediction time point (t_{pred}) to prevent data leakage; or (ii) the patient had fewer than 12 hours of remaining ICU stay at the window start. Delirium labels were assigned using strategy C: a window was labeled positive if the delirium onset time fell within the 8-hour prediction horizon (i.e., between t_{pred} and $t_{\text{pred}} + 8$ hours), capturing the imminent risk period clinically relevant for intervention.

Feature Engineering

Features were extracted from each 12-hour observation window and categorized into static and dynamic components. Static features included age, SOFA score at ICU admission, and flags for sedative medication classes received prior to window start (propofol, midazolam, dexmedetomidine). Dynamic features were computed from time-series data within each window and included:

- *Sedation depth*: Richmond Agitation-Sedation Scale (RASS) mean, standard deviation, slope, deep sedation ratio ($\text{RASS} \leq -3$), and cumulative deep sedation hours.
- *Fever and inflammation*: maximum and mean temperature, fever hours (temperature $>38.3^\circ\text{C}$), and white blood cell count with rate of change.
- *Hemodynamics and oxygenation*: heart rate mean, standard deviation, maximum, tachycardia flag ($\text{HR} >100$ bpm), SpO_2 mean, minimum, hypoxia flag ($\text{SpO}_2 <92\%$), and hypoxia hours.
- *Neurological status*: Glasgow Coma Scale (GCS) mean, minimum, standard deviation, and slope, derived from the sum of eye, verbal, and motor subscale scores.
- *Sedative pharmacology*: propofol, midazolam, dexmedetomidine, and total benzodiazepine dose within the window.
- *Interaction terms*: deep sedation ratio $\times$ maximum temperature; deep sedation ratio $\times$ WBC rate of change, capturing the hypothesized synergistic effect of neuroinflammation and sedation on delirium risk⁵.
- *ICU phase*: categorical indicator of ICU stay phase (early: $\leq 48\text{h}$; mid: $48\text{--}120\text{h}$; late: $>120\text{h}$).

Statistical Analysis and Modeling

Three models of increasing complexity were developed. Model 1 (static logistic regression) used only static admission features as a clinical baseline, analogous to conventional risk scores. Model 2 (dynamic logistic regression) incorporated all static and dynamic features to assess the contribution of time-varying data within a linear framework. Model 3 (XGBoost) used gradient boosting with all features to capture nonlinear interactions and higher-order feature dependencies¹⁸.

The dataset was partitioned chronologically by first ICU admission time, with 70% of patients allocated to training and 30% to testing, to prevent temporal data leakage. Class imbalance was addressed using inverse class-frequency weighting. Missing values were imputed using training-set feature medians.

Model performance was assessed using the area under the receiver-operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and Brier score. The incremental value of dynamic features was quantified as Δ AUROC (Model 2 or 3 minus Model 1). Bootstrap 95% confidence intervals were computed using 500 resampling iterations with replacement. Decision curve analysis (DCA) was performed to evaluate net clinical benefit across a range of decision thresholds¹⁹. Feature importance was assessed using SHapley Additive exPlanations (SHAP) values²⁰.

Competing risk analysis was conducted to account for ICU death as a competing event for delirium, using the Aalen-Johansen estimator to compute cause-specific cumulative incidence functions (CIF)¹⁴. This analysis served as a descriptive characterisation of the study cohort; the machine learning models themselves are standard binary classifiers and do not directly incorporate competing risk penalisation into the loss function.

Sensitivity analyses were conducted across: (i) three outcome definitions (composite, CAM-ICU+ICD, CAM-ICU only); and (ii) three prediction horizons (8h, 16h, 24h). All analyses were performed in Python 3.13 using scikit-learn, XGBoost, and SHAP libraries. The study followed the TRIPOD reporting guideline for prediction model development²¹. Linguistic editing of this manuscript was assisted by Claude (Anthropic, San Francisco, CA), a large language model. The authors take full responsibility for all scientific content.

