DELIRIUM PREDICTION IN THE ICU - DESIGNING A SCREENING TOOL FOR PREVENTIVE INTERVENTIONS

Anirban Bhattacharyya, MD[1#*], Seyedmostafa Sheikhalishahi, MSc[2,3#], Heather Torbic, PharmD[4], Wesley Yeung, MBBS[5], Tiffany Wang, RN[6], Jennifer Birst, OTR/L[7], Abhijit Duggal, MD[8], Leo Anthony Celi, MD[6,9,10], Venet Osmani, PhD[2].

1 Critical Care Services, Mayo Clinic, Jacksonville, FL

2 Fondazione Bruno Kessler Research Institute, Trento, Italy

3 University of Trento, Trento, Italy

4 Department of Pharmacy, Cleveland Clinic, Cleveland OH

5 National University of Singapore, Singapore

6 Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA

7 Physical and Occupational Therapy, Mayo Clinic, Jacksonville, FL

8 Respiratory Institute, Cleveland Clinic, Cleveland, OH

9 Institute of Medical Engineering and Science, Massachusetts Institute of Technology, Boston, MA

10 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

# Joint First Author

* Corresponding author

Confidential: For Review Only

Address for reprints and correspondence:

4500 San Pablo Rd S,

JA_MA_01_CC

Jacksonville FL 32224

Phone: 904 786 3331

Email: bhattacharyya.anirban@mayo.edu

We do not anticipate reprints being ordered.

No financial support including institutional departmental support was used in this research or for preparation of manuscript.

Key words: delirium, clinical decision support, machine learning, artificial intelligence, nursing assessment, predictive modeling

Word count: 3042

Confidential: For Review Only

Confidential: For Review Only

ABSTRACT

Introduction

Delirium occurrence is common and preventive strategies are resource intensive. Screening tools can prioritize patients at risk. Using machine learning we can capture time and treatment effects that pose a challenge to delirium prediction. We aim to develop a delirium prediction model that can be used as a screening tool.

Methods

From the eICU Collaborative Research Database (eICU-CRD) and the Medical Information Mart for Intensive Care version III (MIMIC-III) database, patients with one or more Confusion Assessment Method -Intensive Care Unit (CAM-ICU) values and ICU length of stay greater than 24-hours were included in our study. We validated our model using 21 quantitative clinical parameters and assessed performance across a range of observation and prediction windows, using different thresholds and applied interpretation techniques.

Results

We evaluated 16546 and 6294 patients from eICU-CRD and MIMIC-III databases, respectively. Performance was best in BiLSTM models where, precision and recall changed from 37.52% (95% CI 36.00%-39.05%) to 17.45 (95%CI 15.83%-19.08%) and 86.1% (95% CI 82.49%-89.71%) to 75.58% (95% CI 68.33%-82.83%) respectively as prediction window increased from 12 to 96 hours. After optimizing for higher recall, precision and recall changed from 26.96% (95% CI 24.99%-28.94%) to 11.34% (95% CI 10.71%-11.98%) and 93.73% (95% CI 93.1%-94.37%) to 92.57% (95% CI 88.19%-96.95%) respectively. Comparable results were obtained in the MIMIC-III cohort.

Conclusions

Our model performed comparably to contemporary models using fewer variables. Using techniques like sliding windows, modification of threshold to augment recall and feature ranking for interpretability, we addressed shortcomings of current models.

Confidential: For Review Only

INTRODUCTION

The diagnosis of delirium is common in critically ill patients and depending on the patient population its incidence can be up to 80%.[1] Delirium leads to increased hospital length of stay and need for prolonged institutionalization for critically ill patients.[2–4] Delirium drives up healthcare costs, and its impact often persists beyond the intensive care unit (ICU) including risk for functional decline in daily living activities, and long-term cognitive impairment.[5–9]

Treatment and prevention of delirium is dependent on identifying the complex interplay of multiple triggers in the ICU.[10] A multimodal strategy of evidence-based best-practice recommendations aimed at coordinating multidisciplinary care to reduce delirium risk and expedite ICU discharge commonly referred to as the ABCDEF bundle is effective in both preventing and treating delirium.[11,12] Unfortunately this bundle of interventions requires education of caregivers, coordination between a multidisciplinary team, is labor and resource intensive, and therefore not consistently implemented across all ICU patients and all health care settings.[11,13] A screening tool to prioritize ABCDEF implementation to those who are most vulnerable can be an invaluable tool to maximize the benefit of the resource-intensive preventive measures.

Current assessment tools, such as the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU), only diagnoses delirium after its onset.[14] Administering CAM-ICU requires specialized training. Although each hospital has its own protocol for delirium, because of its time-consuming nature CAM-ICU is infrequently done compared to other vital signs and diagnosis can be delayed. Although certain patient

Confidential: For Review Only

characteristics, such as age, illness severity, and certain medications, are considered high risk for development of delirium or while elevations in inflammatory biomarkers possibly associated with severe disease, these risk factors have been inconsistent in their ability to predict the onset of delirium.[15–17]

Previous prediction models trained on small patient cohorts lacked adequate power to capture the complex relationships between delirium and the time-varying predictor variables.[18,19] To improve accuracy larger administrative datasets were used to develop prediction models, using several hundred variables, but these models lack interpretability, and are almost impossible to adopt in day-to-day practice.[20] Additionally, most of these models are not specific to the critically ill population and cannot be extrapolated to the ICU.[19]

We propose to build an ABCDEF screening tool by developing and fine tuning a delirium prediction model that requires fewer variables than existing models and can predict the risk of delirium in a continuous fashion using a sliding window. Using both conventional machine learning methods and deep learning algorithms we will evaluate performance of our model across various observation and prediction windows to address the issues of variability across time and treatment effects. In addition, we will rank the independent variables in order of their predictive importance to help with interpretability. These attributes should help pave the way for implementation of a screening tool to help caregivers at the bedside.

Confidential: For Review Only

METHODS

Ethical Review

The analysis using the eICU Collaborative Research Database (eICU-CRD) is exempt from institutional review board approval due to the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA) (Health Insurance Portability and Accountability Act Certification no. 1031219-2). The data in the Medical Information Mart for Intensive Care version III (MIMIC-III) is de-identified, and the institutional review boards of the Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) both approved the use of the database for research.

Confidential: For Review Only

Study Population

The eICU-CRD is a freely available multicenter database comprising 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 in over 200 hospitals located throughout the US.[21] The MIMIC-III database is an open-access single-center ICU database including 53,423 distinct hospital admissions for 46,476 unique patients admitted from 2001 to 2012.[22] Both datasets comprise data on patient demographics, vitals, clinical flowsheets, laboratory values, medications, interventions, and outcomes.

Any patient admitted to the ICU for 24 hours or more and with at least one CAM-ICU assessment was included in our study population (eFigure-1).

Confidential: For Review Only

Delirium Assessment

Observation window refers to the period where patient data are collected, and the model is derived. Prediction window refers to the period from when the observation window ends up to the onset of the outcome, CAM-ICU positive in our case. We observed patients from 0 to 12 hours, 0 to 24 hours and 0 to 48 hours. We predicted the incidence of Delirium for the next 12 hours, 24 hours, 48 hours, 72 hours and 96 hours. The diagnosis of delirium was made when at least one CAM-ICU value was positive.[14] In instances with multiple CAM-ICU assessments, onset of delirium was determined from the time of the first positive CAM-ICU.

