Deep Learning–Based Recurrent Delirium Prediction in Critically Ill Patients

Filipe R. Lucini, PhD[1,2]
Henry T. Stelfox, MD, PhD[1,3]
Joon Lee, PhD[2,3,4,5,6]

OBJECTIVES: To predict impending delirium in ICU patients using recurrent deep learning.

DESIGN: Retrospective cohort study.

SETTING: Fifteen medical-surgical ICUs across Alberta, Canada, between January 1, 2014, and January 24, 2020.

PATIENTS: Forty-three thousand five hundred ten ICU admissions from 38,426 patients.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: We used ICU and administrative health data to train deep learning models to predict delirium episodes in the next two 12-hour periods (0–12 and 12–24 hr), starting at 24 hours after ICU admission, and to generate new predictions every 12 hours. We used a comprehensive set of 3,643 features, capturing patient history, early ICU admission information (first 24 hr), and the temporal dynamics of various clinical variables throughout the ICU admission. Our deep learning architecture consisted of a feature embedding, a recurrent, and a prediction module. Our best model based on gated recurrent units yielded a sensitivity of 0.810, a specificity of 0.848, a precision (positive predictive value) of 0.704, and an area under the receiver operating characteristic curve (AUROC) of 0.909 in the hold-out test set for the 0–12-hour prediction horizon. For the 12–24-hour prediction horizon, the same model achieved a sensitivity of 0.791, a specificity of 0.807, a precision of 0.637, and an AUROC of 0.895 in the test set.

CONCLUSIONS: Our delirium prediction model achieved strong performance by applying deep learning to a dataset that is at least one order of magnitude larger than those used in previous studies. Another novel aspect of our study is the temporal nature of our features and predictions. Our model enables accurate prediction of impending delirium in the ICU, which can potentially lead to early intervention, more efficient allocation of ICU resources, and improved patient outcomes.

KEY WORDS: deep learning; delirium; electronic health record; intensive care unit; machine learning; predictive modeling

Delirium is a neuropsychiatric syndrome characterized by the acute onset of fluctuating disturbances in consciousness and cognition, as well as alterations in motor behavior, emotionality, and sleep-wake cycle (1). The prevalence of delirium in critically ill patients has been reported to range from 20% to 83% (2) and is associated with longer hospital and ICU stays, higher rates and longer durations of mechanical ventilation, increased morbidity and mortality, and long-term cognitive impairment (3).

Accurate prediction of delirium may facilitate clinical decision-making and allocation of ICU resources (4). Current prediction models have limited predictive capacity. They generally only use some of the clinical information available

DOI: 10.1097/CCM.0000000000005789

KEY POINTS

Question: Is it possible to accurately predict impending delirium over the next 24 hours in ICU patients using deep learning and electronic health data?

Findings: This retrospective cohort study developed deep learning–based delirium prediction models using comprehensive ICU and administrative health data associated with over 43,000 adult ICU admissions from 15 ICUs in Alberta, Canada. The best model performed at an area under the receiver operating characteristic curve of approximately 0.9.

Meaning: Via early warnings, our delirium prediction model enables early intervention and efficient ICU resource allocation to improve the outcomes of patients with delirium.

(usually restricted to the start of the ICU stay) (5–8), focus on patients without previous delirium (5, 7), and provide no estimates of when delirium is likely to manifest (5–7). Many risk factors used in current models are only measured at a single time point although they change over time. Even when multiple measurements over time are used for recursive prediction updates, each risk factor is usually summarized as one value (e.g., the most abnormal value of the analyzed period [8]), neglecting potentially important temporal patterns.

The present study aimed to develop and validate a multivariable delirium prediction model that recurrently (every 12-hr period) predicts patient delirium in the ICU for the two subsequent 12-hour periods (0–12 and 12–24 hr) by using both data available at ICU admission (historical and admission data) and temporal data recorded throughout the ICU stay.

MATERIALS AND METHODS

Design, Setting, and Population

We conducted a retrospective multicenter cohort study. All adult patients (≥ 18 yr) admitted to 15 medical-surgical ICUs in Alberta, Canada, from January 1, 2014, to January 24, 2020, with an ICU stay longer than 24 hours and shorter than 30 days were included in this study. Patients were excluded if there were no registered assessments of delirium during their ICU stays or ICU admission data did not link with administrative databases.

This study was approved by the Conjoint Health Research Ethics Board at the University of Calgary (REB17-0389) and was reported in accordance with the TRIPOD statement (9) (**Supplemental Digital Content 1**, http://links.lww.com/CCM/H278). The need for informed consent was waived due to the large number of patients involved in the study and the retrospective nature of the research.

All computations were performed in Python 3.7 (Python Software Foundation, Wilmington, DE). The source code used in this study and our best trained model, which can be used as a pretrained model, are available on GitHub (https://github.com/data-intelligence-for-health-lab/delirium_prediction).

Data Sources

The primary data source was eCritical Tracer, an electronic medical record system used in all ICUs across Alberta (10). Additional linked administrative data sources included the Discharge Abstract Database (hospitalization data including up to 25 *International Classification of Diseases*, 9th Edition [ICD-9] codes), National Ambulatory Care Reporting System and Alberta Ambulatory Care Report System (emergency and ambulatory care data including up to 10 *International Classification of Diseases*, 10th Edition [ICD-10] codes), Physician Claims (outpatient data including up to 3 ICD-10 codes), and Vital Statistics (mortality data). The ICD-9 codes from the Discharge Abstract Database were translated to ICD-10 to adhere to the other data sources (11). All clinical and administrative data were extracted by Alberta Health Services and were not publicly available.

