Scientia Psychiatrica
Vol. 7, No. 3, 2026, pp. 171–180 · e-ISSN 2715-9736 journal homepage: scientiapsychiatrica.com

ORIGINAL RESEARCH ARTICLE

Dataset Overlap Constrains Validation of a Postoperative Delirium Model at Intensive Care Unit Admission

Karina Chandra[1*], Danniel Hilman Maulana[1]
[1]*Department of Anesthesiology and Intensive Care, CMHC Research Center, Palembang, Indonesia*

ARTICLE INFO

Article history
Received: May 19, 2026
Revised: June 12, 2026
Accepted: July 7, 2026
Available online: July 26, 2026
Published: July 30, 2026

Keywords
Calibration
Delirium
Intensive care unit
Prediction model
Validation

***Corresponding author**
Karina Chandra
E-mail: karina.chandra@cattleyacenter.id
ORCID: 0009-0004-0194-6110

DOI
10.37275/scipsy.v7i3.219

ABSTRACT

Background. Postoperative delirium is common in intensive care and linked to adverse functional and cognitive outcomes. Valid risk prediction requires transparent predictor timing, outcome definition, validation design, and cohort independence.

Objective. To reproduce and extend a published postoperative delirium model, quantify record overlap between its public development and validation worksheets, and determine what level of validation the data support.

Methods. We analyzed a public prospective-cohort workbook (300 development and 100 designated-validation records). The published five-predictor equation, a refitted logistic model, and a ten-predictor penalized logistic model were compared by repeated nested 10-fold cross-validation. Discrimination, Brier score, and calibration were assessed with bootstrap intervals. Exact cross-sheet matches required agreement on ten predictors, the Intensive Care Delirium Screening Checklist score, and the delirium label. Designated-worksheet and non-overlap analyses were sensitivity analyses.

Results. Delirium occurred in 82/300 development (27.3%) and 29/100 designated-validation records (29.0%). The area under the receiver operating characteristic curve (AUROC) was 0.838 (95% bootstrap interval 0.782–0.886) for the refitted five-predictor model and 0.840 (0.785–0.888) for the penalized model; Brier scores were 0.132 and 0.134, and calibration slopes 0.751 and 1.061. Seventy-six of 100 validation records matched development records exactly. Designated-worksheet AUROC was high (0.957–0.961) but not independent. The 24-record non-overlap subset had 17 events and seven non-events, so estimates were imprecise.

Conclusion. The parsimonious model showed reproducible internal discrimination, and penalization added no material gain. Because record independence is not supported, the validation worksheet does not establish transportability. Prospective evaluation in an independent cohort is required before clinical use.

Copyright: © 2026 The Author(s). Published by HM Publisher in collaboration with CMHC Research Center on behalf of Scientia Psychiatrica. The authors retain copyright of their work.

Access licence: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

Delirium is an acute disturbance of attention, awareness, and cognition. It commonly complicates critical illness and the postoperative period. In intensive care it is more than a transient behavioral event. A greater delirium burden has been linked to poorer functional neurologic recovery, longer hospitalization, cognitive impairment, and mortality.[1–3] Early risk recognition is therefore clinically attractive, particularly when modifiable precipitating factors can be addressed through multicomponent prevention, medication review, sleep promotion, mobilization, pain control, and avoidance of unnecessary deep sedation. The value of prediction lies in directing preventive attention before symptoms become established. It does not lie in replacing repeated bedside assessment. Many delirium risk models exist. They were proposed for medical, surgical, and intensive-care populations, and recent primary studies used conventional regression, simple risk scores, penalized models, machine learning, and electronic-health-record data, and they relied on internal, temporal, geographic, or prospective validation designs.[4–19] Reported performance varies with case mix, prediction time point, outcome surveillance, missing-data procedures, predictor availability, and the independence of the evaluation cohort. Provenance and validation design are therefore central to interpretation, not secondary reporting details.

Xing and colleagues developed a five-predictor equation for postoperative patients admitted to an intensive care unit in China.[20] The model combined a POSSUM score, an acid-base imbalance score, and histories of diabetes, hypertension, and coma. It was developed on one cohort. The report gave an area under the receiver operating characteristic curve (AUROC) of 0.852 during development and evaluated the equation in a separate group of 100 patients. The accompanying workbook, released in Dryad (doi:10.5061/dryad.gb5mkkwk3), allows reanalysis of individual records. Public access is especially valuable for prediction research, because it permits verification of the published equation, assessment of calibration, comparison with regularized alternatives, and a check on whether the designated partitions are analytically independent.

The distinction between internal and external validation matters. Random or resampling-based internal validation estimates optimism within a development sample, whereas temporal, geographic, or setting-based external validation examines transportability to a new data-generating context. Recent development and validation studies show that performance can differ across surgical specialties, ICU populations, institutions, and validation sources.[21–34] When identical records occur in both the development and evaluation partitions, the latter is no longer an independent test, because performance may partly reflect predictor-outcome combinations the model has already seen. A high AUROC does not by itself establish accurate risk estimates or clinical benefit. Calibration and decision consequences count as well. Current TRIPOD+AI and PROBAST+AI guidance stresses transparent data provenance, participant flow, predictor and outcome definitions, separation of model development from evaluation, and reporting of both discrimination and calibration.[35,36]

Penalization offers a principled way to control coefficient instability when candidate predictors are numerous relative to the number of events. Complexity does not guarantee better performance in every finite sample. Tuning must also stay inside the resampling loop to avoid optimistic selection. Recent external-validation, deep-learning, biomarker, and dynamic-prediction studies illustrate that apparent gains depend on the data source and the evaluation design.[37–46] A fair comparison should separate three things: the fixed published equation, an unpenalized refit of its original five predictors, and a penalized model using the wider set of candidate predictors available before outcome ascertainment. The Intensive Care Delirium Screening Checklist (ICDSC) should not enter as a predictor when the target is a delirium label derived from clinical screening, since doing so would introduce outcome information into prediction.