Variable Selection

The rationale for selection of independent variables was based on their ability to predict delirium in prior literature, availability in our databases, ease of extracting and monitoring in a real-time environment. We identified 21 categorical or numerical variables classified into demographic data, vital signs, laboratory values, and vasopressor dose that fulfilled above criteria.[23–34] We also calculated daily sequential organ failure assessment (SOFA) scores to provide overall patient status. Since admission diagnoses or past medical history were not consistently available in the applied datasets, we excluded them. Downstream variables such as outcomes would not be available in real-time and similarly excluded. Initiation of delirium therapies like antipsychotic drugs could be a reaction to onset of delirium, and hence excluded to avoid confounding. Table-1 lists all the variables used.

Table 1: Variables included in the prediction models

Demographic Data
age, gender, height, weight
Vital Signs
oxygen saturation (SpO_2), heart rate (HR), temperature
Other Measurements
sofa, sofa without GCS, Ventilation
Laboratory Measurements
white blood cell count (WBC), sodium (Na), blood urea nitrogen (BUN), glucose, hemoglobin, platelets, potassium, chloride, bicarbonate, creatinine
Medications as continuous drips
Dopamine, epinephrine, norepinephrine, phenylephrine (all calculated as norepinephrine equivalent)

Confidential: For Review Only

Data Preprocessing

All variables were aggregated into hourly intervals, where the last known value was used as a candidate for that interval. In cases where the last value for each variable is not measured in the interval, the representative of that interval was computed by averaging the available measurements in the interval. Missing values that were collected hourly like vital signs were imputed by forward and then if needed backward imputation. Categorical variables were converted into a vector to capture the semantics of each category at the model derivation phase. Specifically, we reshaped the data that was fed in 3 dimensions for BILSTM to 2-dimension input for LR and RF in order to have the same input for each model and to provide a fair comparison for each model. For all continuous variables, we utilized the recorded value in the database without any adaptation. Heatmap further details the set of variables, including linear correlations between each variable.

Confidential: For Review Only

Model Derivation and Validation

We evaluated the results based on 5-fold stratified cross-validation. This method divides data into 5 folds where in each fold there exist 5 subsets of data. Four of these subsets are considered a derivation set and one subset is considered as a mutually exclusive validation set. Typically, metrics calculated based on the k-fold stratified cross-validation can effectively assess overfitting and has lower variance.[35]

We used 3 sets of algorithms to evaluate delirium prediction, namely Logistic Regression (LR), Random Forest (RF), and Bidirectional Long Short-Term Memory

(BiLSTM). BiLSTM represents an evolution from recurrent neural network-based LSTM, and with a backward input, preserves information from both past and future. This produces more accurate predictions.[36,37]

Considering that both LR and RF are unable to process time series variables efficiently, we pre-processed the clinical variables and all-time steps and corresponding variables were flattened into a single record. This was done to ensure that both LR and RF have access to the same data about the changes in patient state as BiLSTM, to ensure a fair performance comparison.

Confidential: For Review Only

Statistical Analysis

The classification results for delirium prediction are reported using the Area Under Receiver Operating Characteristic (AUROC), Area Under Precision Recall Curve (AUPRC), Recall, Precision (Positive Predictive Value) and Negative Predictive Value. Furthermore, we also investigated calibration quality of our models.

Model Interpretability

Although there are many definitions of interpretability, we focused on how the model ranks each input variable with respect to outcome prediction. LR and RF have been successfully employed in the clinical domain due to their ease of interpretation, however they require additional processing to handle high-dimensional, longitudinal and irregular EHR datasets.[38] In this context, we employed the Shapley Value Sampling (SVS) method to probe the Bi-LSTM model.[39] SVS is a perturbation-based method to compute variable attribution, which is based on sampling theory that can be used to estimate Shapley values.[40] The SVS produces feature ranking with respect to each feature input, allowing us to rank these variables based on their predictive power. Given that interpretability of neural networks is still an open research question, especially for temporal neural networks, we also provide results from two other methods, namely Integrated Gradient and Guided Backpropagation, to ensure that the variable importance results are consistent across the three methods.[41–43]

Confidential: For Review Only

Source code

The entire code is available at https://github.com/mostafaalishahi/Delirium_prediction_models

RESULTS

Patient characteristics

The eICU-CRD cohort consisted of 16,546 patients, with a mean age of 62.84 (±16.02) years and 46.53% were female. The incidence of delirium was 19.06% (Table-2). In the first 48 hours of admission, 59.30% of patients presented with delirium. The MIMIC-III cohort consisted of 6,294 patients, with a mean age of 63.58 (±15.79) years and 43.82% were female. The incidence of delirium was 20.15% and 66.34% of patients presented within the first 48 hours of ICU admission (Table-2). For vital signs and laboratory values that were generated hourly, an average of 7% were missing values.

Confidential: For Review Only

Table 2: Characteristics of the included patients divided by their CAM-ICU status.

Variable	eICU			MIMIC		
	CAM-ICU +	CAM-ICU -	*p* value	CAM-ICU +	CAM-ICU -	*p* value
Number of Patients	3153	13393		1268	5026	
Age, mean (SD), years	65.53 (15.14)	62.20 (16.16)	< 0.05	64.81 (15.62)	63.27 (15.82)	< 0.05
Female (%)	1405 (44)	6295 (47)	-	545 (43)	2211 (44)	-
Height, mean (SD), m	168.47 (18.23)	169.25 (15.90)	< 0.05	170.06 (14.22)	168.88 (14.87)	0.054
Weight, mean (SD), kg	83.06 (29.88)	85.00 (25.58)	< 0.05	82.68 (30.25)	81.53 (24.89)	0.15
Heart Rate, mean (SD), bpm	88.22 (18.06)	85.09 (17.73)	< 0.05	88.60 (17.53)	85.12 (17.29)	< 0.05
Oxygen Saturation, mean (SD), %	97.16 (2.72)	96.80 (2.79)	< 0.05	97.17 (2.71)	96.58 (4.50)	< 0.05
Glucose, mean (SD), mg/dL	140.32 (45.97)	146.46 (56.31)	< 0.05	144.51 (58.70)	141.25 (51.43)	< 0.05
Temperature, mean (SD), °C	37.01 (0.69)	36.97 (2.65)	< 0.05	37.06 (0.76)	36.88 (0.76)	< 0.05
Serum Sodium, mean (SD), mEq/L	140.32 (5.80)	138.57 (5.04)	< 0.05	139.39 (5.48)	138.32 (4.89)	< 0.05
BUN, mean (SD), mg/dL	31.93 (22.10)	25.88 (18.64)	< 0.05	33.96 (24.46)	28.10 (20.77)	< 0.05
WBC, mean (SD), per microliter	13.01 (6.47)	11.08 (5.51)	< 0.05	12.13 (7.73)	10.74 (6.29)	< 0.05
Hemoglobin, mean (SD), g/dL	9.73 (1.89)	10.00 (2.08)	< 0.05	9.76 (1.68)	10.27 (1.76)	< 0.05
Platelets, mean (SD), per microliter	201.34 (122.76)	210.23 (108.70)	< 0.05	202.59 (137.23)	199.53 (114.33)	< 0.05
Serum Potassium, mean (SD), mEq/L	3.98 (0.59)	4.00 (0.57)	0.1431	4.03 (0.57)	4.07 (0.56)	< 0.05
Chloride, mean (SD), mEq/L	105.54 (6.86)	103.24 (6.29)	< 0.05	104.57 (6.69)	104.36 (6.37)	< 0.05
Serum Bicarbonate, mean (SD), mEq/L	35.23 (5.02)	25.52 (5.02)	< 0.05	25.16 (5.21)	24.88 (4.95)	< 0.05
Serum creatinine, mean (SD), mg/dL	1.45 (1.16)	1.37 (1.21)	< 0.05	1.63 (1.28)	1.37 (1.05)	< 0.05
Ventilation, mean (SD)	0.87 (0.34)	0.71 (0.45)	< 0.05	0.56 (0.50)	0.33 (0.47)	< 0.05
Total norepinephrine dose (SD), mcg/kg/min	0.02 (0.31)	0.01 (0.28)	< 0.05	0.08 (0.63)	0.06 (0.57)	< 0.05
SOFA, mean (SD)	4.9 (3.3)	3.42 (2.84)	< 0.05	6.46 (3.77)	6.67 (3.34)	< 0.05
SOFA without GCS, mean (SD)	3.27 (2.83)	2.58 (2.33)	< 0.05	5.42 (3.65)	4.99 (3.13)	< 0.05