Feature Representation

Two feature sets (static and temporal) were used as input to our predictive models. The static feature set captured historical and early ICU admission (first 24 hr) data. The temporal feature set captured the temporal dynamics of variables between ICU admission and prediction time.

Historical data comprised information from up to 5 years prior to ICU admission and referred to patients' past diagnoses. It was organized into combinations of ICD-10 groups (n = 277) and time frames (n = 3, from

5 yr to 6 mo before ICU admission, from 6 mo to 48 hr prior ICU admission and from 48 hr before ICU admission to ICU admission), which resulted in 831 historical features. These historical features were binary: a one indicated that at least one documented diagnosis related to the patient, ICD-10 group, and time frame was present in data, whereas a zero represented the absence of documented diagnoses.

Early ICU admission data included demographics (age, sex, weight at ICU admission, and height), reason for admission, and clinical assessments completed in the first 24 hours of ICU admission (Sequential Organ Failure Assessment [SOFA], Acute Physiology And Chronic Health Evaluation [APACHE] II/III/IV, and their components). Sex and reason for ICU admission were one-hot encoded, resulting in a total of 27 early ICU admission features.

Temporal ICU data comprised 222 timestamped variables. Whenever two or more variables presented similar information, they were merged into one variable (e.g., "respiratory rate bedside monitor" and "manual respiratory rate" were merged as "respiratory rate"). The number of temporal variables was reduced to 192, which included medications and prescriptions (n = 127), laboratory test results (n = 30), Intensive Care Delirium Screening Checklist (ICDSC) components (n = 9), vital signs (n = 8), SOFA components (n = 7), Glasgow Coma Scale (GCS) components (n = 3), duration of mechanical ventilation (n = 2) and dialysis (n = 2), pain assessments (n = 2), Richmond Agitation and Sedation Scale (RASS) scores (n = 1), and urine volumes (n = 1). Each ICU stay was split into 12-hour periods starting from ICU admission, and records of the same variable occurring within each period were grouped together. Variables that presented meaningful cumulative values (e.g., duration of invasive mechanical ventilation) were summed, and the remaining variables were aggregated using 12 distribution and trend metrics to capture the temporal dynamics of the unevenly sampled variables in a standardized manner (minimum, first quartile, median, third quartile, maximum, mean, SD, interquartile range [IQR], minimum to maximum range, average difference between subsequent measurements, and difference between the last observed value and minimum and maximum values). In total, 2,435 temporal features (192 variables × 12 distribution and trend metrics + 131 cumulative values) were calculated for each ICU admission and period.

All features, static and temporal, were linearly normalized to the (0–1) range. Missing data were represented as zeros, and a binary missingness indicator was created for each early ICU admission and temporal feature to enable our models to distinguish between missing values and an actual value of zero. As a result, the static and temporal feature sets comprised 885 (831 historical features + 27 early ICU admission features + 27 missingness indicators) and 2,758 features (2,435 temporal features + 323 missingness indicators), respectively. A complete list of features is available in **Supplemental Digital Content 2** (http://links.lww.com/CCM/H278).

The delirium ground-truth label for each 12-hour period was created using ICDSC scores (12–14), a validated screening instrument based on eight dimensions of patients' condition: 1) altered level of consciousness; 2) inattention; 3) disorientation, 4) hallucination, delusion, or psychosis; 5) psychomotor agitation or retardation; 6) inappropriate mood or speech; 7) sleep-wake cycle disturbance; and 8) fluctuations. Abnormal status in four or more dimensions indicated the presence of delirium (12). The ICDSC is performed by trained bedside registered nurses once every 12-hour nursing shift and has been shown to have high internal consistency (15) and a good sensitivity and specificity (16). The Confusion Assessment Method for the ICU is not used in Alberta ICUs. In cases where ICDSC assessments were missing in a period, delirium status was propagated up to one subsequent period (12 hr).

Prediction Model Development and Evaluation

Our modeling approach was sequential over the ICU admission (**Fig. 1**); similar temporal early warning machine learning models have been developed for prediction of sepsis (17, 18) and hemodynamic interventions (19). At the end of each 12-hour period, starting from the end of the second period (i.e., after 24 hr of admission), static and temporal feature sets were input into a model to predict the probability of delirium in the two subsequent 12-hour periods. A probability greater than a prediction threshold indicated a positive delirium prediction. During training, only prediction time points where ground-truth labels for both prediction horizons were available were tested.