The novelty of this study is the combination of leakage-aware repeated nested validation with a record-level examination of cross-sheet independence and of ICDSC–outcome concordance in the openly archived dataset. This approach separates reproducible internal predictive performance from evidence that could support an independent-validation claim. The source report did not address that distinction. We aimed to (1) reproduce the published equation; (2) compare a refitted five-predictor logistic model with a ten-predictor penalized logistic model using repeated nested cross-validation; (3) quantify exact within- and cross-sheet record overlap and the concordance between ICDSC threshold status and the workbook delirium label; and (4) determine which level of validation the public data support.

2. Methods

2.1. Study design and setting

Study design and reporting framework. This was a reproducible secondary analysis of the individual-record workbook that accompanies a prospective postoperative ICU cohort.[20] We evaluated an existing prediction equation, refitted its predictors, and examined a penalized alternative. Reporting followed TRIPOD+AI principles for reporting prediction models and used PROBAST+AI concepts for participants and data sources, predictors, outcome, and analysis.[35,36] The workbook lacked stable participant identifiers, recruitment dates, and site identifiers. We therefore describe the unit of analysis as a record. References to patients are reserved for the design reported in the source article.

2.2. Participants

Data source and study population. The archived Excel workbook held a 300-record development worksheet and a 100-record worksheet designated as validation. The source study prospectively recruited adult postoperative patients admitted to an ICU in China and assessed delirium during the postoperative ICU course.[20] Every complete record in the two analysis worksheets was included. None of the 12 core fields used here had a missing value. No imputation was needed, and no record was excluded. The sample size was fixed by the archived workbook: with 82 events in the development worksheet, there were 8.2 events per candidate predictor and 16.4 events per predictor of the five-predictor model. The readme described the coding scheme and the published equation. The raw workbook was kept unchanged. All calculations used programmatically imported values, not cached spreadsheet formula results.

2.3. Variables and instruments

Outcome. The primary outcome was the binary Delirium field in the public workbook. The ICDSC score was treated as an outcome-assessment field and was never offered to a prediction model. The source article used an ICDSC-based assessment process and reported a threshold of at least four as indicative of delirium.[20] To characterize the archived labels without replacing them, we compared the binary outcome with ICDSC ≥4 and counted discordant records separately in each worksheet. This concordance analysis was descriptive. Because the public data held no adjudication notes, we

could not determine why a binary label and a single recorded ICDSC value differed.

Candidate predictors. Ten candidate predictors were available before or at postoperative ICU admission: sex, POSSUM score, operation duration in minutes, history of alcohol dependence, history of coma, mechanical ventilation, history of hypertension, history of diabetes, history of cerebral apoplexy (stroke), and acid-base imbalance score. Binary fields kept the source coding of 0 for absence and 1 for presence. Sex was coded 0 for female and 1 for male. The total POSSUM score supplied in the workbook was used. According to the readme, acid-base imbalance was graded by the APACHE III scheme and covered metabolic or respiratory acidosis or alkalosis found on the first blood-gas analysis.

Models. Three models were specified. The first was the fixed published equation, reproduced without alteration (the complete equation is presented below), and the second was a logistic regression refitted on the development worksheet with the same five predictors, and its coefficients, odds ratios, Wald 95% confidence intervals, and p values were kept for descriptive interpretation. The third was a penalized binomial logistic model using all ten candidate predictors. Its tuning family included ridge ($\alpha = 0$), elastic net ($\alpha = 0.5$), and lasso ($\alpha = 1$) options, with lambda chosen by inner cross-validation. Continuous variables stayed on their original scale for the conventional logistic model. Glmnet standardized predictors internally during penalized fitting. It returned coefficients on the original scale.

$$\eta_i = -6.963 + 0.310(\text{AcidBase}_i) + 1.228(\text{Diabetes}_i) + 1.308(\text{Hypertension}_i) + 3.428(\text{Coma}_i) + 0.113(\text{POSSUM}_i)$$

$$p_i = 1 / (1 + \exp(-\eta_i))$$

2.4. Data collection procedure

Record multiplicity and cross-sheet independence. An exact record signature joined the ten candidate predictors, ICDSC, and the binary delirium label. Excess exact duplicates within each worksheet were counted as the number of records minus the number of unique signatures. To assess cross-sheet independence, each designated-validation signature was compared with the set of development signatures. A validation record counted as overlapping when at least one complete 12-field signature match existed in the development worksheet. The definition is deliberately strict. It demands agreement on every analyzed predictor, the screening score, and the outcome label. Signatures that excluded the outcome were also examined in the reproducibility log, and they led to a similar conclusion.

2.5. Statistical analysis

Primary internal validation. Primary predictive performance for the refitted and penalized models was estimated by repeated stratified nested 10-fold cross-validation with 10 repeats. Within each repeat, every record served once in an outer test fold. The five-predictor model was estimated only on the matching outer training data, and for the penalized model, α and lambda were chosen by stratified five-fold cross-validation restricted to each outer training partition, and the selected model then produced probabilities for the untouched outer test fold. Predictor scaling, tuning, coefficient estimation, and prediction thus stayed separate from the outer assessment data. Each record gathered one out-of-fold prediction per repeat. Its 10 predictions were averaged for aggregate performance and calibration.

Performance measures and uncertainty. Discrimination was quantified by AUROC. The Brier score summarized the overall accuracy of probabilistic prediction, with lower values meaning smaller mean squared error. Calibration-in-the-large was estimated as the intercept of a logistic model with the logit of predicted probability entered as an offset, where zero is ideal. Calibration slope came from regressing the outcome on the logit of predicted probability. Ideal is a slope of one. A slope below one can indicate predictions that are too extreme. Ninety-five percent intervals were obtained from 2,000 stratified bootstrap samples that preserved the numbers of events and non-events. Calibration curves used deciles of averaged out-of-fold probabilities as a descriptive visualization.

