Daily automated prediction of delirium risk in hospitalized patients: Model development and validation

Kendrick Matthew Shaw, Yu-Ping Shao, Manohar Ghanta, Valdery Moura Junior, Eyal Y Kimchi, Timothy T Houle, Oluwaseun Akeju, Michael Brandon Westover

Submitted to: JMIR Medical Informatics
on: May 15, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Daily automated prediction of delirium risk in hospitalized patients: Model development and validation

Kendrick Matthew Shaw[1, 2] MD, PhD; Yu-Ping Shao[3] MS; Manohar Ghanta[3, 4] MS; Valdery Moura Junior[3, 4] MS; Eyal Y Kimchi[5] MD, PhD; Timothy T Houle[1, 2] PhD; Oluwaseun Akeju[1, 2] MD; Michael Brandon Westover[3, 4]

[1]Department of Anesthesia, Pain, and Critical care Medicine Massachusetts General Hospital Boston US
[2]Harvard Medical School Boston US
[3]Department of Neurology Massachusetts General Hospital Boston US
[4]Department of Neurology Beth Israel Deaconess Medical Center Boston US
[5]Ken & Ruth Davee Department of Neurology Feinberg School of Medicine Northwestern University Chicago US

Corresponding Author:
Kendrick Matthew Shaw MD, PhD
Department of Anesthesia, Pain, and Critical care Medicine
Massachusetts General Hospital
55 Fruit Street
Boston
US

Abstract

Background: Delirium is common in hospitalized patients and correlated with increased morbidity and mortality. Despite this, delirium is underdiagnosed, and many institutions do not have sufficient resources to consistently apply effective screening and prevention.

Objective: To develop a machine learning algorithm to identify patients at highest risk of delirium in the hospital each day in an automated fashion based on data available in the electronic medical record, reducing the barrier to large-scale delirium screening.

Methods: We developed and compared multiple machine learning models on a retrospective dataset of all hospitalized adult patients with recorded Confusion Assessment Method (CAM) screens at a major academic medical center from April 2nd, 2016 to January 16th 2019, comprising 23006 patients. The patient's age, gender, and all available laboratory values, vital signs, prior CAM screens, and medication administrations were used as potential predictors. Four machine learning approaches were investigated: logistic regression with L1-regularization, multilayer perceptrons, random forests, and boosted trees. Model development used 80% of the patients; the remaining 20% were reserved for testing the final models. Lab values, vital signs, medications, gender, and age were used to predict a positive CAM screen in the next 24 hours.

Results: The boosted tree model achieved the greatest predictive power, with a 0.92 area under the receiver operator characteristic curve (AUROC) (95% Confidence Interval (CI) 0.913-9.22), followed by the random forest at 0.91 (95% CI 0.909-0.918), multilayer perceptron at 0.86 (95% CI 0.850-0.861), and logistic regression at 0.85 (95% CI 0.841-0.852). These AUROCs decreased to 0.78-0.82 and 0.74-0.80 when limited to patients not currently or never delirious, respectively.

Conclusions: A boosted tree machine learning model was able to identify hospitalized patients at elevated risk for delirium in the next 24 hours. This may allow for automated delirium risk screening and more precise targeting of proven and investigational interventions to prevent delirium.

(JMIR Preprints 15/05/2024:60442)
DOI: https://doi.org/10.2196/preprints.60442

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Kendrick M Shaw, MD PhD[1,2], Yu-Ping Shao, MS[3], Manohar Ghanta, MS[3,4], Valdery Moura Junior, MS MBA[3,4], Eyal Y Kimchi, MD PhD[5], Timothy T Houle, PhD[1,2], Oluwaseun Akeju, MD[1,2], M. Brandon Westover, MD PhD[2,3,4]

[1]Department of Anesthesia, Pain, and Critical care Medicine, Massachusetts General Hospital, Boston Massachusetts, USA

[2]Harvard Medical School, Boston Massachusetts, USA

[3]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston Massachusetts, USA

[4]Department of Neurology, Beth Israel Deaconess Medical Center, Boston Massachusetts, USA