Abbreviations: CAM-ICU: confusion assessment method in the ICU, +: present, -: absent, SD: standard deviation, m: meter, kg: kilogram, bpm: beats/minute, mg/dL: milligrams/deciliter, °C: degree Celcius, mEq/L: milli equivalents per liter, g/dL: gram per deciliter, mcg/kg/min: micrograms per kilogram per minute, SOFA: sequential organ failure assessment, GCS: Glasgow coma scale.

Confidential: For Review Only

Confidential: For Review Only

Performance of Machine Learning Models

The BiLSTM algorithm was noted to have had the highest AUROC and AUPRC values for most of the observation-prediction combinations. With 24 hour observation of the eICU-CRD cohort and 48 hour prediction, the AUROC of BiLSTM model was 84.87% (95% CI, 83.32%-86.41%), LR 82.57% (95% CI, 79.64%-85.47%) and RF 83.24% (95% CI, 81.83%-84.67%), and AUPRC of 34.97% (95% CI,32.22%-37.27%), 31.07% (95% CI, 27.62%-33.81%) and 32.82% (95% CI, 28.89%-36.75%) respectively (Figure-1).

Since BiLSTM had the best AUROCs and AUPRCs, we calculated the precision and recall values in each observation-prediction window using BiLSTM. In the eICU-CRD derivation cohort, for the 12 hour observation window, the precision and recall decreased from 37.52% (95% CI,36.00%-39.05%) to 28.68% (95% CI, 24.88%-32.49%) and from 86.1% (95% CI, 82.49%-89.71%) to 63.49% (95% CI, 52.91%-74.08%) respectively when the prediction window changed from 12 hour to 96 hours (Figure-2 and Table-3). When increasing the observation window for 48-hour prediction, the precision and recall changed from 32.82% (95% CI, 29.6%-36 .04%) to 17.9% (95% CI, 15.37%-20.44%) and from 82.22% (95% CI, 78.16%-86.27%) to 73.95% (95% CI, 64.8%-83.11%).

As we were interested in making our model more sensitive for screening, we changed thresholds to have higher recall at the expense of precision. For a 12-hour observation window, while recall changed slightly from 93.73% (95% CI, 93.1% - 94.37%) to 92.57% (95% CI, 88.19%-96.95%) as the prediction window changed from 12 hour to 96 hours, the precision decreased from 26.96% (95% CI, 24.99%-28.94%) to 11.34% (95% CI, 10.71%-11.98%) (Table-3). For the 48-hour prediction window as we increased the observation window from 12 hours to 48 hours, the precision and recall changed from 16.82% (95% CI, 15.61%-18.02%) to 15.64% (95% CI, 13.96%-17.42%) and 92.15% (95% CI, 88.47%-95.82) to 91.13% (95% CI, 89.57%- 92.69%) respectively. Comparable results for the MIMIC-III cohort are presented in supplement (Table-4, Figure-3, and Figure-4). A heat map demonstrating correlation among features is presented in eFigure-3 for the eICU-CRD for the MIMIC-III populations.

Confidential: For Review Only

Table 3: Performance metrics of derived model in eICU-CRD cohort.

Prediction window	12 hours	24 hours	48 hours	72 hours	96 hours
Observation window					
A.	**Unmodified Threshold**				
	Unmodified threshold: Area under receiver operating curve				
12 hours (95% CI), %	87.82 (87.17-88.30)	86.82 (85.15-88.64)	84.00 (81.68-86.13)	81.45 (78.61-84.10)	79.03 (76.69-82.11)
24 hours (95% CI), %	88.39 (86.41-89.96)	86.68 (85.79-88.15)	84.87 (83.32-86.41)	81.99 (80.66-83.38)	79.93 (76.57-83.34)
48 hours (95% CI), %	88.00 (75.78-89.59)	87.23 (86.30-88.20)	84.51 (82.14-86.92)	82.19 (80.99-83.41)	79.78 (75.37-84.25)
	Unmodified threshold: Area under precision recall curve				
12 hours (95% CI), %	46.86 (42.52-50.85)	40.92 (37.03-44.46)	34.04 (28.99-38.24)	26.78 (25.24-27.71)	24.90 (18.48-30.22)
24 hours (95% CI), %	44.62 (39.11-50.02)	40.85 (38.38-43.10)	34.97 (32.22-37.27)	28.68 (23.78-33.02)	26.37 (21.00-31.28)
48 hours (95% CI), %	41.67 (37.52-45.62)	39.64 (37.00-42.07)	33.35 (27.58-38.88)	29.75 (26.06-32.90)	26.43 (19.65-32.71)
	Unmodified threshold: Precision				
12 hours (95% CI), %	37.52 (36.00-39.05)	32.68 (29.09-36.28)	25.01 (22.73-27.28)	21.30 (20.09-22.49)	17.45 (15.83-19.08)
24 hours (95% CI), %	35.27 (33.51-37.03)	30.69 (28.71-32.66)	24.84 (23.35-26.32)	20.69 (18.24-23.15)	19.08 (17.85-20.31)
48 hours (95% CI), %	32.82 (29.60-36.04)	29.37 (25.18-33.56)	24.17 (21.68-26.67)	22.25 (67.85-82.37)	17.90 (15.37-20.44)
	Unmodified threshold: Recall				
12 hours (95% CI), %	86.10 (82.49-89.71)	84.09 (81.81-86.37)	80.53 (76.76-84.30)	77.96 (69.87-86.05)	75.58 (68.33-82.83)
24 hours (95% CI), %	84.74 (81.57-87.90)	83.87 (81.24-86.50)	79.44 (75.53-83.35)	78.73 (72.41-85.05)	71.20 (61.95-80.45)
48 hours (95% CI), %	82.22 (78.16-86.27)	82.06 (78.55-85.56)	80.38 (75.53-85.24)	75.11 (67.85-82.37)	73.95 (64.80-83.11)
B.	**Threshold optimized favoring a higher recall**				
	Threshold optimized favoring higher recall: Area under receiver operating curve				
12 hours (95% CI), %	87.45 (86.87-88.03)	86.41 (84.12-88.71)	83.63 (81.43-85.83)	81.19 (78.49-83.89)	79.01 (76.10-81.92)
24 hours (95% CI), %	87.93 (86.39-89.48)	86.63 (85.41-87.86)	84.25 (82.92-85.62)	81.50 (80.13-82.90)	79.66 (76.61-82.72)
48 hours (95% CI), %	87.24 (85.34-89.15)	85.93 (84.29-87.60)	83.94 (81.72-85.90)	81.76 (81.03-82.61)	78.99 (74.90-83.01)
	Threshold optimized favoring higher recall: Area under precision recall curve				
12 hours (95% CI), %	46.63 (42.17-50.93)	39.52 (34.88-43.89)	33.21 (28.84-36.83)	26.55 (23.55-28.97)	24.30 (18.52-29.21)
24 hours (95% CI), %	44.55 (39.40-49.02)	39.95 (38.25-41.47)	33.70 (30.96-36.07)	27.49 (22.86-31.46)	26.11 (22.02-29.85)
48 hours (95% CI), %	40.96 (36.55-44.72)	36.98 (32.38-41.33)	32.12 (26.50-37.45)	29.55 (25.26-33.35)	24.65 (15.60-33.20)
	Threshold optimized favoring higher recall: Precision				
12 hours (95% CI), %	26.96 (24.99-28.94)	22.04 (20.66-23.42)	16.82 (15.61-18.02)	13.33 (13.03-13.60)	11.34 (10.71-11.98)
24 hours (95% CI), %	23.61 (22.55-24.66)	21.73 (20.63-22.83)	16.57 (15.74-17.38)	13.46 (12.29-14.62)	12.60 (11.81-13.39)
48 hours (95% CI), %	23.18 (20.49-25.87)	18.70 (14.49-22.87)	15.64 (13.96-17.42)	14.02 (12.06-16.04)	11.69 (10.75-12.73)
	Threshold optimized favoring higher recall: Recall				
12 hours (95% CI), %	93.73 (93.10-94.37)	93.08 (90.42-95.75)	92.15 (88.47-95.82)	92.08 (90.25-93.91)	92.57 (88.19-96.95)
24 hours (95% CI), %	93.59 (91.69-95.48)	92.29 (88.83-95.76)	91.65 (89.07-94.23)	89.72 (86.74-92.69)	90.40 (88.58-92.23)
48 hours (95% CI), %	90.49 (86.48-94.50)	91.46 (89.97-92.95)	91.13 (89.57-92.69)	89.37 (84.87-93.41)	90.20 (82.79-97.61)