Predictions were made using three sequential modules: 1) embedding, 2) recurrent, and 3) prediction. The

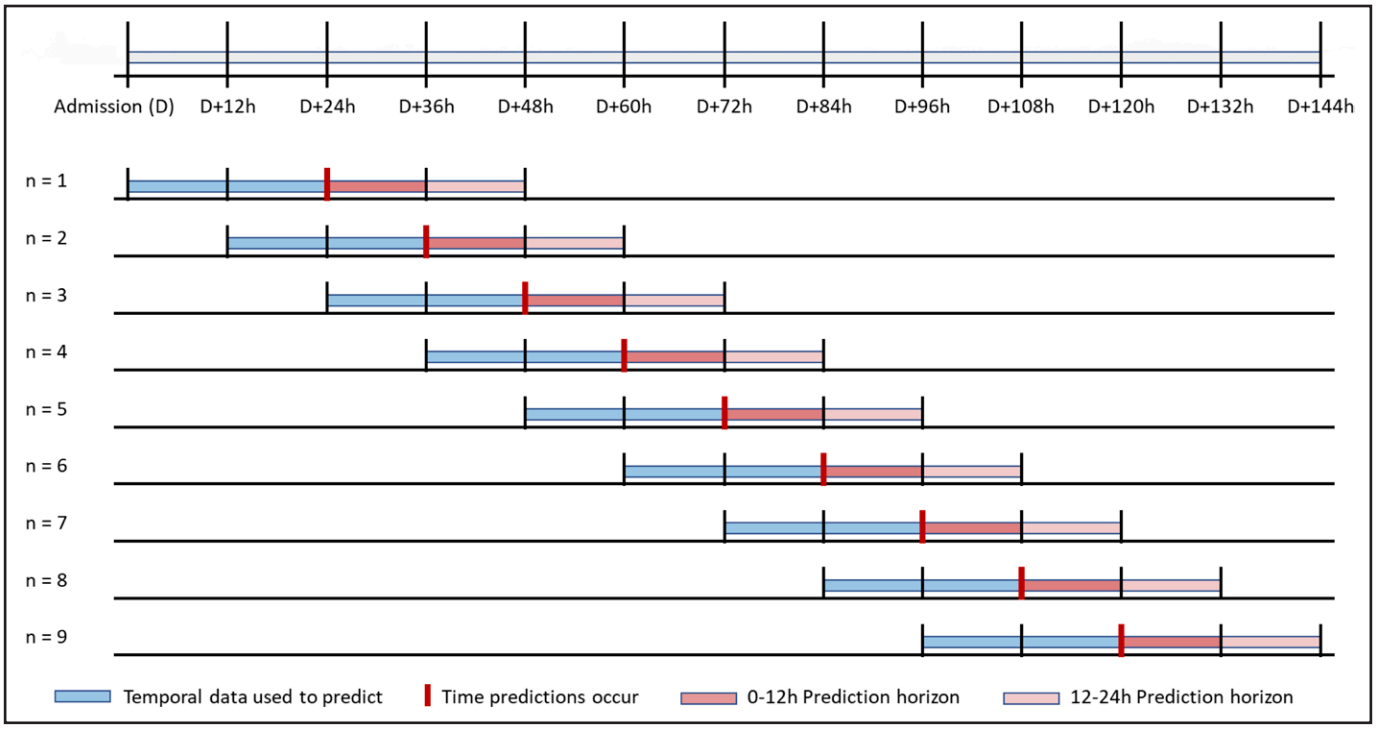

Figure 1. Sample sequential analysis over a 6-d ICU admission. *n* represents individual prediction instances. In addition to temporal data (presented in *blue*), models also used static data (historical and early ICU admission).

embedding module transformed the high-dimensional features into a parsimonious lower dimensional representation. The recurrent module modeled the temporal dynamics of the features and previous predictions. The prediction module generated predictions for the 0–12 and 12–24-hour prediction horizons. **Supplemental Digital Contents 3** and **4** (http://links.lww.com/CCM/H278) provide machine learning details and all evaluated model architectures and hyperparameter values, respectively.

The data were split by patients into training (80%), validation (5%), calibration (5%), and test (10%) sets. The data were used in this order: training, validation, calibration, and test. The training set was used to train the proposed models. The validation set was used to compare trained models and identify the best model architecture and hyperparameters. The 30 models (10 for each recurrent neural network architecture) with the best mean area under the receiver operating characteristics curve (AUROC) on the validation set were calibrated and compared using the calibration set. The model with the best mean AUROC with the prediction threshold resulting in the best F1-score in the calibration set was evaluated on the test set in terms of precision (positive predictive value), recall (sensitivity), specificity, F1-score, AUROC, and area under the precision-recall curve (AUPRC). To quantify the uncertainty of the performance of the best model, we calculated 95% CIs using the pivot bootstrap estimator (20) which resampled the test dataset 200 times with replacement. Because bootstrapping assumes independent events, we resampled at the patient level rather than at the admission or prediction instance level.

It was possible that our model may simply learn to predict that the delirium state will remain the same, since delirium state is not expected to change too frequently. Hence, to evaluate the performance of our best model (i.e., the model with the best architecture, hyperparameter values, and prediction threshold from the model development described above) in delirium state transitions as opposed to constant delirium state, we further analyzed the test set performance in the four possible scenarios for both prediction horizons (delirium state at prediction time-delirium state in the prediction horizon): 1) no delirium-no delirium, 2) no delirium-delirium, 3) delirium-no delirium, and 4) delirium-delirium.

Deep learning models are known to be highly complex "black box" models that are challenging to interpret (21). To mitigate this, we estimated feature importance by applying SHapley Additive exPlanations (SHAP) to our best model for each prediction horizon.