Data availability

The datasets analyzed during the current study are available in the Medical Information Mart for Intensive Care IV (MIMIC-IV, version 2.2) database on PhysioNet (<https://physionet.org/content/mimiciv/2.2/>). Access to the database requires completion of a data use agreement and human subjects training.

Code availability

The analysis code is publicly available at https://github.com/EdwardoStarks/ICU_Delirium.

References

1. Ely, E. W., Margolin, R., Francis, J. *et al.* Evaluation of delirium in critically ill patients: validation of the confusion assessment method for the intensive care unit (CAM-ICU). *Critical Care Medicine* **29**, 1370–1379 (2001).
2. Salluh, J. I., Wang, H., Schneider, E. B. *et al.* Outcome of delirium in critically ill patients: systematic review and meta-analysis. *BMJ* **350**, h2538 (2015).
3. Wu, N.-N., Zhang, Y.-B., Wang, S.-Y., Zhao, Y.-H. & Zhong, X.-M. Incidence, prevalence and risk factors of delirium in icu patients: A systematic review and meta-analysis. *Nursing in Critical Care* **28**, 653–669 (2023).
4. Pandharipande, P. P., Girard, T. D., Jackson, J. C. *et al.* Long-term cognitive impairment after critical illness. *New England Journal of Medicine* **369**, 1306–1316 (2013).
5. Wilson, J. E., Mart, M. F., Cunningham, C. *et al.* Delirium. *Nature Reviews Disease Primers* **8**, 54 (2022).
6. Ruppert, M. M., Lipori, J., Patel, S. *et al.* ICU delirium-prediction models: a systematic review. *Critical Care Explorations* **2**, e0296 (2020).
7. Devlin, J. W., Skrobik, Y., Gelinas, C. *et al.* Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. *Critical Care Medicine* **46**, e825–e873 (2018).
8. Sosnowski, K. *et al.* The effect of the abcde/abcdef bundle on delirium, functional outcomes, and quality of life in critically ill patients: a systematic review and meta-analysis. *International Journal of Nursing Studies* **138**, 104410 (2023).
9. Wong, A. *et al.* Development and validation of an electronic health record-based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Network Open* **1**, e181018 (2018).
10. Al-Jabri, M. M. & Anshasi, H. Performance of machine and deep learning models for predicting delirium in adult icu patients: A systematic review. *International Journal of Medical Informatics* **203**, 106008 (2025).

11. Gong, K. D., Lu, R., Bergamaschi, T. S. *et al.* Predicting intensive care delirium with machine learning: model development and external validation. *Anesthesiology* **138**, 299–311 (2023).
12. Zhang, Y. *et al.* Development of a machine learning-based prediction model for sepsis-associated delirium in the intensive care unit. *Scientific Reports* **13**, 12697 (2023).
13. Johnson, A. E., Bulgarelli, L., Shen, L. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* **10**, 1 (2023).
14. Austin, P. C. & Fine, J. P. Practical recommendations for reporting fine-gray model analyses for competing risk data. *Statistics in Medicine* **36**, 4391–4400 (2017).
15. Fan, Y.-Y. *et al.* Mechanisms underlying delirium in patients with critical illness. *Frontiers in Aging Neuroscience* **16**, 1446523 (2024).
16. Devlin, J. W., Skrobik, Y., Gélinas, C. *et al.* Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the icu. *Critical Care Medicine* **46**, e825–e873 (2018).
17. Jaworska, N. *et al.* Facilitators and barriers influencing antipsychotic medication prescribing and deprescribing practices in critically ill adult patients: a qualitative study. *Journal of General Internal Medicine* **38**, 2262–2271 (2023).
18. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).
19. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **26**, 565–574 (2006).
20. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30**, 4765–4774 (2017).
21. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594 (2015).
22. Goldberger, A. L. *et al.* Physiobank, physiotoolkit, and physionet. *Circulation* **101**, DOI: [10.1161/01.cir.101.23.e215](https://doi.org/10.1161/01.cir.101.23.e215) (2000). [Online; accessed 2026-01-23].