Sensitivity analyses. The fixed published equation, the refitted five-predictor model, and the full-development penalized model were applied to the 100-record designated validation worksheet. These results were labeled sensitivity analyses, not external-validation estimates, because cross-sheet independence was tested and not assumed. We repeated the performance calculations in the subset of validation records without an exact 12-field match, to explore whether discrimination collapsed once exact matches were removed. The subset was not read as independent validation. The residual records came from the same archived structure and included only seven non-events.

Descriptive and coefficient analyses. Continuous variables were summarized by median and interquartile range, as the workbook held bounded or skewed scores and procedure durations. Binary variables were summarized by count and percentage. Adjusted odds ratios from the refitted five-predictor model describe the association structure of the development worksheet. They assert no causal effects. We performed no univariable screening, automated stepwise selection, post hoc threshold optimization, or significance-driven model modification. Model comparisons emphasized the size and uncertainty of predictive performance and not isolated p values.

Software and reproducibility. Analyses were run in R with base statistical functions and glmnet version 4.1-10. A fixed seed (20260901) governed fold assignment, tuning, and bootstrap resampling. The executable script imports the unmodified workbook, checks sample sizes, event counts, completeness, and field structure, and writes all predictions, coefficients, tuning results, performance tables, and figures. The analysis script and machine-readable outputs are included in the reproducibility materials.

2.6. Ethical approval

The analysis used a publicly archived, deidentified dataset released for reuse. The CMHC Research Center Ethics Committee Indonesia granted ethical approval for this secondary analysis (No. 2026/266). No participants were contacted, and no attempt was made to re-identify records.

3. Results

3.1. Cohort characteristics

All 400 records were complete for the analyzed fields. The development worksheet held 82 delirium events among 300 records (27.3%), and the designated validation worksheet held 29 events among 100 records (29.0%). Median POSSUM scores were 37.0 (interquartile range [IQR] 32.0–43.0) and 38.0 (31.8–45.0), respectively. Mechanical ventilation was recorded for 75.0% and 72.0%, hypertension for 39.3% and 36.0%, and diabetes for 16.7% and 17.0%. Coma history was more frequent in the designated validation worksheet than in the development worksheet (8.0% versus 4.7%). Stroke history was less frequent (4.0% versus 8.3%). Table 1 details the cohort characteristics.

Table 1. Characteristics of the archived development and designated validation worksheets.

Characteristic	Development (n = 300)	Designated validation (n = 100)
POSSUM score, median (IQR)	37.0 (32.0–43.0)	38.0 (31.8–45.0)
Operation duration, min, median (IQR)	200.0 (138.8–241.2)	195.0 (130.0–251.2)
Acid-base imbalance score, median (IQR)	0.0 (0.0–3.0)	0.5 (0.0–3.0)
ICDSC, median (IQR)	0.0 (0.0–2.0)	0.0 (0.0–1.2)
Male sex, n (%)	158 (52.7)	58 (58.0)
Alcohol-dependence history, n (%)	25 (8.3)	11 (11.0)
Coma history, n (%)	14 (4.7)	8 (8.0)
Mechanical ventilation, n (%)	225 (75.0)	72 (72.0)
Hypertension history, n (%)	118 (39.3)	36 (36.0)
Diabetes history, n (%)	50 (16.7)	17 (17.0)
Stroke history, n (%)	25 (8.3)	4 (4.0)
Delirium, n (%)	82 (27.3)	29 (29.0)

Values are median (IQR) or n (%). ICDSC, Intensive Care Delirium Screening Checklist; IQR, interquartile range; POSSUM, Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity.

3.2. Reproduction and model coefficients

The fixed published equation reproduced an apparent development-worksheet AUROC of 0.852 (95% bootstrap interval 0.799–0.897), matching the discrimination reported by Xing et al.[20] Its apparent Brier score was 0.125. Every predictor kept a positive coefficient in the refitted five-predictor model. Adjusted odds ratios were 1.37 per unit of acid-base imbalance score (95% CI 1.21–1.56), 3.55 for diabetes history (1.61–7.85), 3.75 for hypertension history (1.91–7.36), 31.34 for coma history (3.20–306.72), and 1.12 per POSSUM point (1.07–1.17). The wide interval for coma reflects limited precision. Only 14 development records had a coma history, so the point estimate is probably inflated by sparse-data bias. Figure 1 shows the complete coefficient profile.

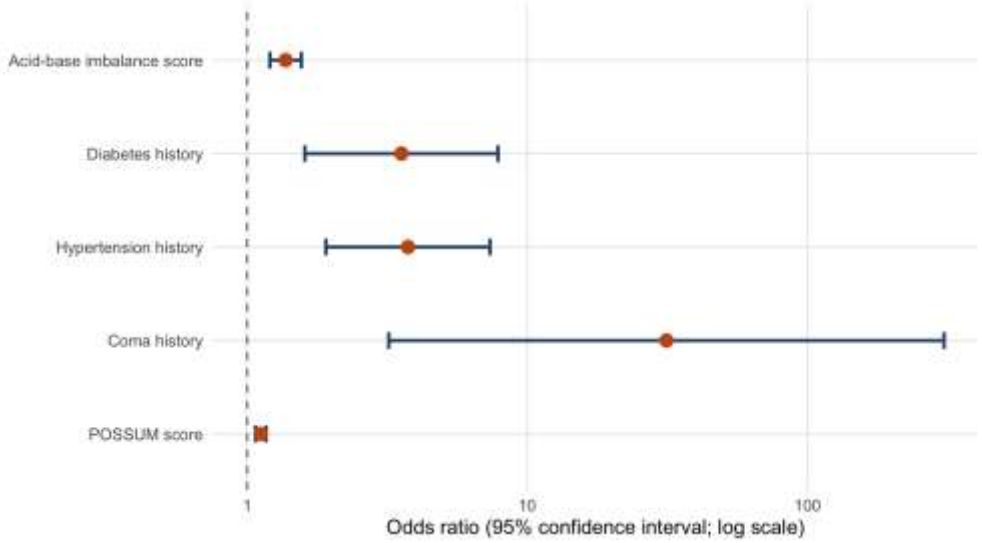


Figure 1. Adjusted odds ratios from the refitted five-predictor logistic model. The x-axis is logarithmic; bars show 95% Wald confidence intervals. Odds ratios are per unit for the acid-base imbalance score and the POSSUM score and for presence versus absence for the three histories.