[5]Ken & Ruth Davee Department of Neurology, Feinberg School of Medicine, Northwestern University, Chicago Illinois, USA

Corresponding author: Kendrick Shaw, MD PhD <kmshaw@mgh.harvard.edu> +1-216-256-1285, Department of Anesthesiology, Massachusetts General Hospital 55 Fruit Street, Boston, MA 02114

Support: KMS was supported by NIH (T32-GM007592) and MGH departmental funds. MBW was supported by the Glenn Foundation for Medical Research and American Federation for Aging Research (Breakthroughs in Gerontology Grant); American Academy of Sleep Medicine (AASM Foundation Strategic Research Award); Football Players Health Study (FPHS) at Harvard University; Department of Defense through a subcontract from Moberg ICU Solutions, Inc; NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, R01AG062989); and NSF (award SCH-2014431).

-*Author Contributions*:

-Kendrick M Shaw: This author helped with study design, data preparation, implementation of the models, analysis, and drafting the manuscript.

-Yu-Ping Shao: This author helped with data acquisition and preparation, and review of the manuscript.

-Manohar Ghanta: This author helped with data acquisition and preparation, and review of the manuscript.

-Valdery Moura Junior: This author helped with data acquisition and preparation, and review of the manuscript.

-Eyal Y Kimchi: This author helped with study design and review of the manuscript.

-Timothy T Houle: This author helped with study design, interpretation, and review of the manuscript.

-Oluwaseun Akeju: This author helped with study design, interpretation, and review of the manuscript.

-M. Brandon Westover: This author helped with the conception of the work, study design, interpretation, and review of the manuscript.

Daily automated prediction of delirium risk in hospitalized patients: Model development and validation

Abstract

Background: Delirium is common in hospitalized patients and correlated with increased morbidity and mortality. Despite this, delirium is underdiagnosed, and many institutions do not have sufficient resources to consistently apply effective screening and prevention.
Objective To develop a machine learning algorithm to identify patients at highest risk of delirium in the hospital each day in an automated fashion based on data available in the electronic medical record, reducing the barrier to large-scale delirium screening.

Methods: We developed and compared multiple machine learning models on a retrospective dataset of all hospitalized adult patients with recorded Confusion Assessment Method (CAM) screens at a major academic medical center from April 2nd, 2016 to January 16th 2019, comprising 23006 patients. The patient's age, gender, and all available laboratory values, vital signs, prior CAM screens, and medication administrations were used as potential predictors. Four machine learning approaches were investigated: logistic regression with L1-regularization, multilayer perceptrons, random forests, and boosted trees. Model development used 80% of the patients; the remaining 20% were reserved for testing the final models. Lab values, vital signs, medications, gender, and age were used to predict a positive CAM screen in the next 24 hours.

Results: The boosted tree model achieved the greatest predictive power, with a 0.92 area under the receiver operator characteristic curve (AUROC) (95% Confidence Interval (CI) 0.913-9.22), followed by the random forest at 0.91 (95% CI 0.909-0.918), multilayer perceptron at 0.86 (95% CI 0.850-0.861), and logistic regression at 0.85 (95% CI 0.841-0.852). These AUROCs decreased to 0.78-0.82 and 0.74-0.80 when limited to patients not currently or never delirious, respectively.

Conclusions: A boosted tree machine learning model was able to identify hospitalized patients at elevated risk for delirium in the next 24 hours. This may allow for automated delirium risk screening and more precise targeting of proven and investigational interventions to prevent delirium.