Acknowledgements

We thank the MIT Laboratory for Computational Physiology and PhysioNet for granting access to the MIMIC-IV database²². We are grateful to the developers of the open-source MIMIC-IV derived tables for their contributions to data preprocessing. Linguistic editing of this manuscript was assisted by Claude (Anthropic, San Francisco, CA), an artificial intelligence language model. The authors take full responsibility for the integrity and accuracy of all scientific content.

Funding

This work was supported by the Guangdong College Students' Innovation and Entrepreneurship Training Program (Grant No. S202510571066).

Author contributions

Q.L., W.H., M.Z., and X.L. contributed equally to this work. Q.L. and W.H. conceived the study design and performed the data extraction and preprocessing. M.Z. and X.L. conducted the statistical analyses and developed the prediction models. Y.C. and H.L. reviewed the clinical relevance of the findings and contributed to result interpretation. W.Q.H., C.C., and J.Z. contributed to literature review and manuscript revision. H.Z. supervised the project, provided critical revisions to the manuscript, and is responsible for the overall integrity of the work. All authors reviewed and approved the final manuscript.

Additional information

Competing interests The authors declare no competing interests.

Figure 1 Study Flowchart

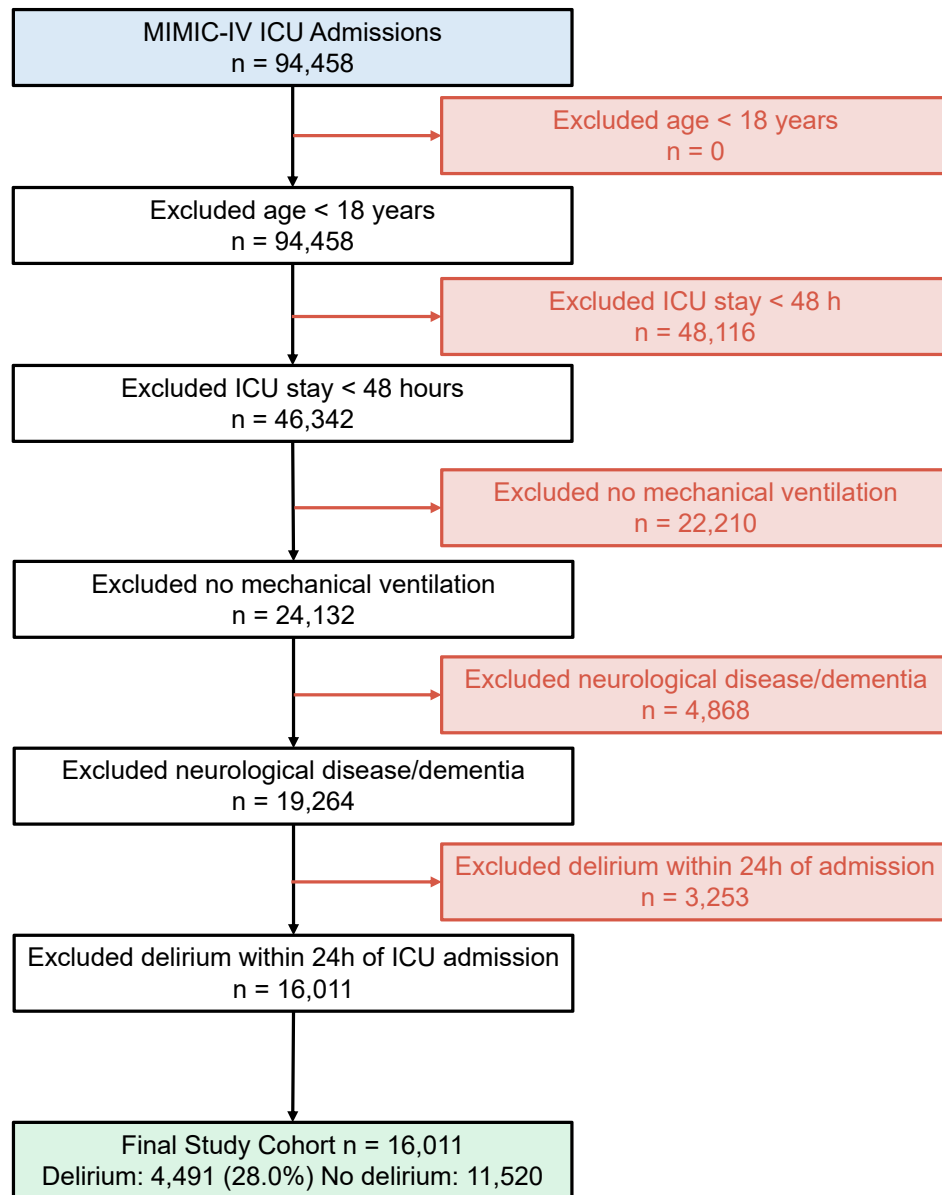


Figure 1. Study flowchart | Consecutive exclusion criteria applied to MIMIC-IV ICU admissions to derive the final study cohort.

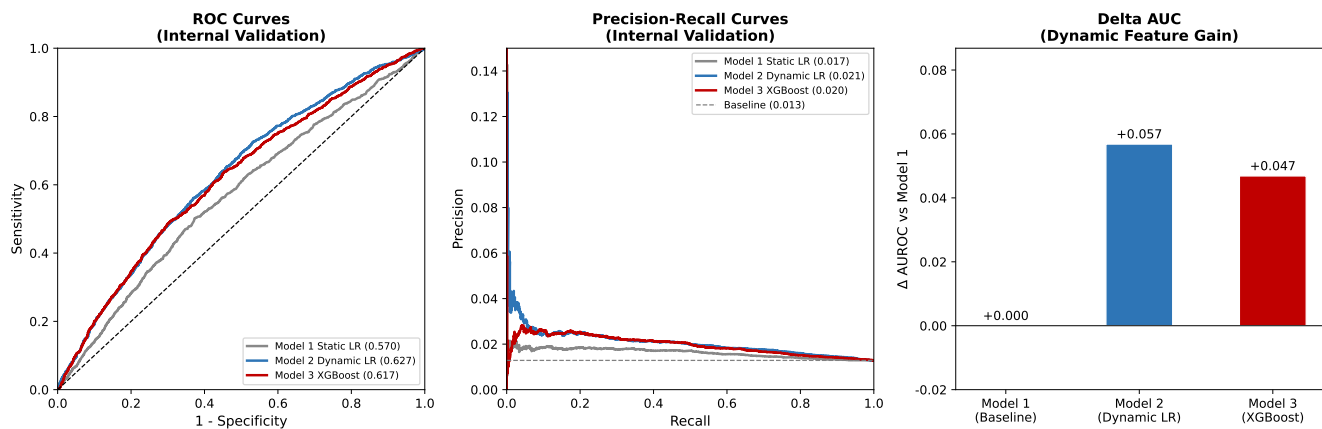


Figure 2. Model performance | Receiver-operating characteristic (ROC) curves, precision-recall curves, and incremental AUROC (Δ AUROC) for the three prediction models. Model 1: static logistic regression; Model 2: dynamic logistic regression; Model 3: XGBoost.