Panel A. Unmodified thresholds. Panel B. After thresholds were optimized favoring higher recall. Abbreviations: eICU-CRD: eICU Collaborative Research Database, 95% CI: 95 percent confidence interval, %: percentage.

Confidential: For Review Only

Table 4: Performance metrics of LSTM model in MIMIC-III cohort.

Prediction window	12 hours	24 hours	48 hours	72 hours	96 hours
Observation window					
A.			**Unmodified Threshold**		
			Area under receiver operating curve		
12 hours (95% CI), %	80.34 (78.31-82.21)	77.64 (75.92-79.28)	73.38 (69.43-77.15)	71.47 (66.24-76.77)	69.21 (63.95-74.41)
24 hours (95% CI), %	81.72 (78.09-85.36)	78.25 (75.97-80.63)	72.14 (64.37-79.61)	69.06 (61.33-77.29)	66.26 (56.31-76.38)
48 hours (95% CI), %	81.15 (79.46-82.30)	77.90 (74.96-80.84)	70.38 (64.35-76.59)	65.87 (58.36-73.44)	67.20 (61.93-72.46)
			Area under precision recall curve		
12 hours (95% CI), %	41.61 (36.14-46.56)	40.97 (34.96-46.05)	33.52 (30.06-37.08)	34.93 (29.39-39.89)	31.16 (26.65-35.69)
24 hours (95% CI), %	48.00 (43.11-52.94)	42.54 (36.27-48.58)	34.19 (27.39-40.66)	32.76 (24.17-41.03)	27.29 (19.35-34.66)
48 hours (95% CI), %	48.08 (42.59-53.32)	43.48 (36.68-50.22)	34.15 (29.67-38.03)	28.02 (22.75-32.56)	29.33 (24.80-33.66)
			Precision		
12 hours (95% CI), %	30.14 (26.54-33.74)	35.12 (31.85-38.39)	30.99 (27.91-34.07)	30.86 (26.90-34.82)	28.68 (24.88-32.49)
24 hours (95% CI), %	34.07 (31.36-36.79)	33.35 (29.82-36.88)	30.21 (27.00-33.41)	28.36 (23.08-33.65)	24.71 (19.89-29.52)
48 hours (95% CI), %	36.05 (32.37-39.74)	34.27 (32.22-36.32)	30.61 (28.27-32.95)	26.69 (21.06-32.32)	26.92 (22.57-31.26)
			Recall		
12 hours (95% CI), %	71.75 (68.75-74.74)	64.80 (57.11-72.49)	65.36 (62.42-68.29)	62.91 (58.26-67.57)	63.49 (52.91-74.08)
24 hours (95% CI), %	73.93 (67.53-80.32)	69.23 (66.58-71.89)	65.38 (59.77-70.99)	60.35 (49.13-71.57)	60.42 (46.04-74.80)
48 hours (95% CI), %	74.35 (67.09-81.61)	70.00 (66.75-73.25)	64.04 (53.35-74.74)	60.69 (48.72-72.66)	64.00 (49.96-78.04)
B.			**Threshold optimized favoring a higher recall**		
			Area under receiver operating curve		
12 hours (95% CI), %	80.25 (78.31-82.21)	77.61 (75.92-79.28)	73.27 (69.43-77.15)	71.51 (66.25-76.76)	69.12 (63.96-74.38)
24 hours (95% CI), %	81.67 (78.09-85.36)	78.26 (75.97-80.63)	71.99 (64.37-79.62)	69.31 (61.33-77.28)	66.35 (56.31-76.38)
48 hours (95% CI), %	80.89 (79.46-82.30)	77.87 (74.96-80.83)	70.47 (64.35-76.60)	65.86 (58.35-73.44)	67.09 (61.94-72.40)
			Area under precision recall curve		
12 hours (95% CI), %	41.30 (36.12-46.04)	41.02 (35.64-45.60)	33.56 (29.61-37.58)	35.07 (29.90-39.51)	31.07 (26.63-35.49)
24 hours (95% CI), %	47.35 (43.17-51.63)	42.61 (36.53-48.48)	34.07 (26.91-40.81)	32.74 (24.69-40.53)	27.24 (19.51-34.35)
48 hours (95% CI), %	47.30 (43.15-51.42)	43.50 (36.80-50.05)	34.86 (30.14-39.06)	29.53 (21.64-36.90)	29.44 (24.71-33.94)
			Precision		
12 hours (95% CI), %	20.98 (19.31-22.64)	23.78 (20.54-27.01)	21.67 (19.08-24.26)	23.27 (21.34-25.21)	23.30 (21.45-25.14)
24 hours (95% CI), %	25.67 (24.48-26.86)	25.41 (21.16-29.65)	23.09 (21.71-24.47)	23.35 (21.12-25.57)	20.90 (18.66-23.15)
48 hours (95% CI), %	28.08 (24.45-31.75)	26.67 (25.20-28.14)	24.57 (23.20-25.93)	22.51 (19.89-25.05)	23.70 (22.20-25.20)
			Recall		
12 hours (95% CI), %	86.63 (83.32-90.01)	76.95 (73.08-80.82)	81.46 (71.94-90.98)	84.47 (77.72-91.22)	87.38 (73.19-99.05)
24 hours (95% CI), %	82.22 (76.40-88.05)	81.14 (79.88-82.40)	87.36 (74.76-92.75)	84.11 (74.56-93.66)	86.14 (73.04-99.24)
48 hours (95% CI), %	83.18 (76.24-90.13)	83.79 (76.92-90.66)	82.24 (71.18-93.30)	83.20 (70.13-96.27)	87.38 (78.40-96.36)

Confidential: For Review Only

Panel A. Unmodified thresholds. Panel B. After thresholds were optimized favoring higher recall. Abbreviations: LSTM: long short term memory, SD: standard deviation, %: percentage.