3.3. Repeated nested cross-validation

The refitted five-predictor model reached a repeated nested cross-validated AUROC of 0.838 (95% bootstrap interval 0.782–0.886) and a Brier score of 0.132 (0.111–0.155). Its calibration intercept was −0.002 (−0.205 to 0.192) and its calibration slope was 0.751 (0.439–1.162). The ten-predictor penalized model reached an AUROC of 0.840 (0.785–0.888), a Brier score of 0.134 (0.115–0.154), a calibration intercept of 0.006 (−0.157 to 0.165), and a calibration slope of 1.061 (0.813–1.425). Table 2 details the primary and sensitivity performance estimates. Figures 2 and 3 show repeated nested cross-validated discrimination and calibration, respectively. The AUROC difference was 0.002 and the Brier difference was 0.002. The additional predictors therefore gave no evidence of material incremental performance.

Table 2. Predictive performance of the published, refitted, and penalized models.

Dataset, model	AUROC (95% interval)	Brier score	Calibration intercept	Calibration slope
Development, apparent, published equation	0.852 (0.799–0.897)	0.125	0.009	1.009
Nested CV, refit five-predictor	0.838 (0.782–0.886)	0.132	−0.002	0.751
Nested CV, penalized ten-predictor	0.840 (0.785–0.888)	0.134	0.006	1.061
Designated worksheet, published equation*	0.957 (0.888–0.997)	0.059	−0.434	1.663
Designated worksheet, refit five-predictor*	0.957 (0.888–0.997)	0.059	−0.445	1.648
Designated worksheet, penalized ten-predictor*	0.961 (0.891–0.999)	0.059	−0.364	1.807

Intervals are stratified percentile bootstrap intervals. *Sensitivity analyses only because 76% of designated-validation records exactly matched development signatures. AUROC, area under the receiver operating characteristic curve; CV, cross-validation; Brier score, mean squared error of the predicted probabilities (lower is better).

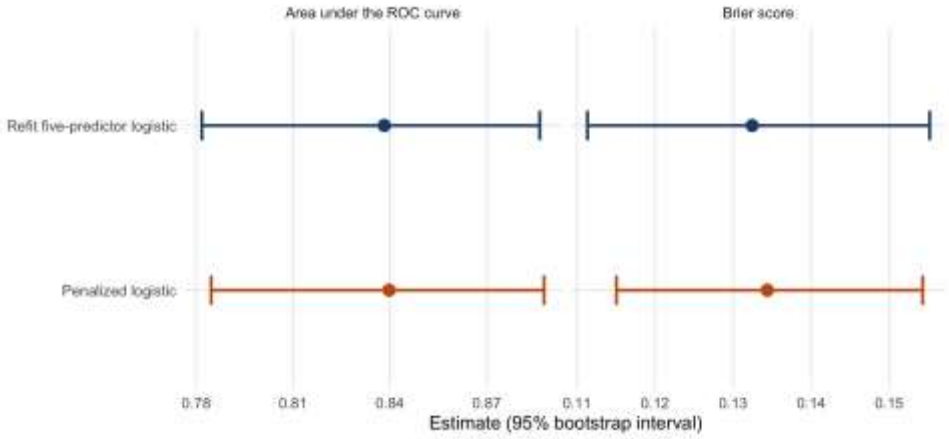


Figure 2. Repeated nested cross-validated discrimination and Brier score. Points show estimates; bars show 95% stratified bootstrap intervals. Blue, refitted five-predictor logistic model; orange, penalized logistic model; ROC, receiver operating characteristic.

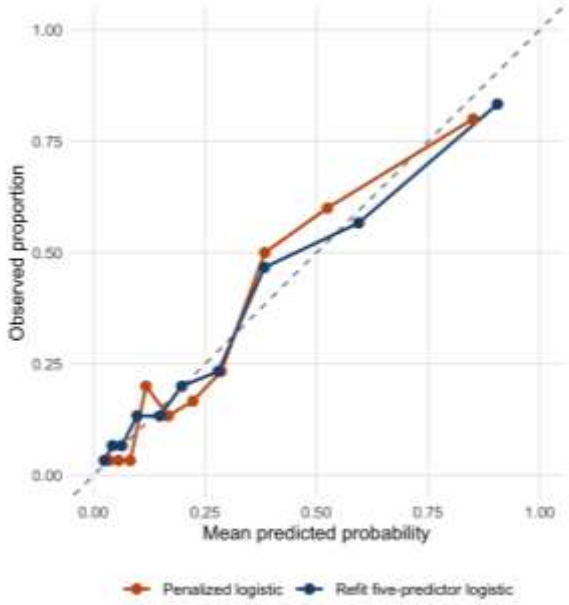


Figure 3. Calibration of averaged repeated out-of-fold predictions. Points represent groups defined by deciles of predicted probability; the dashed line indicates perfect calibration. Blue, refitted five-predictor logistic model; orange, penalized logistic model.

3.4. Penalization stability

Across the 100 outer training partitions, the inner procedure selected ridge penalization in 69, α = 0.5 elastic net in 16, and lasso in 15. Median selected lambda values were 0.0199 for ridge, 0.0155 for α = 0.5, and 0.0147 for lasso. When tuned on the complete development worksheet for the designated-worksheet sensitivity analysis, the selected model was ridge with lambda 0.0176. Because the selected α varied, the data supported shrinkage but not a uniquely stable sparsity pattern. Performance, not variable-selection frequency, was the primary basis of interpretation.