Sensitivity Analyses

We also conducted sensitivity analyses where new models were trained and evaluated based on data partitions split by year and site, as well as random data partitions without propagation of previous ICDSC assessments, using the best model architecture, prediction threshold, and hyperparameters selected based on the random data split described above. For stratification by site, data from 11 and 4 ICUs were used as training and test sets, respectively, with a random 15% of the admissions in the training set used for calibration (**Supplemental Digital Content 5**, http://links.lww.com/CCM/H278). For stratification by year, data from 2014 to 2018 and 2019 to 2020 were used as training and test sets, respectively, again with a random 15% of the admissions in the training set used for calibration (**Supplemental Digital Content 6**, http://links.lww.com/CCM/H278). The analysis without ICDSC propagation followed the same random data partitioning as the main analysis (**Supplemental Digital Content 7**, http://links.lww.com/CCM/H278).

RESULTS

A total of 48,672 unique patients (55,689 admissions) were admitted to the 15 ICUs during the study period. Of these 38,426 patients (79.0%) with 43,510 admissions (78.1%) satisfied the inclusion criteria and were included in the analysis. The most frequent reason for exclusion was an ICU length of stay (LOS) less than 1 day (6,576 patients [13.5%]; 7,908 admissions [14.2%]), followed by admissions with no registered delirium assessment (2,933 patients [6.0%]; 3,244 admissions [5.8%]) and an ICU LOS greater than 30 days (737 patients [1.5%]; 1,027 admissions [1.8%]) (**Fig. 2**).

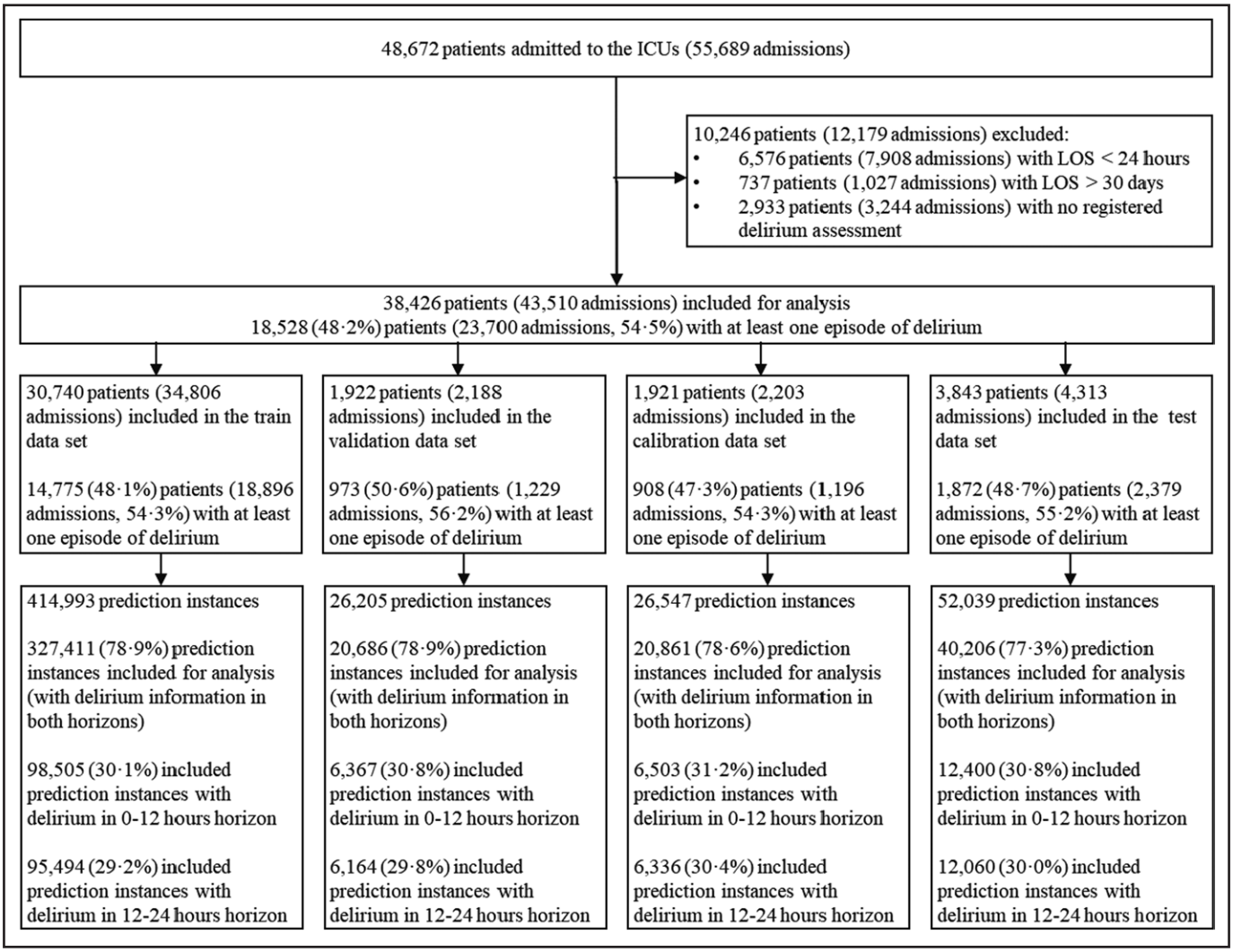

Figure 2. Flow diagram of patients in the study cohort. LOS = length of stay.