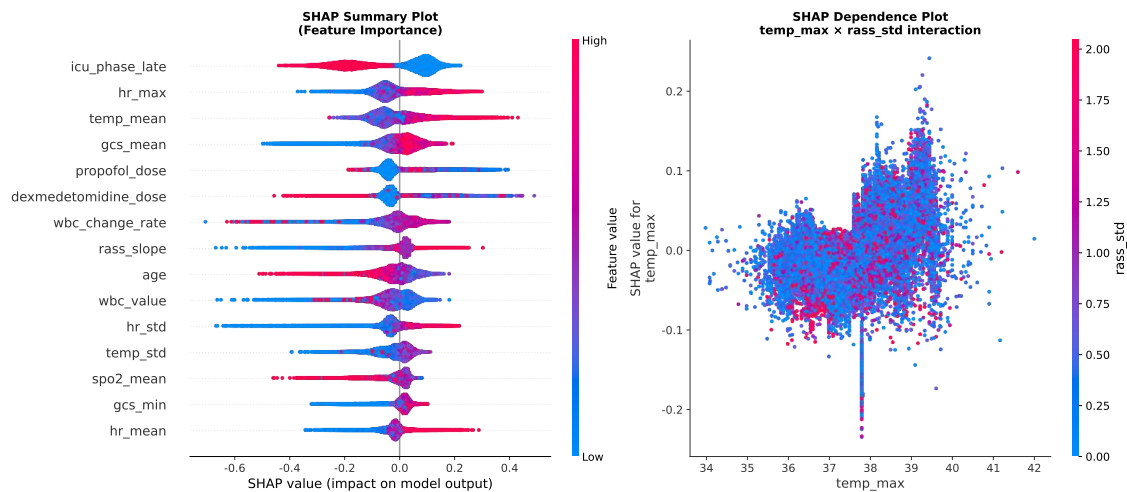


Figure 3. SHAP feature importance | Features are ranked by mean absolute SHAP value for the XGBoost model (Model 3). The interaction between maximum temperature and RASS standard deviation ($\text{temp_max} \times \text{rass_std}$) is highlighted.

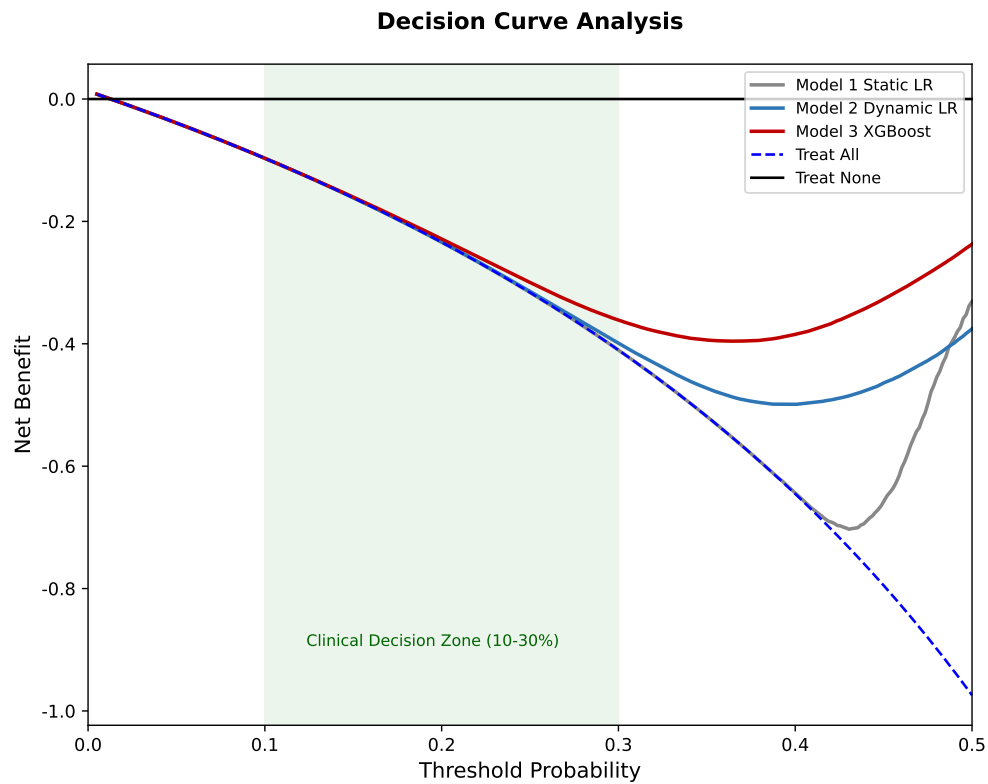


Figure 4. Decision curve analysis | Net benefit of the three prediction models across a range of threshold probabilities. The treat-all and treat-none strategies are shown as reference lines.

Supplementary Files

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

- [supplementary.docx](#)