3.5. Cross-sheet overlap and outcome concordance

Seventy-six of the 100 designated-validation records had an exact match in the development worksheet across all ten candidate predictors, ICDSC score, and delirium label. Seventy-two unique validation signatures were represented in the development data, and multiset matching accounted for 75 records because some signatures occurred with different multiplicities. Within the development worksheet, 73 records were excess exact duplicates beyond the first occurrence of a signature; the corresponding count in the validation worksheet was four. The binary delirium label differed from ICDSC ≥4 in 15 development records and nine designated-validation records. Table 3 details record overlap, within-sheet multiplicity, and ICDSC–outcome concordance. These findings were kept as distinct provenance and outcome-concordance results. No labels were overwritten.

Table 3. Record overlap, multiplicity, and ICDSC–outcome concordance.

Dataset-provenance measure	Value
Development records	300
Designated-validation records	100
Validation records exactly matching development	76 (76.0%)
Validation records without an exact match	24 (24.0%)
Excess exact duplicates within development	73 (24.3%)
Excess exact duplicates within validation	4 (4.0%)
Development records discordant for ICDSC ≥4 vs label	15 (5.0%)
Validation records discordant for ICDSC ≥4 vs label	9 (9.0%)

Exact signatures included all ten candidate predictors, ICDSC, and the binary delirium label. Excess duplicates are records beyond the first occurrence of each signature. ICDSC, Intensive Care Delirium Screening Checklist.

3.6. Designated-sheet sensitivity analyses

Applied to all 100 designated-validation records, AUROCs were 0.957 (0.888–0.997) for the fixed equation, 0.957 (0.888–0.997) for the refitted five-predictor model, and 0.961 (0.891–0.999) for the penalized model. The Brier score was 0.059 for all three models. Calibration intercepts ranged from −0.445 to −0.364, which indicates that predicted risks were, on average, too high in this worksheet. Calibration slopes ranged from 1.648 to 1.807 and had wide intervals. Because 76% of these records overlapped the development worksheet exactly, the high discrimination estimates cannot be read as independent-validation performance.

3.7. Non-overlap subset

The 24 validation records without an exact development-worksheet match comprised 17 delirium events and seven non-events. AUROC was 0.929 (95% bootstrap interval 0.794–1.000) for both the fixed and refitted five-predictor models and 0.958 (0.849–1.000) for the penalized model. Brier scores ranged from 0.088 to 0.090. Calibration-slope intervals were extremely wide, extending above 70 for the five-predictor models and above 290 for the penalized model. Removing exact matches thus did not erase rank separation. The subset was small and event-heavy, however, and it was too limited for stable calibration or transportability inference.

4. Discussion

4.1. Principal findings

This reproducible reanalysis produced three principal findings. First, the published five-predictor equation was reproducible. It retained good discrimination under repeated nested cross-validation after refitting. Second, a broader penalized model did not materially improve AUROC or Brier score, which supports the informational efficiency of the parsimonious predictor set in these data. Third, 76% of records in the worksheet designated for validation matched development records exactly across predictors, ICDSC, and outcome. The high designated-worksheet AUROC should therefore be read as a sensitivity result from a non-independent partition and not as evidence of external transportability.

4.2. Comparison with the existing literature

The internally validated AUROC of 0.838 is clinically interesting because it lies within the broad range reported by recent delirium prediction studies. Models developed for neurosurgical, cardiac-surgical, orthopedic, non-cardiac surgical, and general ICU populations have reported moderate-to-high discrimination, depending on timing and cohort.[6,10–15,21–30] Numerical comparison across studies needs caution, though. A model that predicts delirium after ICU admission from postoperative severity variables answers a different question from a preoperative model,

a dynamic 12-hour forecast, or an electronic-health-record alert. Performance shifts with predictor availability, outcome surveillance, preventive treatment, and delirium prevalence.

That penalization produced almost no gain is understandable. The original five predictors capture acute physiologic burden (POSSUM and acid-base disturbance), chronic vulnerability (diabetes and hypertension), and impaired consciousness (coma). Sex, operation duration, alcohol history, mechanical ventilation, and stroke history may add clinically relevant context. In this 300-record sample, however, they offered limited incremental ranking information. Penalization reduced estimation instability and did not create new signal. Recent comparisons of conventional, machine-learning, deep-learning, biomarker, and dynamic models likewise show that greater complexity does not ensure more transferable performance.[4,28,30–34,37–46] Complexity is not usefulness.

Recent external-validation studies underline the need for setting-specific evaluation. Gong and colleagues developed dynamic ICU delirium models in a multicenter database and evaluated them in two separate single-center databases, an example of the stronger design that true source separation allows.[21] Prospective, geographic, and clinical-routine validations have also shown that promising development performance can weaken or become miscalibrated after transport.[11,14–16,24–27,32,37,38,40,47] These examples show why the distinction matters in practice. Model readiness depends on how performance changes across populations, not only on how well a model reproduces its development signal.

4.3. Interpretation and mechanism

Calibration. Calibration separated the two internally evaluated models more clearly than discrimination did. The slope of 0.751 for the five-predictor model suggests somewhat over-extreme predictions after refitting, although its interval included one. The penalized model's slope of 1.061 was closer to the ideal, again with uncertainty. Both calibration intercepts were near zero because the repeated out-of-fold predictions reflected the development prevalence. The results show why AUROC alone is insufficient. AUROC evaluates ranking, whereas calibration determines whether a stated 30% risk matches an observed frequency near 30%, and current reporting and appraisal guidance therefore asks for both domains to be reported transparently.[35,36] Preventive pathways often rely on risk thresholds. Systematic over- or underprediction could therefore change how many patients receive intensified surveillance or multicomponent prevention.

Designated validation worksheet. The designated validation worksheet appeared to perform exceptionally well, with AUROCs near 0.96 and Brier scores near 0.06. Read in isolation, these values might suggest a highly transportable model. The record-level comparison changes that reading. An evaluation partition cannot represent a new sample when most of its complete predictor-outcome signatures have already appeared in the development data. Exact matches do not show why the duplication occurred, and they may reflect copied rows, repeated exports, deterministic construction, or genuine patients who share all recorded values. Without stable participant and recruitment identifiers, however, independence cannot be established. The burden of evidence favors the conservative interpretation.