Interpretability

Figure-5 ranks the features that have contributed to delirium prediction according to their relative importance in the eICU-CRD derived model. Ventilation, heart rate, age, white blood cell count, SOFA score and vasopressor use are the highest ranked features across different prediction windows. Most of these features are also the highest ranked features when assessing interpretability in the MIMIC III cohort (Figure-6).

Confidential: For Review Only

DISCUSSION

Our study shows that a machine learning model using only a few routine clinical variables replicated the performance of previously reported models that were developed using hundreds of variables. Our study successfully demonstrated that we could modify the performance of a model to fit our clinical needs as an effective screening tool. We took the following steps that helped us achieve our goal: 1.) we studied the peak delirium onset time in our population and optimized the model to maximize predictive accuracy in that time frame; 2.) we incorporated sliding windows in our model for continuous prediction across time and address drop in performance associated with predictions further ahead; and 3.) We adjusted our thresholds to favor a high recall to ensure the model detects all patients at risk of delirium. Furthermore, we demonstrated that performance across different datasets diminish in accuracy and needs to be individualized to the population. Our features when ranked suggest older and more critically ill patients are at greater risk of delirium, especially in combination with mechanical ventilation and vasopressor therapy. Our model's ranking of features is consistent with what we already know as high-risk features. We have shared our code for replicating the results and recommend adjustments be made according to the specific setting and their needs.[44]

Screening tools like CAM-ICU describe a snapshot in time and do not give an idea of the patient's progress nor are predictive. Strategies based on established best practices such as ABCDEF are resource intensive and challenging to implement universally.[11] Despite effective prevention strategies, delirium is still commonplace in the ICU

Confidential: For Review Only

highlighting a need for a screening tool that prioritizes patients at risk and allows us to exclude patients who are low risk from these time-consuming therapies. Few models exist that can both accurately predict and be easy to implement. Most models use several hundred variables or use only a snapshot of features that can vary with time. Also, many of these models were trained on small datasets and use inconsistent approaches for collecting and/or stratifying data into training and validation cohorts limiting generalizability.[18–20] The Pre-Deliric and e-Pre-Deliric, were built with a handful of predictor variables from a large patient cohort, and been externally validated.[45,46] However, they employ data from admission variables that change and lose predictive power with time. Recent machine learning based algorithms were able to predict delirium accurately but using over 700 predictor variables and were also criticized for their analysis methods.[20,47] Importantly, these models have not investigated how their performance changes with different prediction windows, optimal time of observation, capture the evolution of a patient's state through time and unable to adjust delirium risk estimation temporally. In our knowledge, our report is one of the first instances of delirium prediction, where we have not only tried to predict accurately across different scenarios but also addressed the issues with prior prediction models. Notably, we have iteratively developed our model to address the challenges that are posed by low incidence of delirium, temporal progression of disease and different patient populations. Additionally, we have ventured into the realm of explaining how our features contribute, something that is rare in models using deep learning.

Confidential: For Review Only

The BiLSTM-based model, which has the advantage of capturing temporal dependencies, performed the best of the three models evaluated, suggesting that the trajectory of predictive features is more informative than a single value. The simpler LR model is an attractive option if implementation is determined by computational limitations of a deep learning model. A longer observation window gained little in terms of model performance. A 48-hour observation window even led to a drop in accuracy, but this is due to a decrease in the size of the training cohort. Another possibility is that factors contributing to delirium are proximal to its onset, further justifying the use of continuous prediction using a sliding window. The decay in performance of the algorithm as it predicts delirium with longer lead time is similar in both MIMIC-III and eICU-CRD.

A screening tool needs to be sensitive. This is best addressed by a model with a high recall. We adjusted thresholds favoring a high recall while sacrificing precision (Table-3 and Table-4) to achieve this purpose. Also, it is desirable to have prediction algorithms that have short observation duration and predict the furthest ahead. In our case, since most (59% in eICU-CRD, 66% in MIMIC-III) delirium cases occurred within 48 hours of ICU admission (eFigure-2), hence we targeted performance for a 48-hour prediction window with a 12- or 24-hour observation window. We also demonstrated that as the prediction window moved beyond 48 hours the model-maintained recall, but with a precipitous drop in precision. Non-trivial tuning of hyperparameters is required when algorithms are ported across populations. We suggest the performance of different

Confidential: For Review Only

observation and prediction times be studied on the local dataset and depending on the objective of the algorithm the optimal windows are determined.

Delirium is precipitated through many factors, some that are unique to the ICU. Our variables were chosen *a priori* based on literature review. We only included variables that can be easily extracted in real time. Instead of using static values, we employed a sliding window for prediction and incorporated the trajectory of each variable over time. Our results indicate that this strategy predicts delirium more accurately than values captured at a moment in time and eliminates the need for long term prediction.

Since we conducted a retrospective study, causality between the features and delirium cannot be established. Other limitations include selection bias (we excluded observations with missing CAM-ICU values) and interpreter bias (the data recorded in the databases might have been collected after the onset of delirium, given the noncontinuous nature of CAM-ICU measurement). Additionally, CAM-ICU was scored by different nurses at separate times and in different units, potentially resulting in inter-operator variability.

Confidential: For Review Only

Conclusion

We successfully designed a delirium prediction model as a potential screening tool for ABCDEF bundle implementation. Using a few clinically relevant predictor variables we were able to achieve comparable performance to contemporary and well reported models. We were able to tackle the challenge presented by evolving temporal and treatment effects by using methods that captured temporal trends in data rather than static values and sliding observation windows, threshold adjustments to ensure consistently high recall. Additionally, we peeked at interpreting the model and shared our code online for reproducibility. We believe our model will help with identifying patients at risk of delirium early and will allow us to target preventive therapies, which is often time consuming and personnel-intensive, to the patients who are most likely to benefit.

Confidential: For Review Only

REFERENCES:

1 Krewulak KD, Stelfox HT, Leigh J, *et al.* Incidence and Prevalence of Delirium Subtypes in an Adult ICU. *Crit Care Med* 2018;46:2029–35. doi:10.1097/ccm.0000000000003402

2 Dasgupta M, Brymer C. Poor functional recovery after delirium is associated with other geriatric syndromes and additional illnesses. *Int Psychogeriatr* 2015;27:793–802. doi:10.1017/s1041610214002658

3 Kamdar BB, Combs MP, Colantuoni E, *et al.* The association of sleep quality, delirium, and sedation status with daily participation in physical therapy in the ICU. *Crit Care* 2016;20:261. doi:10.1186/s13054-016-1433-z

4 Sakusic A, O'Horo JC, Dziadzko M, *et al.* Potentially Modifiable Risk Factors for Long-Term Cognitive Impairment After Critical Illness: A Systematic Review. *Mayo Clin Proc* 2018;93:68–82. doi:10.1016/j.mayocp.2017.11.005

5 Schubert M, Schürch R, Boettger S, *et al.* A hospital-wide evaluation of delirium prevalence and outcomes in acute care patients - a cohort study. *Bmc Health Serv Res* 2018;18:550. doi:10.1186/s12913-018-3345-x

6 Ely EW, Shintani A, Truman B, *et al.* Delirium as a Predictor of Mortality in Mechanically Ventilated Patients in the Intensive Care Unit. *Jama* 2004;291:1753–62. doi:10.1001/jama.291.14.1753