Keywords: delirium; prediction model; machine learning; boosted trees

Introduction

Delirium is a common condition in hospitalized patients, and has been been recognized as an independent risk factor for poor clinical outcomes, including mortality, institutionalization, and cognitive impairment following hospital discharge [1–3]. The US annual national costs attributable to delirium have been estimated to be as high as 152 billion dollars, rivaling costs attributable to diabetes and falls [4]. As a result, a basic assessment for delirium is recommended for all hospitalized patients aged 65 or older [5], and formal screening for delirium is recommended for critically ill patients [6].
Despite these recommendations, delirium frequently remains undiagnosed [7]. An automated delirium prediction tool could help address this, by alerting clinicians to at risk patients so that they could be more

carefully assessed for delirium. Such screening tools could also help focus interventions aimed at prevention of delirium (e.g. components of the hospital elder life program (HELP) [8]) and provide an enriched patient sample for future delirium prevention studies.

In particular, we intend to use an automated tool to identify hospitalized patients at our institution who are at high risk of delirium in the next 24 hours. These patients will then be visited by a member of a delirium service for further evaluation and identification of interventions which may reduce the patients risk of delirium. For this purpose near term risk (24 hour risk) is more useful than risk of delirium at some point during this hospitalization, and any history of prior or current delirium is relevant to identifying the patients at risk of ongoing delirium who should be seen (as reducing the duration of ongoing delirium is still likely to benefit the patient).

Although multiple prior delirium prediction tools have been described [9–12] (for a recent systematic review see [13]) most have properties that have limited their use as a tool to be applied daily to every patient in the hospital. Most prediction tools are designed to allow a risk score to be easily calculated by a clinician by hand, limiting the model's performance compared to larger models with more features and favoring features that are easy for a human to produce over those easily extracted from the medical record. In addition, most prior models were developed using datasets of only a few hundred to a few thousand patients, limiting the complexity of the models that could be developed without overfitting.

To address these limitations, we have developed a model that can provide automated delirium screening based on data readily available from the electronic health record, emphasizing predictive power over ease of manual computation or ease of interpretation. Because current and prior delirium are known risk factors for future delirium, we also explore the performance of the model in patients without these risk factors. This tool achieves state of the art accuracy for delirium prediction in this automated setting, and maintains good performance even when restricted to patients without current or without prior delirium.

Methods

This study was reviewed and approved by the Mass General Brigham institutional review board (IRB), approval number 2013P001024. The IRB determined that informed consent was not required for this retrospective study. This manuscript adheres to the applicable TRIPOD guidelines.

Study cohort

Data were obtained for all patients who received any variation of a Confusion Assessment Method (CAM) screen [14] (eg the CAM-ICU) [15]) in our hospital between April 2nd, 2016 and January 16th 2019, for a total of 23,006 patients. No specific exclusion criteria were used, as we wished the results to be applicable to the typical population of the hospital. Approximately 20% of patients (4511) were randomly selected and set aside for final evaluation of the model (the "test dataset"); we remained blinded to this data set until after all model choices and parameters had been fixed in preparation for publication. The remaining 80% of patients (the "training dataset") was used for model selection, model training, and hyperparameter tuning.

Model development overview

We provide an overview of the model development here; additional details can be found in Supplemental Text 1.

For the outcome to be predicted, we used presence of at least one positive CAM screen within a given day where CAM screens were performed. The CAM screen is a validated and widely used tool for assessing delirium where an observer assesses for a change in cognition with an acute onset and fluctuating course involving inattention and either disorganized thinking or an altered level of consciousness. [14] For each patient, we first identified all 24 hour intervals from 5am to 5am during which at least one CAM screen or CAM screen variation had been performed. For each such interval, the model was required to predict whether at least one CAM screen variant would be positive during that interval (vs all negative CAM screens).

As model inputs, we used the patient's age, gender, and all prior recorded vital signs, laboratory values, medications, and prior CAM assessments present in the medical record at 5 am before the 24 hours in which delirium was to be predicted. Categorical values were converted to integers (e.g "1" for "Positive", "0" for "Negative"). This data was reduced to summary statistics for each measurement (e.g. minimum, maximum, and mean systolic blood pressure in the past 24 hours), which were used to form fixed-length feature vectors for each prediction interval. These feature vectors were then normalized by subtracting the median and dividing by the interquartile interval, with both the median and the quartiles estimated by the P2 algorithm [16]. Because the P2 algorithm provides only an approximation of the quantiles, the resulting values were generally not exact integers even for categorical values (e.g. a binary measure that was mostly negative would have an estimated median that was slightly above 0). Features that were missing in more then 95% of the patients were discarded. The remaining missing values were imputed to be the (P2-estimated) median value for the feature.