Non-overlap subset. The non-overlap subset does not solve the problem. Its high AUROC shows that rank ordering stayed favorable among the remaining records, but the subset contained only seven non-events. AUROC then rests on 119 event–non-event pairs, and bootstrap samples keep recycling a small number of patterns. Calibration slopes became extremely unstable, which shows how little information was available for absolute-risk evaluation. Deleting matching signatures after inspecting both worksheets also does not recreate a prospectively sampled external cohort. The subset is useful as a robustness check. It is not a replacement validation set.

ICDSC concordance. The ICDSC comparison adds a separate layer of interpretation. A threshold of four is widely used to indicate delirium in ICU screening, but the workbook's binary label did not coincide perfectly with ICDSC ≥4.[20] This discordance should not automatically be treated as an error. Delirium fluctuates, so a patient can meet diagnostic criteria at one assessment while a later or summarized ICDSC value falls below the threshold. Clinical adjudication, repeated measurements, and the timing of sedation or ventilation can also matter. Those details were unavailable, so the analysis kept the original label and reported threshold concordance openly. ICDSC was kept out of the predictors. It is part of outcome ascertainment and would have created direct leakage.

Coefficient pattern. The coefficient pattern makes clinical sense but should be read as predictive and not causal. Higher POSSUM and acid-base imbalance scores represent systemic physiologic stress. Diabetes and hypertension may mark vascular and metabolic vulnerability. Coma history had the largest odds ratio, in line with impaired neurologic reserve. Its confidence interval was wide because the exposure was uncommon. Similar constructs appear across recent ICU and postoperative models, which often include cognitive vulnerability, physiologic severity, procedural burden, mechanical ventilation, or neurologic status.[6,9,11,12,21,25,28,30] Agreement at the level of constructs adds credibility, but it does not show that the present coefficients will transfer unchanged to another hospital.

4.4. Implications for practice and policy

For clinical practice, the current model should be considered a candidate risk-stratification tool. Its five inputs are interpretable and likely to be available around ICU admission, and its internal discrimination justifies further study. It should not yet determine care by itself. A responsible next evaluation would define the prediction time in advance and ensure that every

predictor is measured before that time, collect repeated delirium assessments by trained staff, retain stable patient and encounter identifiers, and recruit from hospitals not represented in development. The fixed equation should be applied without refitting as the primary external test, and AUROC, calibration-in-the-large, calibration slope, calibration plot, Brier score, and decision-curve analysis should be reported. Any recalibration should be prespecified and reported as model updating, not as validation of the unchanged equation.

Sample-size planning for the next study should follow the expected event rate and the precision needed for calibration and discrimination, not a simple train–test percentage. Multiple centers and a temporal holdout would allow assessment of heterogeneity. Preventive interventions started after risk prediction should be documented. They can reduce observed delirium and alter apparent model performance. Fairness evaluation should examine clinically relevant sex, age, surgical, and comorbidity strata while avoiding unstable claims in small subgroups. A locked analysis protocol, a data dictionary, and a versioned model object would also guard against accidental leakage and analytical drift.

4.5. Strengths and limitations

The study has clear strengths. It uses individual-record public data, reproduces the published equation, confines tuning to inner folds, reports calibration as well as discrimination, and releases record-level out-of-fold predictions and tuning histories. Its overlap definition is transparent and deliberately stringent. Primary conclusions come from repeated nested cross-validation and not from the non-independent validation worksheet. The comparison between parsimonious and penalized approaches is clinically interpretable, and reporting followed TRIPOD+AI principles.[35,36]

Several limitations set the scope of inference. The public workbook contains 400 records and 111 events, and only 82 events were available for primary model development and internal validation. Participant identifiers, recruitment dates, surgical categories, hospital units, and repeated ICDSC assessments were unavailable, so duplication mechanisms and temporal ordering could not be resolved. The outcome definition could not be independently adjudicated. Bootstrap intervals reflect sampling variability conditional on the archived records and do not capture uncertainty from dataset construction. The repeated nested procedure evaluates internal reproducibility and cannot replace external validation. Finally, decision-curve analysis was not emphasized because non-independent or poorly calibrated evaluation data can give a misleading impression of clinical net benefit.

These limitations do not negate the predictive signal. They locate it precisely. The workbook supports good internal discrimination for a compact postoperative ICU model. The calibration slope of the penalized model was nearer one, but the intervals of both slopes include one, so the data do not show that penalization improves calibration. The workbook does not demonstrate geographic, temporal, or setting transportability. The distinction becomes more central as delirium prediction moves toward automated electronic-health-record systems. Transparent provenance and independent evaluation are not clerical refinements. They decide whether an apparently accurate model is truly being tested against new clinical information.

5. Conclusion

A five-predictor postoperative delirium model was reproducible and achieved good repeated nested cross-validated discrimination in the public development data. Expanding to a ten-predictor penalized model did not materially improve AUROC or Brier score. Its calibration slope was closer to one, though. The worksheet designated for validation overlapped extensively with development records, so its very high discrimination cannot support an independent-validation claim.

The next step is prospective, protocolized evaluation of the locked equation in a genuinely independent, sufficiently sized, multi-setting cohort, with repeated delirium assessment and complete calibration reporting. That evidence is still missing. Until it exists, the model should be regarded as a promising and interpretable candidate for research-based risk stratification. It is not a stand-alone clinical decision rule.

Declarations

Use of Generative Artificial Intelligence. None. The authors declare that no generative artificial intelligence tool was used in the conception, research, analysis, drafting, revision or preparation of this manuscript, and no AI tool is listed as an author.

Author Contributions. K.C.: conceptualization, methodology, software, formal analysis, data curation, visualization, writing – original draft. D.H.M.: validation, investigation, supervision, writing – review & editing. All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Funding. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest. The authors declare that they have no competing interests, financial or non-financial, in relation to this work.

Ethical Approval. Ethical approval was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2026/0266). The source publication describes the ethical procedures of the original data collection.

Informed Consent. Not applicable to the present secondary analysis of publicly archived, deidentified records. The source publication reports the consent and ethics procedures for original data collection.