7 Pandharipande P, Cotton BA, Shintani A, *et al.* Prevalence and Risk Factors for Development of Delirium in Surgical and Trauma Intensive Care Unit Patients. *J Trauma Inj Infect Critical Care* 2008;65:34–41. doi:10.1097/ta.0b013e31814b2c4d

8 McPherson JA, Wagner CE, Boehm LM, *et al.* Delirium in the Cardiovascular ICU. *Crit Care Med* 2013;41:405–13. doi:10.1097/ccm.0b013e31826ab49b

9 Pandharipande PP, Girard TD, Jackson JC, *et al.* Long-term cognitive impairment after critical illness. *New Engl J Medicine* 2013;369:1306–16. doi:10.1056/nejmoa1301372

10 Fong TG, Tulebaev SR, Inouye SK. Delirium in elderly adults: diagnosis, prevention and treatment. *Nat Rev Neurol* 2009;5:210–20. doi:10.1038/nrneurol.2009.24

11 Marra A, Ely EW, Pandharipande PP, *et al.* The ABCDEF Bundle in Critical Care. *Crit Care Clin* 2017;33:225–43. doi:10.1016/j.ccc.2016.12.005

Confidential: For Review Only

12 Devlin JW, Skrobik Y, Gélinas C, *et al.* Clinical Practice Guidelines for the Prevention and Management of Pain, Agitation&sol;Sedation, Delirium, Immobility, and Sleep Disruption in Adult Patients in the ICU. *Crit Care Med* 2018;46:e825–73. doi:10.1097/ccm.0000000000003299

13 Hsieh SJ, Otusanya O, Gershengorn HB, *et al.* Staged Implementation of Awakening and Breathing, Coordination, Delirium Monitoring and Management, and Early Mobilization Bundle Improves Patient Outcomes and Reduces Hospital Costs*. *Crit Care Med* 2019;47:885–93. doi:10.1097/ccm.0000000000003765

14 Ely WE, 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). *Crit Care Med* 2001;29:1370–9. doi:10.1097/00003246-200107000-00012

15 Girard TD, Pandharipande PP, Ely WE. Delirium in the intensive care unit. *Critical Care Lond Engl* 2008;12 Suppl 3:S3. doi:10.1186/cc6149

16 Brummel NE, Vasilevskis EE, Han J, *et al.* Implementing delirium screening in the ICU: secrets to success. *Crit Care Med* 2013;41:2196–208. doi:10.1097/ccm.0b013e31829a6f1e

17 Khan BA, Perkins AJ, Prasad NK, *et al.* Biomarkers of Delirium Duration and Delirium Severity in the ICU*. *Crit Care Med* 2020;48:353–61. doi:10.1097/ccm.0000000000004139

18 Ruppert MM, Lipori J, Patel S, *et al.* ICU Delirium-Prediction Models: A Systematic Review. *Critical Care Explor* 2020;2:e0296. doi:10.1097/cce.0000000000000296

19 Lindroth H, Bratzke L, Purvis S, *et al.* Systematic review of prediction models for delirium in the older adult inpatient. *Bmj Open* 2018;8:e019223. doi:10.1136/bmjopen-2017-019223

20 Wong A, Young AT, Liang AS, *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 Netw Open* 2018;1:e181018. doi:10.1001/jamanetworkopen.2018.1018

21 Pollard TJ, Johnson AE, Raffa JD, *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 2018;5:180178. doi:10.1038/sdata.2018.178

22 Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035. doi:10.1038/sdata.2016.35

Confidential: For Review Only

23 Inouye SK. A Predictive Model for Delirium in Hospitalized Elderly Medical Patients Based on Admission Characteristics. *Ann Intern Med* 1993;119:474. doi:10.7326/0003-4819-119-6-199309150-00005

24 Pompei P, Foreman M, Rudberg MA, *et al.* Delirium in Hospitalized Older Persons: Outcomes and Predictors. *J Am Geriatr Soc* 1994;42:809–15. doi:10.1111/j.1532-5415.1994.tb06551.x

25 Kim MY, Park UJ, Kim HT, *et al.* DELirium Prediction Based on Hospital Information (Delphi) in General Surgery Patients. *Medicine* 2016;95:e3072. doi:10.1097/md.0000000000003072

26 Pendlebury ST, Lovett NG, Smith SC, *et al.* Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of a susceptibility score based on factors identified externally in pooled data for use at entry to the acute care pathway. *Age Ageing* 2016;46:226–31. doi:10.1093/ageing/afw198

27 Rudolph JL, Jones RN, Levkoff SE, *et al.* Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. *Circulation* 2008;119:229–36. doi:10.1161/circulationaha.108.795260

28 Carrasco MP, Villarroel L, Andrade M, *et al.* Development and validation of a delirium predictive score in older people. *Age Ageing* 2013;43:346–51. doi:10.1093/ageing/aft141

29 Leung JM, Sands LP, Lim E, *et al.* Does Preoperative Risk for Delirium Moderate the Effects of Postoperative Pain and Opiate Use on Postoperative Delirium? *Am J Geriatric Psychiatry* 2013;21:946–56. doi:10.1016/j.jagp.2013.01.069

30 O'KEEFFE ST, LAVAN JN. Predicting Delirium in Elderly Patients: Development and Validation of a Risk-stratification Model. *Age Ageing* 1996;25:317–21. doi:10.1093/ageing/25.4.317

31 Kalisvaart KJ, Vreeswijk R, Jonghe JFMD, *et al.* Risk Factors and Prediction of Postoperative Delirium in Elderly Hip-Surgery Patients: Implementation and Validation of a Medical Risk Factor Model. *J Am Geriatr Soc* 2006;54:817–22. doi:10.1111/j.1532-5415.2006.00704.x

32 Jang S, Jung K-I, Yoo W-K, *et al.* Risk Factors for Delirium During Acute and Subacute Stages of Various Disorders in Patients Admitted to Rehabilitation Units. *Ann Rehabilitation Medicine* 2016;40:1082–91. doi:10.5535/arm.2016.40.6.1082

33 Rudolph JL, Doherty K, Kelly B, *et al.* Validation of a Delirium Risk Assessment Using Electronic Medical Record Information. *J Am Med Dir Assoc* 2016;17:244–8. doi:10.1016/j.jamda.2015.10.020

Confidential: For Review Only

34 Rudolph JL, Harrington MB, Lucatorto MA, *et al.* Validation of a medical record-based delirium risk assessment. *J Am Geriatr Soc* 2011;59 Suppl 2:S289-94. doi:10.1111/j.1532-5415.2011.03677.x

35 Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J Mach Learn Res*;11:2079–107.https://jmlr.org/papers/v11/cawley10a.html

36 Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput* 1997;9:1735–80. doi:10.1162/neco.1997.9.8.1735

37 Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *Ieee T Signal Proces* 1997;45:2673–81. doi:10.1109/78.650093

38 Miotto R, Wang F, Wang S, *et al.* Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2017;19:1236–46. doi:10.1093/bib/bbx044

39 Castro J, Gómez D, Tejada J. Polynomial calculation of the Shapley value based on sampling. *Comput Oper Res* 2009;36:1726–30. doi:10.1016/j.cor.2008.04.004

40 Strumbelj E, Kononenko E. An Efficient Explanation of Individual Classifications using Game Theory | The Journal of Machine Learning Research. *J Mach Learn Res* 3AD;11:1–18.https://dl.acm.org/doi/10.5555/1756006.1756007

41 Ismail AA, Gunady M, Bravo HC, *et al.* Benchmarking Deep Learning Interpretability in Time Series Predictions. *Arxiv* 2020.