The XGBoost library [17] was used to fit boosted tree models [18] for the cleaned and normalized training data sets. For comparison, random forest models [19] and logistic regression models using L1 regularization [20] were also fit to the data using scikit learn [21]. In addition, a deep neural network model was developed using TensorFlow [22]. The final network had a 32-node Rectified Linear unit (RELU) [23] input layer, two hidden layers of 16 and 8 RELU nodes respectively, and an output layer with a single sigmoidal node. All layers were fully connected, and a 50% dropout [24] was used between layers. For the logistic regression and random forest models, Platt scaling was used to improve calibration of the model. We used 10-fold cross validation [25] for hyperparameter tuning to minimize overfitting.

All development was done using the Ubuntu 18.04 distribution of Gnu-Linux. Data processing and analysis was performed using the Python [26] and Julia [27] programming languages. The code used to generate the models and figures is available at https://github.com/kms15/DeliriumPredictor2022a. The datasets used to develop and test the model contain personally identifiable health information and thus are not publicly available; the authors can be contacted for more information.

Statistical analysis

We provide an overview of the statistical analysis here; additional details can be found in Supplemental Text 2. The receiver operator characteristic curve (ROC) and the area underneath it (AUROC) were used to evaluate the performance of each model. To capture the effects of population prevalence on the performance, we also used performance-recall curves (PRC) and the area underneath them (AUPRC). Calibration curves were used to qualitatively evaluate model calibration, and the expected calibration error (ECE) and maximum calibration error (MCE) were used to quantify the degree of calibration. Confidence intervals were calculated in python using bootstraping with 1000 rounds, resampling by prediction day. To interpret the final behavior of the models, we used SHapley Additive exPlanations (SHAP) value estimation methods as described by Lundberg et al [28,29].

The final models were trained on the full training data set (80% of patients). Once the final models were trained the test dataset was unblinded and the model performance was measured on the test dataset. The performance of the cross-validated version of the models on the training dataset was similar to the performance of the final models on the test dataset and is not reported here.

Although current and prior delirium are useful predictors for our intended use of the model, these are well-known risk factors for delirium and it is thus useful to explore how the model performs on patients without these risk factors To test the model performance these populations with a lower initial probability of delirium, versions of the final models were trained and tested on only snapshots of patients who were not delirious (i.e. no positive CAM screens in the past 24 hours) or never delirious (no prior positive CAM screens). Conceptually, all patients start in the “never delirious” state, and then potentially transition to a “currently delirious” state and then potentially between this state and a “previously but not currently delirious” state. For a summary of the assignment of patient snapshots to these groups, please see Supplemental Figure 1.

Results

Of the 20,006 patients in the dataset, 4583 patients (19.9%) had at least one positive CAM screen (Table 1). The average age of the patients was 65 years old, and slightly higher in patients with a positive CAM screen (69.8). Approximately 54% of patients were male and 46% were female; no other genders were recorded in this dataset. The fraction of males was slightly higher (57%) in patients with a positive CAM screen. An average of 12.6 CAM screens were recorded per patient, with more (24.8) recorded for patients with a positive CAM screen than for patients with no positive CAM screens (9.5). An average of 8.3% of CAM screens per patient were positive, which rose to an average of 42% of CAM screens per patient that were positive in patients who had at least one positive CAM screen. CAM screens were performed on an average of 8.1 days per patient, of which 8.1 percent of days with CAM screens had at least one positive CAM screen.