Data Availability. The source data and data dictionary are openly available from the Dryad Digital Repository at doi:10.5061/dryad.gb5mkkwk3.

Acknowledgements. The authors acknowledge Xing and colleagues for making the source dataset publicly available.

References

1. Paixao L, Sun H, Hogan J, et al. ICU delirium burden predicts functional neurologic outcomes. PLoS One. 2021;16(12):e0259840. doi:10.1371/journal.pone.0259840
2. Al Huraizi AR, Al-Maqbali JS, Al Farsi RS, et al. Delirium and its association with short- and long-term health outcomes in medically admitted patients: a prospective study. J Clin Med. 2023;12(16):5346. doi:10.3390/jcm12165346
3. Tsui A, Searle SD, Bowden H, et al. The effect of baseline cognition and delirium on long-term cognitive impairment and mortality: a prospective population-based study. Lancet Healthy Longev. 2022;3(4):e232-e241. doi:10.1016/S2666-7568(22)00013-7
4. Racine AM, Tommet D, D'Aquila ML, et al. Machine learning to develop and internally validate a predictive model for post-operative delirium in a prospective, observational clinical cohort study of older surgical patients. J Gen Intern Med. 2021;36(2):265–73. doi:10.1007/s11606-020-06238-7
5. Pagali SR, Miller D, Fischer K, et al. Predicting delirium risk using an automated Mayo Delirium Prediction Tool: development and validation of a risk-stratification model. Mayo Clin Proc. 2021;96(5):1229–35. doi:10.1016/j.mayocp.2020.08.049
6. Huang HW, Zhang GB, Li HY, et al. Development of an early prediction model for postoperative delirium in neurosurgical patients admitted to the ICU after elective craniotomy (E-PREPOD-NS): A secondary analysis of a prospective cohort study. J Clin Neurosci. 2021;90:217–24. doi:10.1016/j.jocn.2021.06.004
7. Castro VM, Sacks CA, Perlis RH, et al. Development and external validation of a delirium prediction model for hospitalized patients with coronavirus disease 2019. J Acad Consult Liaison Psychiatry. 2021;62(3):298–308. doi:10.1016/j.jaclp.2020.12.005
8. Zhong X, Lin JY, Li L, et al. Derivation and validation of a novel comorbidity-based delirium risk index to predict postoperative delirium using national administrative healthcare database. Health Serv Res. 2021;56(1):154–65. doi:10.1111/1475-6773.13565
9. Hur S, Ko RE, Yoo J, et al. A machine learning-based algorithm for the prediction of intensive care unit delirium (PRIDE): retrospective study. JMIR Med Inform. 2021;9(7):e23401. doi:10.2196/23401
10. Bishara A, Chiu C, Whitlock EL, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. BMC Anesthesiol. 2022;22(1):8. doi:10.1186/s12871-021-01543-y
11. Chaiwat O, Chittawatanarat K, Mueankwan S, et al. Validation of a delirium predictive model in patients admitted to surgical intensive care units: a multicentre prospective observational cohort study. BMJ Open. 2022;12(6):e057890. doi:10.1136/bmjopen-2021-057890
12. Cai S, Cui H, Pan W, et al. Two-stage prediction model for postoperative delirium in patients in the intensive care unit after cardiac surgery. Eur J Cardiothorac Surg. 2022;63(1):ezac573. doi:10.1093/ejcts/ezac573
13. Menzenbach J, Kirfel A, Guttenthaler V, et al. PRe-Operative Prediction of postoperative DElirium by appropriate SCreening (PROPDESC) development and validation of a pragmatic POD risk screening score based on routine preoperative data. J Clin Anesth. 2022;78:110684. doi:10.1016/j.jclinane.2022.110684
14. Wong CK, van Munster BC, Hatseras A, et al. Head-to-head comparison of 14 prediction models for postoperative delirium in elderly non-ICU patients: an external validation study. BMJ Open. 2022;12(4):e054023. doi:10.1136/bmjopen-2021-054023
15. Kim SE, Ko RE, Na SJ, et al. External validation and comparison of two delirium prediction models in patients admitted to the cardiac intensive care unit. Front Cardiovasc Med. 2022;9:947149. doi:10.3389/fcvm.2022.947149
16. Jauk S, Veeranki SPK, Kramer D, et al. External validation of a machine learning based delirium prediction software in clinical routine. Stud Health Technol Inform. 2022;293:93–100. doi:10.3233/SHTI220353
17. Huang W, Wu Q, Zhang Y, et al. Development and validation of a nomogram to predict postoperative delirium in type B aortic dissection patients underwent thoracic endovascular aortic repair. Front Surg. 2022;9:986185. doi:10.3389/fsurg.2022.986185
18. Yang Y, Wang T, Guo H, et al. Development and validation of a nomogram for predicting postoperative delirium in patients with elderly hip fracture based on data collected on admission. Front Aging Neurosci. 2022;14:914002. doi:10.3389/fnagi.2022.914002
19. Jung JW, Hwang S, Ko S, et al. A machine-learning model to predict postoperative delirium following knee arthroplasty using electronic health records. BMC Psychiatry. 2022;22(1):436. doi:10.1186/s12888-022-04067-y
20. Xing H, Zhou W, Fan Y, et al. Development and validation of a postoperative delirium prediction model for patients admitted to an intensive care unit in China: a prospective study. BMJ Open. 2019;9(11):e030733. doi:10.1136/bmjopen-2019-030733
21. Gong KD, Lu R, Bergamaschi TS, et al. Predicting intensive care delirium with machine learning: model development and external validation. Anesthesiology. 2023;138(3):299–311. doi:10.1097/ALN.0000000000004478
22. Lee DY, Oh AR, Park J, et al. Machine learning-based prediction model for postoperative delirium in non-cardiac surgery. BMC Psychiatry. 2023;23(1):317. doi:10.1186/s12888-023-04768-y
23. Ren Y, Zhang Y, Zhan J, et al. Machine learning for prediction of delirium in patients with extensive burns after surgery. CNS Neurosci Ther. 2023;29(10):2986–97. doi:10.1111/cns.14237
24. Wueest AS, Berres M, Bettex DA, et al. Independent external validation of a preoperative prediction model for delirium after cardiac surgery: a prospective observational cohort study. J Cardiothorac Vasc Anesth. 2023;37(3):415–22. doi:10.1053/j.jvca.2022.11.038
25. Guo R, Zhang S, Yu S, et al. Inclusion of frailty improved performance of delirium prediction for elderly patients in the cardiac intensive care unit (D-FRAIL): A prospective derivation and external validation study. Int J Nurs Stud. 2023;147:104582. doi:10.1016/j.ijnurstu.2023.104582
26. Ali MIM, Kalkman GA, Wijers CHW, et al. External validity of an automated delirium prediction model (DEMO) and comparison to the manual VMS-questions: a retrospective cohort study. Int J Clin Pharm. 2023;45(5):1128–35. doi:10.1007/s11096-023-01641-6
27. Sweerts L, Dekkers PW, van der Wees PJ, et al. External validation of prediction models for surgical complications in people considering total hip or knee arthroplasty was successful for delirium but not for surgical site infection, postoperative bleeding, and nerve damage: a retrospective cohort study. J Pers Med. 2023;13(2):277. doi:10.3390/jpm13020277
28. Zhang Y, Hu J, Hua T, et al. Development of a machine learning-based prediction model for sepsis-associated delirium in the intensive care unit. Sci Rep. 2023;13(1):12697. doi:10.1038/s41598-023-38650-4