The median age of included patients was 59.9 years (IQR 46.3–70.3 yr), and most were male (57.9%). Patients had admission diagnoses that were medical (59.8%), surgical (21.4%), neuroscience (8.6%), trauma (6.2%), or not available (3.9%). A minority of patients were admitted after emergency (15.5%) or elective (7.4%) surgeries. The patients had moderate illness severity as measured by their median and IQR admission APACHE II (19 [13–25]), APACHE III (62 [44–82]), and SOFA (6 [3–9]) scores. The median ICU LOS was 4.1 days (IQR 2.4–7.6 d), and approximately half of the patients (54.5%) experienced at least one episode of delirium during their ICU stay. **Table 1** compares the characteristics of the patients with and without delirium.

The average number of ICDSC assessments per admission was 10.2. The median time between consecutive ICDSC assessments was 11 hours and 43 minutes, with an IQR of 5 hours.

The median and IQR of the percentages of the features with missing values per prediction instance in the train, validation, calibration, and test sets were 83.3% (81.0–85.7%), 83.1% (80.9–85.7%), 83.3% (81.0–85.8%), and 83.1% (80.9–85.7%), respectively. The number of prediction instances where one or both outcomes (delirium status in the next 0–12 hr and 12–24 hr) were missing in the train, validation, calibration, and test sets, after propagating the previous delirium status to the subsequent 12-hour period, were 87,582 (21.10%), 5,519 (21.06%), 5,686 (21.42%), and 11,833 (22.74%), respectively. The majority of the missing delirium values (79.5%, 79.3%, 77.5%, and 72.9%, respectively) were for the last two time periods

TABLE 1.
Patient Characteristics at the Admission Level

Variable	With Delirium (*N* = 23,700)	Without Delirium (*N* = 19,810)	*p*[a]
Patient age, yr, median (IQR)	59.7 (46.7–70.3)	60.0 (45.8–70.4)	0.467
Patient sex, male, *n* (%)	14,058 (59.3)	11,150 (56.3)	< 0.001
Admission type, *n* (%)			< 0.001
Nonsurgical	18,582 (78.4)	13,288 (67.1)	
Emergency surgery	3,558 (15.0)	3,167 (16.0)	
Elective surgery	1,042 (4.4)	2,190 (11.0)	
Not available	518 (2.2)	1,165 (5.9)	
Admission class, *n* (%)			< 0.001
Medical	14,606 (61.6)	11,430 (57.7)	
Surgical	4,210 (17.8)	5,122 (25.8)	
Neuroscience	2,437 (10.3)	1,311 (6.6)	
Trauma	1,915 (8.1)	767 (3.9)	
Unavailable	532 (2.2)	1,180 (6.0)	
APACHE II, median (IQR)	21 (16–27)	16 (11–21)	< 0.001
APACHE III/IV, median (IQR)	71 (52–91)	52 (37–70)	< 0.001
Sequential Organ Failure Assessment, median (IQR)	7 (5–10)	4 (2–7)	< 0.001
Invasive ventilation, *n* (%)	16,663 (70.3)	6,741 (34.0)	< 0.001
Noninvasive ventilation, *n* (%)	3,003 (12.7)	2,270 (11.5)	< 0.001
ICU length of stay, d, median (IQR)	5.8 (3.3–10.2)	2.9 (1.9–4.8)	< 0.001
ICU mortality, *n* (%)	2,030 (8.6)	618 (3.1)	< 0.001

APACHE = Acute Physiology And Chronic Health Evaluation, IQR = interquartile range.

[a]Wilcoxon rank-sum tests and Pearson's χ^2 tests were used for the continuous and categorical variables, respectively.

of patients' ICU stays as delirium status was not captured after discharge.

The best performance in terms of the calibration set AUROC was achieved by a gated recurrent unit-based model and a prediction threshold of 0.37 led to the best calibration set F1-score for this model. In the embedding module, it used two layers, 64 neurons on the static data input, 512 neurons on the temporal data input, hyperbolic tangent as the activation function, residual connection, and no dropout. The recurrent module used three layers with 128 neurons each and a dropout of 0.2. Calibration presented best overall results using isotonic regression, with Brier scores of 0.111 and 0.127 for the 0–12 and 12–24-hour prediction horizons in the test set, respectively (**Supplemental Digital Content 8**, http://links.lww.com/CCM/H278). Since the naive Brier score from a random predictor at a prevalence of 30% as in our dataset is 0.333, isotonic regression led to good calibration (the smaller the Brier score the better, and a Brier score of zero represents perfect calibration).

The AUROC and AUPRC, as well as the performances of the best model at various prediction thresholds, are presented in **Figure 3** and **Supplemental Digital Content 9** (http://links.lww.com/CCM/H278), respectively. The test set results show general (mean between both prediction horizons) precision, recall, specificity, F1-score, AUROC and AUPRC were 0.670, 0.800, 0.828, 0.729, 0.895, and 0.766, respectively (**Table 2**).