Table 1: Patient demographics

Measure	All Patients	Training Set	Testing Set	One or more positive CAM	All CAM Negative
Number of patients	23006	18495	4511	4583	18423
Mean age (std	65.5	65.5	65.5	69.8	64.4 (17.4)

Measure	All Patients	Training Set	Testing Set	One or more positive CAM	All CAM Negative
dev)	(17.3)	(17.3)	(17.2)	(16.1)	
Male (%)	12502 (54.3)	10037 (54.3)	2465 (54.6)	2614 (57.0)	9888 (53.7)
Female (%)	10500 (45.6)	8455 (45.7)	2045 (45.3)	1968 (42.9)	8532 (46.3)
Patients with at least one positive CAM screen	4583 (19.9)	3656 (19.8)	927 (20.5)	4583 (100)	0 (0)
Mean CAM evaluations per patient (std dev)	12.6 (17.9)	12.5 (18.0)	12.9 (17.6)	24.8 (29.0)	9.5 (12.0)
Average percent of positive CAM screens per patient (std dev)	8.3 (22)	8.3 (22)	8.3 (22)	42 (33)	0 (0)
Mean CAM evaluation days per patient (std dev)	8.1 (11.1)	8.1 (11.3)	8.3 (10.5)	15.6 (18.0)	6.2 (7.6)
Average percent positive CAM days (std dev)	8.1 (21)	8.1 (21)	8.2 (21)	40.8 (31)	0 (0)

All models provided significant predictive power for delirium (a positive CAM screen in the next 24 hours) when applied to all hospitalized patients in the dataset (Figure 1). The boosted tree model had the highest AUROC at 0.92 (95% confidence interval (95% CI) 0.913-9.22). This was followed by the random forest model, the multi-layer perceptron, and the logistic regression model with L1-regularization (with an AUROC of 0.85 (95% CI 0.841-0.852)).

Figure 1: Receiver operator characteristic (ROC) curves for different model types (rows) and patient subsets (columns) showing the true positive rate (i.e. recall) as a function of the false positive rate. The thin light gray region around the line shows the bootstrap 95% confidence interval.

The models were then retrained and evaluated with patients who were not currently delirious (most recent CAM screen was negative) and with patients that had no history of delirium (no prior positive CAM screens). Although the models did not perform as well on these more difficult subsets, they still provided good predictive power. The boosted tree model declined from an AUROC of 0.92 to an AUROC of 0.82 (95% CI

0.815-0.834) and 0.80 (95% CI 0.79-0.81) when limited to patients who were not currently delirious and patients with no prior delirium, respectively. The other models showed a similar decrement, with the AUROC decreasing to 0.77-0.81 and 0.74-0.77 when limited to not currently delirious and never delirious patients, respectively. The boosted tree model outperformed the other models in all three patient groups.

The models significantly varied in their ability to maintain a high positive predictive value as sensitivity was increased (Figure 2). The Boosted Tree model performed well with an AUPRC of 0.73 (95% CI 0.72-0.75), declining significantly to an AUPRC of 0.32 (95% CI 0.30-0.34) for patients who are not currently delirious and 0.22 (95% CI 0.20-0.25) with no prior delirium. The incidence of delirium in all patients was 13%, not delirious patients 6%, and never delirious patients 4%; thus the decrement in AUPRC appears to be largely driven by the decreased incidence in these subgroups. While the random forest model performs relatively well with an AUPRC of 0.70 (95% CI 0.68-0.71), it also suffers significant decrements in performance with patient who are not currently not delirious (AUPRC 0.25) or have no history of delirium (AUPRC 0.14). The multilayer perceptron models perform somewhat worse than the tree-based models, with AUPRCs of 0.50, 0.23, and 0.15 in all patients, not currently delirious, and no prior delirium groups. Logistic regression performed similarly to the multilayer perceptron with AUPRCs of 0.48, 0.22, and 0.17.

Figure 2: Precision-recall curves for different model types (rows) and patient subsets (columns) showing precision (i.e. positive predictive value) as a function of recall (i.e. true positive rate). The grey reigion indicates the bootstrap 95% confidence interval.