42 Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. *Arxiv* 2017.

43 Springenberg JT, Dosovitskiy A, Brox T, *et al.* Striving for Simplicity: The All Convolutional Net. *Arxiv* 2014.

44 Futoma J, Simons M, Panch T, *et al.* The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Heal* 2020;2:e489–92. doi:10.1016/s2589-7500(20)30186-2

45 Boogaard M v d, 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:e420–e420. doi:10.1136/bmj.e420

46 Wassenaar A, Boogaard M van den, Achterberg T van, *et al.* Multinational development and validation of an early prediction model for delirium in ICU patients. *Intens Care Med* 2015;41:1048–56. doi:10.1007/s00134-015-3777-2

Confidential: For Review Only

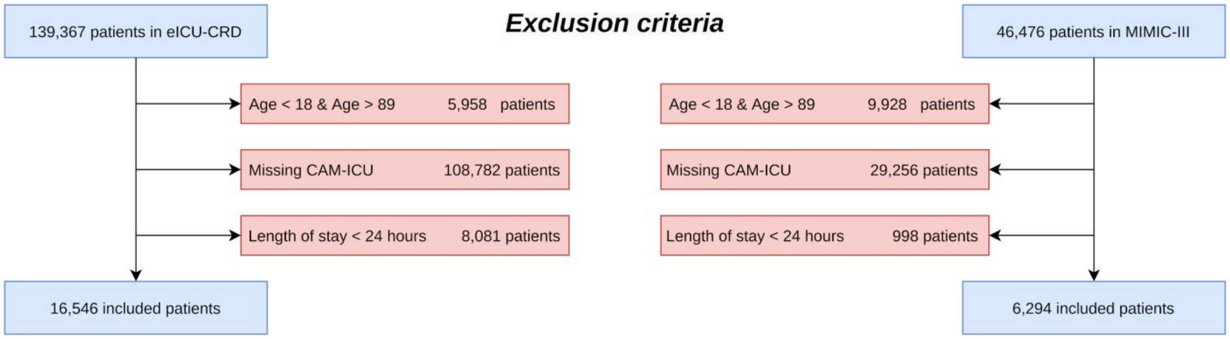
47 Rose S. Machine Learning for Prediction in Electronic Health Data. *Jama Netw Open* 2018;1:e181404–e181404. doi:10.1001/jamanetworkopen.2018.1404

Confidential: For Review Only

Contents:

Confidential: For Review Only

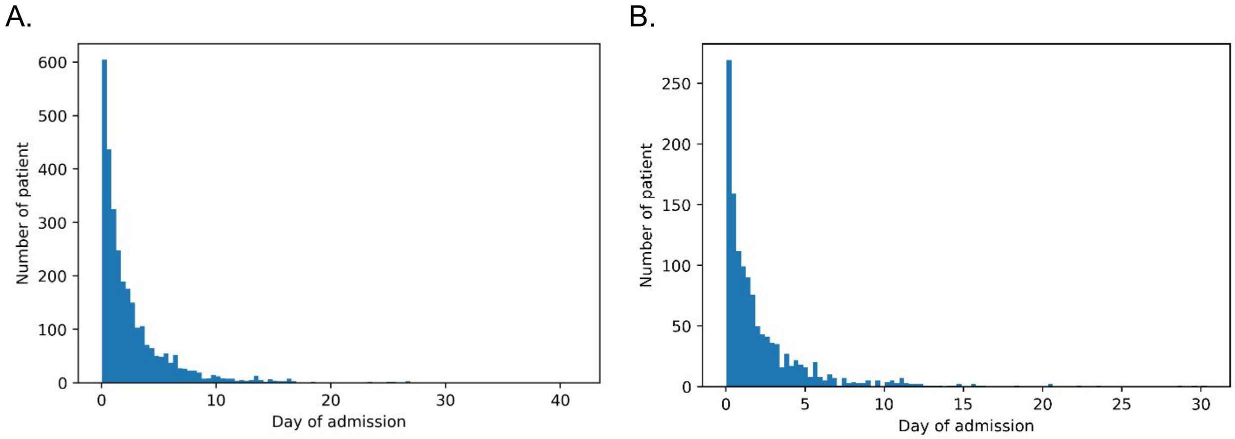
eFigure 1: Consort Diagram


Exclusion criteria
139,367 patients in eICU-CRD
Age < 18 & Age > 89 5,958 patients
Missing CAM-ICU 108,782 patients
Length of stay < 24 hours 8,081 patients
16,546 included patients
46,476 patients in MIMIC-III
Age < 18 & Age > 89 9,928 patients
Missing CAM-ICU 29,256 patients
Length of stay < 24 hours 998 patients
6,294 included patients