Supplemental Digital Content 10 (http://links.lww.com/CCM/H278) breaks down the best model

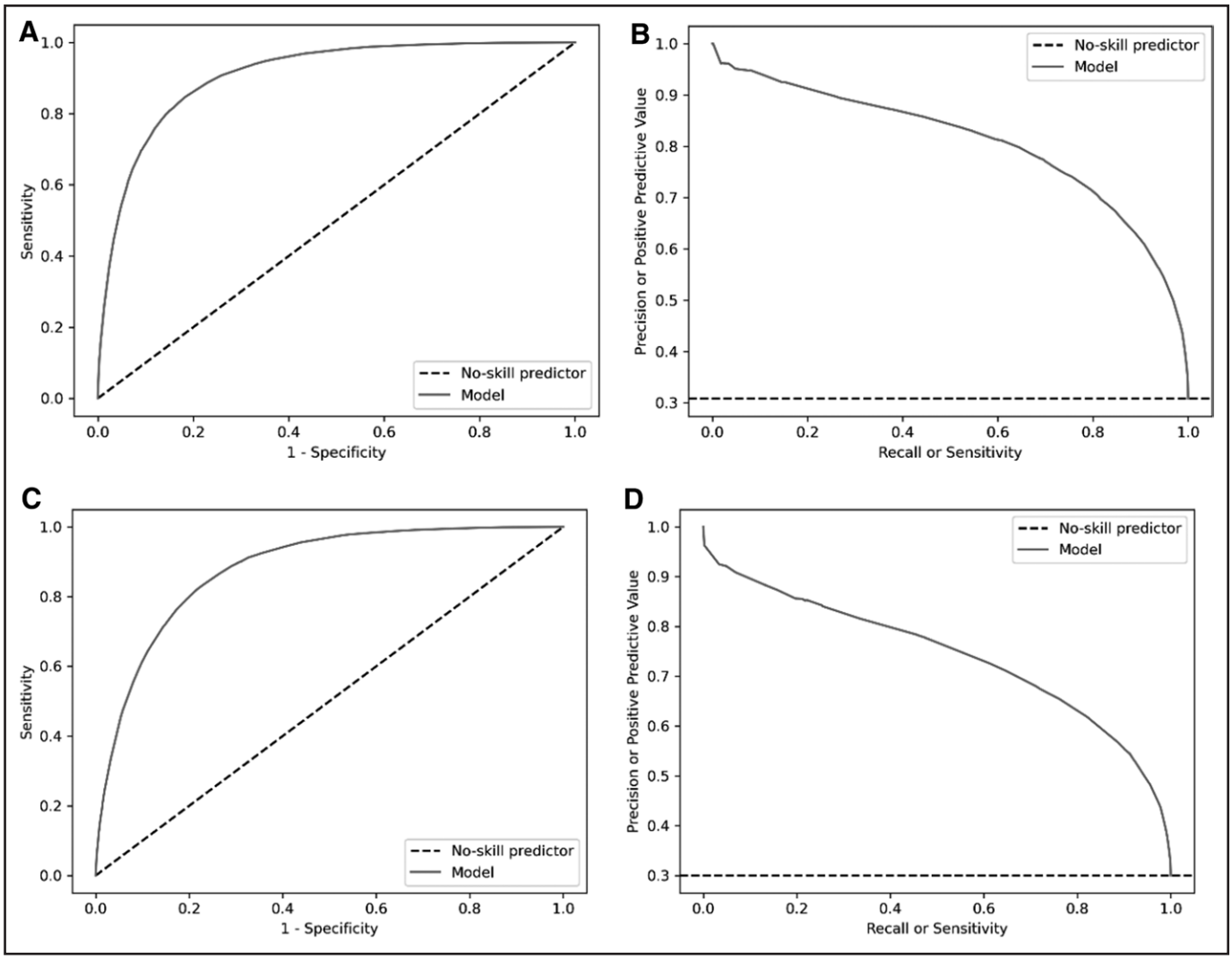

Figure 3. Receiver operating characteristic and precision-recall curves of the best model performance on the test set. **A**, Receiver operating characteristic curve for the 0–12 hr prediction horizon. **B**, Precision-recall curve for the 0–12 hr prediction horizon. **C**, Receiver operating characteristic curve for the 12–24 hr prediction horizon. **D**, Precision-recall curve for the 12–24 hr prediction horizon.

TABLE 2.
Best Model Prediction Performance on the Test Set

Prediction Horizon	Metric	Test Set Performance, Mean (95% CI)
0–12 hr	Precision	0.704 (0.702–0.706)
	Recall (sensitivity)	0.810 (0.809–0.812)
	Specificity	0.848 (0.847–0.850)
	F1-score	0.753 (0.752–0.755)
	AUROC	0.909 (0.908–0.910)
	AUPRC	0.786 (0.785–0.788)
12–24 hr	Precision	0.637 (0.635–0.639)
	Recall (sensitivity)	0.791 (0.789–0.793)
	Specificity	0.807 (0.805–0.808)
	F1-score	0.705 (0.704–0.707)
	AUROC	0.895 (0.894–0.896)
	AUPRC	0.745 (0.743–0.747)
General (mean between the two prediction horizons)	Precision	0.670 (0.669–0.672)
	Recall (sensitivity)	0.800 (0.799–0.802)
	Specificity	0.828 (0.826–0.829)
	F1-score	0.729 (0.728–0.731)
	AUROC	0.895 (0.894–0.896)
	AUPRC	0.766 (0.764–0.767)

AUROC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve.

performance on the test set in terms of the delirium states at prediction time and in the prediction horizon. In general, the performance was excellent when the delirium state was constant. The model performed reasonably well in predicting delirium onsets (sensitivities of 0.757 and 0.782 for the 0–12 and 12–24-hour prediction horizons, respectively, with no calibration) but failed to predict most recoveries from delirium.