We next investigated the calibration of the prediction models. All the models were well calibrated (ECE ≤ 0.02, MCE ≤ 0.11) when applied to all hospitalized patients (Figure 3), with the exception of logistic regression model which over-estimated the risk of delirium in the highest-scored group (ECE 0.03, MCE 0.28). The boosted tree model and the random forest model both identified a larger number of high-risk patients while still maintaining good calibration in this higher risk group. When restricted to patients who were not currently delirious or to patients with no prior delirium, all models classified very few patients as high risk (consistent with the earlier PRCs), and the random forest did not assign higher probabilities to any patients in these subgroups.

Figure 3: Reliability diagrams for different model types (rows) and patient subsets (columns) showing the actual fraction of patient snapshots with delirium for groups with a given predicted risk of delirium (blue squares, left y-axis). Error bars show the bootstrap 95% confidence interval. The grey bars in the background show the number of patient snapshots in each predicted probability bin (y-axis on right). ECE is the expected calibration error, with 95% confidence interval. MCE is the maximum calibration error, with 95% confidence

interval.

We finally turn to an examination of the features influencing the predictions of the most successful model (the boosted tree model). Ordering the features by average SHAP magnitude (Figure 4), we first note that the range of SHAP values for any of the top 40 features is smaller than the range of SHAP values for the sum of the remaining 1901 features; thus the predictions of the model across can not easily be simplified to a small number of driving features that are the same for all patients. The features with the highest average SHAP magnitude appear to fall into several known risk factors for delirium. Current and prior delirium is a known predictor of future delirium, and 6 of the top 40 features relate to prior CAM screen (including 4 of the top 5 features). Of note, the model considers a patient with no prior CAM screens to be at higher risk than a patient with prior negative CAM screens, and this feature remains important even for the “no prior delirium” case (not shown). Antipsychotic administration is also as expected a risk-predicting feature, as it is often used to convert hyperactive delirium to hypoactive delirium. The majority of the top 40 risk features fall into other categories, however, which include known risk factors such as age, liver failure (e.g. AST, Ammonia levels and HCV), infection (e.g. WBC, Monocytes, and Cefepime (which is also neurotoxic)), and malnutrition/frailty (amino acid supplementation, albumin levels).

Figure 4: SHAP beeswarm plots[28,29] of the 40 features with the highest SHAP magnitude for patients in the holdout dataset. Each dot shows a single prediction for a patient, the color of the dot indicates how high (red) or low (blue) the feature was for this patient, and the horizontal position of the dot indicates the relative effect of this feature on the predicted risk for the given patient.

Discussion

Principal Results

We have described the development of a prediction model that can provide automated daily predictions of the risk of delirium for a general population of hospitalized patients. We found that a boosted tree model performed best for this dataset and was able to identify a group of high-risk patients even when limited to patients that were not delirious or had no prior history of delirium. However, the random forest, multi-layer perceptron, and logistic regression models, while less effective, still provided substantial predictive power. All of the models showed good calibration on the full dataset, but showed poorer maximum calibration error when applied to subsets of the data where delirium was less common - patients who were not currently delirious and patients with no prior history of delirium.

The relative performance of the various models may reflect the relative match between the flexibility of each of the types of models and the size of the dataset we used. Although the L1 regularization used for the logistic regression model allowed for some tuning of flexibility by adjusting how many features were used, the logistic model can only capture monotonic relationships. Both boosted tree models and random forest models can better capitalize on non-monotonic relationships which likely underlies their better performance on this dataset. In contrast, it was difficult to prevent overfitting with the multi-layer perceptron model; preventing this overfitting would likely require either a larger dataset or additional methods of regularization.