29. Guo Y, Ji H, Liu J, et al. Development and validation of a delirium risk prediction model for elderly patients undergoing elective orthopedic surgery. Neuropsychiatr Dis Treat. 2023;19:1641–54. doi:10.2147/NDT.S416854
30. Nagata C, Hata M, Miyazaki Y, et al. Development of postoperative delirium prediction models in patients undergoing cardiovascular surgery using machine learning algorithms. Sci Rep. 2023;13(1):21090. doi:10.1038/s41598-023-48418-5
31. Zhao X, Li J, Xie X, et al. Online interpretable dynamic prediction models for postoperative delirium after cardiac surgery under cardiopulmonary bypass developed based on machine learning algorithms: A retrospective cohort study. J Psychosom Res. 2024;176:111553. doi:10.1016/j.jpsychores.2023.111553
32. Jauk S, Kramer D, Sumerauer S, et al. Machine learning-based delirium prediction in surgical in-patients: a prospective validation study. JAMIA Open. 2024;7(3):ooae091. doi:10.1093/jamiaopen/ooae091
33. Yoshimura M, Shiramoto H, Koga M, et al. Development and validation of a machine learning model to predict postoperative delirium using a nationwide database: A retrospective, observational study. J Clin Anesth. 2024;96:111491. doi:10.1016/j.jclinane.2024.111491
34. Lin N, Lv M, Li S, et al. A nomogram for predicting postoperative delirium in pediatric patients following cardiopulmonary bypass: A prospective observational study. Intensive Crit Care Nurs. 2024;83:103717. doi:10.1016/j.iccn.2024.103717
35. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
36. Moons KGM, Damen JAA, Kaul T, 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. doi:10.1136/bmj-2024-082505
37. Kocar TD, Wolf P, Leinert C, et al. SURGE-ahead postoperative delirium prediction: external validation and open-source library. Eur Geriatr Med. 2025;16(3):851–9. doi:10.1007/s41999-025-01180-5
38. Reeve KA, Schmutz Gelsomino N, Venturini M, et al. Prospective external validation of the automated PIPRA multivariable prediction model for postoperative delirium on real-world data from a consecutive cohort of non-cardiac surgery inpatients. BMJ Health Care Inform. 2025;32(1):e101291. doi:10.1136/bmjhci-2024-101291
39. Ahn JH, Lee H, Gambus P, et al. Development of a deep learning-based prediction model for postoperative delirium using intraoperative electroencephalogram in adults. NPJ Digit Med. 2025;8(1):661. doi:10.1038/s41746-025-02033-y
40. Li H, Zang Q, Li Q, et al. Development of a machine learning-based predictive model for postoperative delirium in older adult intensive care unit patients: retrospective study. J Med Internet Res. 2025;27:e67258. doi:10.2196/67258
41. Wan H, Tian H, Wu C, et al. Development of a disease model for predicting postoperative delirium using combined blood biomarkers. Ann Clin Transl Neurol. 2025;12(5):976–85. doi:10.1002/acn3.70029
42. Zhang W, Tang N, Song J, et al. Development and validation of a machine learning-based risk prediction model for postoperative delirium in older patients with hip fracture. J Gerontol A Biol Sci Med Sci. 2025;80(12):glaf200. doi:10.1093/gerona/glaf200
43. Zhao Y, Jiang Y, Xu S, et al. Development and validation of a delirium prediction model for hospitalized older medical patients: a prospective cohort study. BMC Geriatr. 2025;25(1):904. doi:10.1186/s12877-025-06597-y
44. Park C, Han C, Jang SK, et al. Development and validation of a machine learning model for early prediction of delirium in intensive care units using continuous physiological data: retrospective study. J Med Internet Res. 2025;27:e59520. doi:10.2196/59520
45. Contreras M, Kapoor S, Zhang J, et al. A large language model for delirium prediction in the intensive care unit using structured electronic health records. Sci Rep. 2025;15(1):38890. doi:10.1038/s41598-025-22634-7
46. Back JS, Jin Y, Jin T, et al. A data-driven model to predict delirium based on dynamic patterns of clinical deterioration in critically ill patients. Nurs Crit Care. 2025;30(6):e70193. doi:10.1111/nicc.70193
47. Cowan SL, Preller J, Goudie RJB. Evaluation of the E-PRE-DELIRIC prediction model for ICU delirium: a retrospective validation in a UK general ICU. Crit Care. 2020;24(1):123. doi:10.1186/s13054-020-2838-2