Left: eICU-CRD, Right: MIMIC III

Confidential: For Review Only

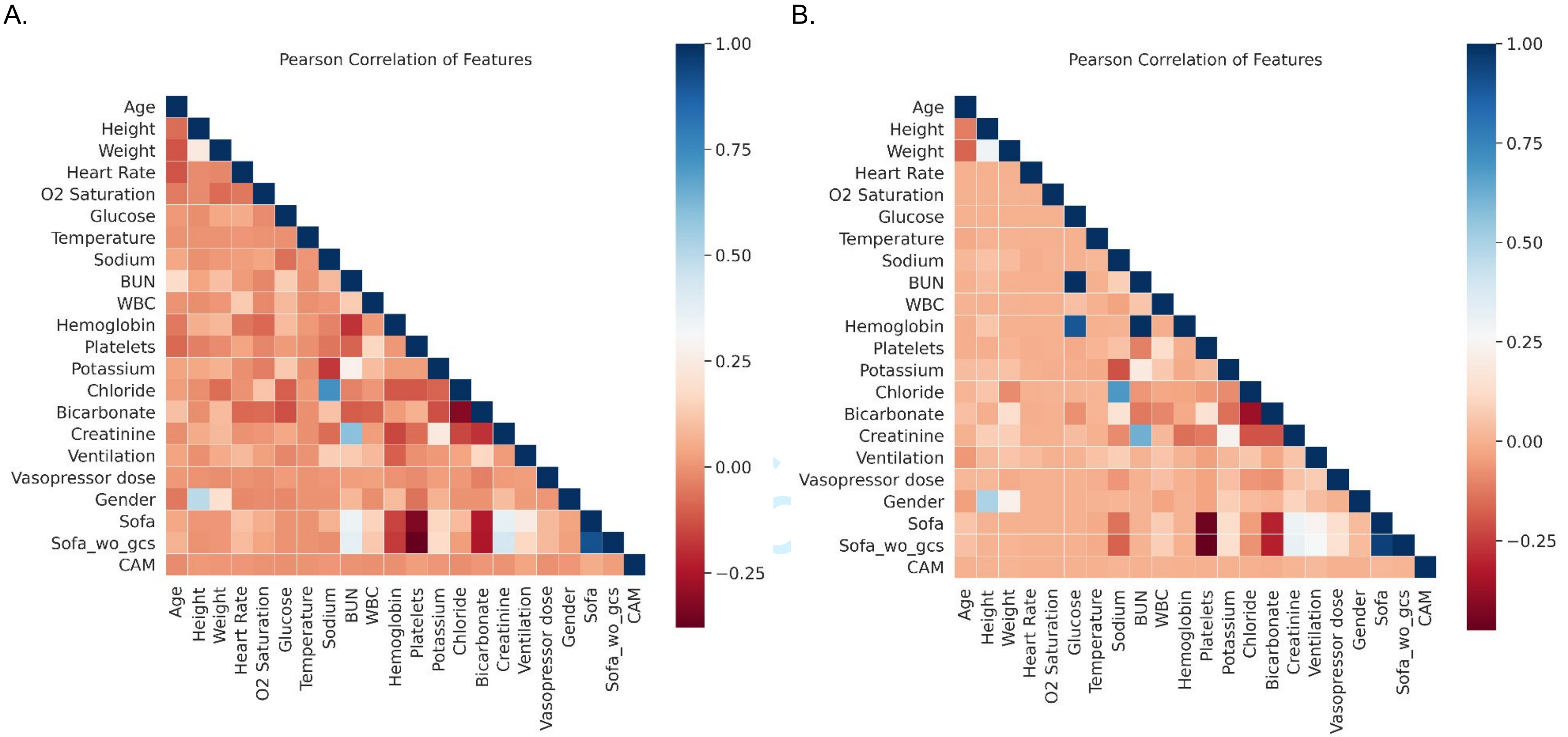
eFigure 2: Delirium incidence by day



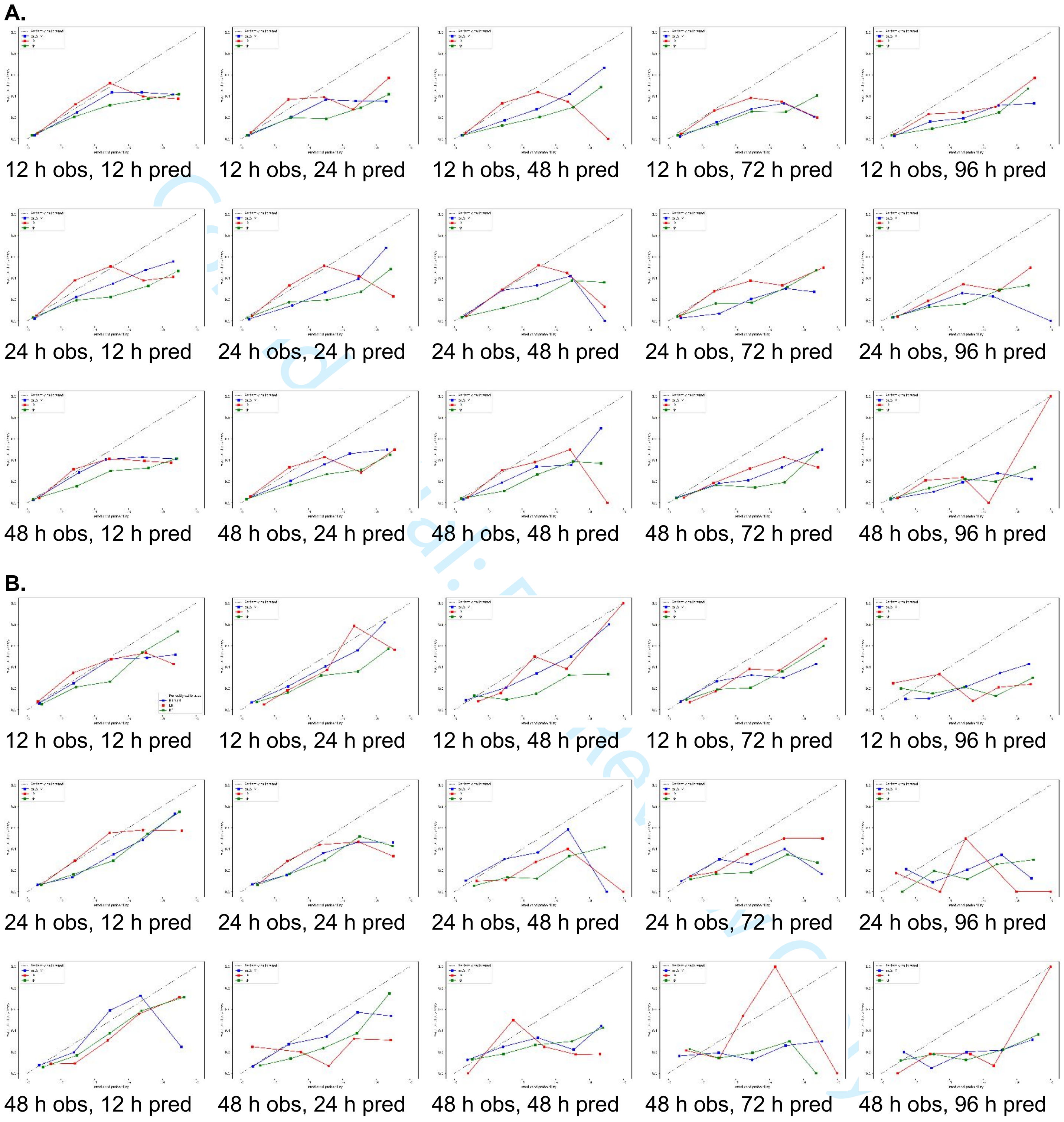
Panel A: eICU-CRD, Panel B: MIMIC-III

Confidential: For Review Only

eFigure 3: Heat-map showing correlation between variables.



Blue shows strong positive correlation, Red shows strong negative correlation. Panel A: eICU-CRD, Panel B: MIMIC-III eFigure 4: Calibration Graphs for Machine Learning Models



12 h obs, 12 h pred
12 h obs, 24 h pred
12 h obs, 48 h pred
12 h obs, 72 h pred
12 h obs, 96 h pred

24 h obs, 12 h pred
24 h obs, 24 h pred
24 h obs, 48 h pred
24 h obs, 72 h pred
24 h obs, 96 h pred

48 h obs, 12 h pred
48 h obs, 24 h pred
48 h obs, 48 h pred
48 h obs, 72 h pred
48 h obs, 96 h pred

12 h obs, 12 h pred
12 h obs, 24 h pred
12 h obs, 48 h pred
12 h obs, 72 h pred
12 h obs, 96 h pred

24 h obs, 12 h pred
24 h obs, 24 h pred
24 h obs, 48 h pred
24 h obs, 72 h pred
24 h obs, 96 h pred

48 h obs, 12 h pred
48 h obs, 24 h pred
48 h obs, 48 h pred
48 h obs, 72 h pred
48 h obs, 96 h pred

Panel A: eICU-CRD, Panel B: MIMIC III. Abbreviations: LR: logistic regression, RF: random forest, LSTM: long short term memory.

Confidential: For Review Only

eFigure 5: Brier Scores

Observation Window	Prediction Window	BiLSTM	RF	LR
12 hours	12 hours	0.0912	0.0938	0.0945
12 hours	24 hours	0.0901	0.1058	0.0986
12 hours	48 hours	0.1161	0.1272	0.1179
12 hours	72 hours	0.1344	0.1365	0.1296
12 hours	96 hours	0.1390	0.1492	0.1483
24 hours	12 hours	0.0947	0.0961	0.0957
24 hours	24 hours	0.1075	0.1113	0.1125
24 hours	48 hours	0.1287	0.1322	0.1320
24 hours	72 hours	0.1442	0.1473	0.1447
24 hours	96 hours	0.1440	0.1459	0.1468
48 hours	12 hours	0.1109	0.1101	0.1206
48 hours	24 hours	0.1141	0.1256	0.1358
48 hours	48 hours	0.1481	0.1470	0.1511
48 hours	72 hours	0.1533	0.1509	0.1497
48 hours	96 hours	0.1428	0.1401	0.1423

Abbreviations: LR: logistic regression, RF: random forest, LSTM: long short term memory.

r Review Only