In both prediction horizons, the most important features based on SHAP were related to ICDSC, GCS, RASS, and mechanical ventilation, with the maximum ICDSC score being the most important feature overall (**Supplemental Digital Content 11**, http://links.lww.com/CCM/H278). The most important static features for both prediction horizons were medical history represented by ICD codes. **Supplemental Digital Contents 12** and **13** (http://links.lww.com/CCM/H278) present the five most important features per feature category for the 0–12 and 12–24-hour horizons, respectively.

The complete sensitivity analysis results, including the training and test set statistics, as well as prediction and calibration performances, are reported in Supplemental Digital Contents 5–7 (http://links.lww.com/CCM/H278) for stratification by site and year, as well as without ICDSC propagation. Overall, the sensitivity analyses resulted in similar prediction performances. The mean AUROCs from the data splits by site and year were 0.910 and 0.902 for the 0–12-hour horizon and 0.883 and 0.877 for the 12–24-hour horizon, respectively. The dataset without ICDSC propagation resulted in the mean AUROCs of 0.883 and 0.860 for the 0–12 and 12–24-hour horizons, respectively. Furthermore, all sensitivity analyses yielded similar performance results to those from the main analysis in terms of changing and constant delirium states.

DISCUSSION

Novelty of This Study

Our study is novel in several ways. First, to the best of our knowledge, our study is the first major application of state-of-the-art deep learning methods to ICU

delirium prediction. Second, our model was developed using data from 43,510 ICU admissions from 38,426 patients, which is considerably larger than those used in previous studies (1,613 [7], 1,824 [6], 2,914 [5], and 560 patients [8]). Third, our model makes new predictions every 12 hours (as opposed to just one prediction at a specific time point) for specific prediction horizons (as opposed to any time in the rest of the ICU admission), which is novel in the context of delirium prediction in the ICU. Fourth, the feature set used in this study is much more comprehensive than those used in most previous studies (a total of 3,643 features vs 7–10 in previous studies [5, 7, 8]) and includes both historical and temporal ICU data. Some studies, such as Moon et al (22), have used feature sets of similar magnitude to this study, but they are rare. Fifth, our model was able to predict delirium onsets reasonably well, although it was unable to predict recovery from delirium. Clinically, it is much more useful to predict delirium onsets than recoveries. Our model's performance was excellent when the delirium state was constant. It should be noted that in general, our model overestimated probabilities of delirium and calibration scaled them down, as seen in Supplemental Digital Contents 5–8 (http://links.lww.com/CCM/H278). As a result, when the same threshold of 0.37 as the main analysis was used in the delirium state transition analysis, the prediction probabilities without calibration resulted in more positive predictions than those with calibration. This is why isotonic regression and Platt scaling underperformed no calibration in the no delirium-delirium and delirium-delirium transitions, whereas improving performance in the no delirium-no delirium and delirium-no delirium transitions (Supplemental Digital Contents 5–7 and 10, http://links.lww.com/CCM/H278). This is a trade-off, however, and performance in any of the four delirium state transition scenarios can be improved by adjusting the prediction threshold, at the expense of performances in other scenarios. Finally, we have made our best performing model publicly available, so that other researchers can use it as is or as a pretrained model at other institutions.

Comparisons With Related Works

There are several ICU delirium prediction models in the literature. The PRE-DELIRIC (7) model was developed and originally validated using data of 1,613 consecutive adult ICU patients and yielded an AUROC of 0.87 (95% CI 0.85–0.89). This model was later recalibrated in a multinational observational study (6) of 1,824 patients from eight ICUs in six countries resulting in an AUROC of 0.77. The Early PRE-DELIRIC (5) model predicts ICU delirium using only information available at ICU admission. It is based on 2,914 consecutive ICU patients and resulted in an AUROC of 0.75. More recently, the DYNAMIC-ICU (8) model was developed based on 560 consecutive adult patients admitted to four ICUs and achieved an AUROC of 0.900 (95% CI 0.858–0.941) in the validation cohort. The performances reported in the present study are similar (8) or better (5–7) than those from these previous studies.

Clinical Implications

Advance warnings about impending delirium could meaningfully inform clinical decision-making such as the efficient allocation of limited resources (23). For example, nursing assignments to patients, which are based on care needs (ICU ratios commonly range from 1:1 to 1:4), could be informed in advance as patients with delirium require greater personal attention. Similarly, prevention (e.g., noise reduction, minimizing interventions at night) and management (e.g., coordinating family-caregiver visitation times) strategies could be planned based on evolving delirium risk. Finally, given that delirium is often an early manifestation of new organ dysfunction and clinical deterioration (24–26) (e.g., resulting from hospital acquired infection), effective prediction models could both inform surveillance strategies and provide a mechanism for early detection and management.

Limitations and Future Work

The results of this study need to be interpreted within the context of its limitations. First, our models used an extensive list of features as input. Some of these variables may not be available in other clinical settings, which limits the generalizability of our model. It is unknown how the models would function with fewer variables. Future work may develop a more widely applicable model using only ICDSC, GCS, and RASS which were the most important predictors. Second, our data preprocessing and

predictive modeling methods were carefully selected but not exhaustive. Other approaches could have led to better prediction performance. Third, we used patient data from a single population (Alberta, Canada), and future work could focus on validation in other populations. Last, our model was developed based on retrospective and hence should not be used for causal inference. Feature importance is solely based on association rather than causation.

CONCLUSIONS

Our delirium prediction model showed promising performance and can meaningfully inform clinical decision-making, potentially leading to an optimal use of ICU resources and improved patient outcomes.

1 *Department of Critical Care Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.*

2 *Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.*

3 *O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.*

4 *Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.*

5 *Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.*

6 *Department of Preventive Medicine, School of Medicine, Kyung Hee University, Seoul, South Korea.*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Supported, in part, by University of Calgary Eyes High Postdoctoral Scholar Program (Filipe Lucini recipient).

Dr. Lee received funding from the University of Calgary; he received support for article research from the University of Calgary. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: joonwu.lee@ucalgary.ca

REFERENCES

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. Fifth Edition. Arlington, VA, American Psychiatric Association, 2013
2. Cavallazzi R, Saad M, Marik PE: Delirium in the ICU: An overview. *Ann Intensive Care* 2012; 2:1–11
3. Salluh JIF, Wang H, Schneider EB, et al: Outcome of delirium in critically ill patients: Systematic review and meta-analysis. *BMJ* 2015; 350:h25381–hh2538
4. Wassenaar A, Schoonhoven L, Devlin JW, et al: Delirium prediction in the intensive care unit: Comparison of two delirium prediction models. *Crit Care* 2018; 22:1–9
5. Wassenaar A, van den Boogaard M, van Achterberg T, et al: Multinational development and validation of an early prediction model for delirium in ICU patients. *Intensive Care Med* 2015; 41:1048–1056
6. Van Den Boogaard M, Schoonhoven L, Maseda E, et al: Recalibration of the delirium prediction model for ICU patients (PRE-DELIRIC): A multinational observational study. *Intensive Care Med* 2014; 40:361–369
7. Van Den Boogaard M, Pickkers P, Slooter AJC, et al: Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: Observational multicentre study. *BMJ* 2012; 344:17
8. Fan H, Ji M, Huang J, et al: Development and validation of a dynamic delirium prediction rule in patients admitted to the intensive care units (DYNAMIC-ICU): A prospective cohort study. *Int J Nurs Stud* 2019; 93:64–73
9. Collins GS, Reitsma JB, Altman DG, et al: Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Br J Surg* 2015; 102:148–158
10. Brundin-Mather R, Soo A, Zuege DJ, et al: Secondary EMR data for quality improvement and research: A comparison of manual and electronic data collection from an integrated critical care electronic medical record system. *J Crit Care* 2018; 47:295–301
11. National Bureau of Economic Research: ICD-9-CM to and From ICD-10-CM and ICD-10-PCS Crosswalk or General Equivalence Mappings. 2010. Available at: https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings. Accessed May 1, 2021
12. Bergeron N, Dubois MJ, Dumont M, et al: Intensive care delirium screening checklist: Evaluation of a new screening tool. *Intensive Care Med* 2001; 27:859–864
13. Krewulak KD, Rosgen BK, Ely EW, et al: The CAM-ICU-7 and ICDSC as measures of delirium severity in critically ill adult patients. *PLoS One* 2020; 15:e02423781–e02423715
14. Van Eijk MMJ, Van Marum RJ, Klijn IAM, et al: Comparison of delirium assessment tools in a mixed intensive care unit. *Crit Care Med* 2009; 37:1881–1885
15. Detroyer E, Timmermans A, Segers D, et al: Psychometric properties of the intensive care delirium screening checklist when used by bedside nurses in clinical practice: A prospective descriptive study. *BMC Nurs* 2020; 19:1–10
16. Gusmao-Flores D, Figueira Salluh JI, Chalhub RT, et al: The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: A systematic review and meta-analysis of clinical studies. *Crit Care* 2012; 16:1–10
17. Mohammed A, Van Wyk F, Chinthala LK, et al: Temporal differential expression of physiomarkers predicts sepsis in critically ill adults. *Shock* 2021; 56:58–64
18. van Wyk F, Khojandi A, Mohammed A, et al: A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. *Int J Med Inform* 2019; 122:55–62

19. Rahman A, Chang Y, Dong J, et al: Early prediction of hemodynamic interventions in the intensive care unit using machine learning. *Crit Care* 2021; 25:1–9
20. Efron B, Tibshirani RJ: An Introduction to the Bootstrap. Boca Raton, FL, CRC Press, 1994
21. Koh PW, Liang P: Understanding Black-Box Predictions Via Influence Functions. *In:* Proceedings of the 34th International Conference on Machine Learning. Sydney, New South Wales, Australia, August 6-11, 2017, pp 1885–1894
22. Moon KJ, Jin Y, Jin T, et al: Development and validation of an automated delirium risk assessment system (Auto-DelRAS) implemented in the electronic health record system. *Int J Nurs Stud* 2018; 77:46–53
23. Pun BT, Badenes R, Heras La Calle G, et al: Prevalence and risk factors for delirium in critically ill patients with COVID-19 (COVID-D): A multicentre cohort study. *Lancet Respir Med* 2021; 9:239–250
24. Poloni TE, Carlos AF, Cairati M, et al: Prevalence and prognostic value of delirium as the initial presentation of COVID-19 in the elderly with dementia: An Italian retrospective study. *EClinicalMedicine* 2020; 26100490
25. Hsieh SJ, Madahar P, Hope AA, et al: Clinical deterioration in older adults with delirium during early hospitalisation: A prospective cohort study [Internet]. *BMJ Open* 2015; 5:e007496
26. Atterton B, Paulino MC, Povoa P, et al: Sepsis associated delirium. *Medicina (Kaunas)* 2020; 56:240