Limitations

All of the models showed a decrement in performance when restricted to patients who are not currently delirious and a further decrement when restricted without prior delirium. Except for the multilayer perceptron in the no prior delirium case, however, the AUROC of all of the models remained greater than 0.75, which many would consider the threshold for “good” performance of a clinical test [30]. The boosted tree model, in particular, maintained an AUROC of 0.80, which compares very favorably to commonly used diagnostic tests such as D-dimer levels in the setting of a suspected pulmonary embolism (with a reported AUROC of 0.71[31]). The decrement of performance in these subgroups likely reflects the increased difficulty of the task - those patients with significant risk factors are likely to have had delirium on prior days or hospitalizations and most of those who remain are relatively unlikely to become delirious in the next 24 hours. Even within the patients with no prior delirium, however, the boosted tree model was able to identify high-risk patients who would likely warrant further evaluation and interventions - for example from the precision-recall curves (figure 2) we can see that even if one were to demand a 50% true-positive threshold for an intervention, the model would correctly identify over 10% of the never delirious patients who would become delirious as candidates for that intervention.

Comparison with Prior Work

Multiple prior prediction models for delirium have been developed for use in intensive care unit patients [11,12,32] and in hospitalized elderly patients [9]. In general, these models use a small number of predictors (4-11) identified using logistic regression, including such predictors as age, history of cognitive impairment, history of alcohol abuse, respiratory failure, blood urea nitrogen, mean arterial pressure, use of corticosteroids, admission category, admission urgency, and vision impairment. Our prediction goal (predicting a positive CAM screen within the next 24 hours) is somewhat different than the existing models we are aware of, as it is aimed at the specific task of helping determine which hospitalized patients should be seen by a delirium service that day. With that caveat, the performance of our model appears to compare well with other models on similar tasks. Chua et all [13] provides a good review of similar models; reported AUROCs in this review range from 0.71 to 0.91 (compared to 0.92 for the boosted tree model we report). As delirium is an infrequent event, however, the AUPRC may provide a better estimate of the model’s performance as a screening tool. Comparing our model to the best performing models in the review by Chua et al[13], only the model described by Corradi et al [33] (one of the two models with an AUROC of

0.91) reported a AUPRC, which was 0.60 (compared to 0.73 for the boosted tree model we report). One of the challenges machine learning has faced in medicine is translating predictions into improvements in patient outcomes[34]. We plan to use this model will be used to screen all of the patients in a 1000 bed hospital (which would be prohibitively labor intensive to do by hand), and identify a set of high-risk patients to be visited by a delirium service. The members of this delirium service will then evaluate patients and provide recommendations to the team caring for the patient on how to reduce that patient's risk of delirium. By focusing this additional clinical effort and possible interventions on the patients who would most likely benefit from them, we hope to use this tool to improve care at a lower cost per patient than providing the same interventions to every patient (including those at much lower risk of delirium).

Because of this intended use, we have made trade-offs that may limit the use of this model in other contexts. For example, our focus was on maximizing the ability of the model to identify high-risk hospitalized patients rather than on identifying the causal mechanism for a given patient's delirium. Thus, for example, an ABG showing mild hyperoxia might be used for prediction by the model because it is correlated with intubation, sedation, and critical illness rather than because it is directly increasing the patient's risk for delirium, and blindly attempting to correct this laboratory value may not decrease the patient's risk of delirium. In addition, some features, such as administering an anti-psychotic medication, may happen to treat an agitated delirium that has not been documented in the medical record; while this may still be quite useful for a delirium service, it may be less useful for a responding clinician who is treating the agitation who likely already knows he or she is treating a symptom of delirium. Thus while the model works well for finding high- risk patients, interpreting these risk factors and identifying appropriate interventions will still require clinical expertise.

Although our model can be applied in its current form, there are limitations that a user will need to be mindful of. For some machine learning models, such as logistic regression, it can be relatively easy to understand a prediction from the model in terms of the individual features contributing to the prediction. For many others, however, including boosted tree models such as our best-performing model, the non-linear interaction of many features can make it difficult to understand why a given patient was assigned a high or low risk score. Providing interpretability for these more complex, non-linear models is an active area of research in machine learning, and while tools such as SHAP can provide some insight into a model, for some uses a less accurate but more interpretable model (such as logistic regression) may be preferred.

Another limitation of our model is that many of the risk factors such as sleep disruption may only be documented in clinical notes and not in the structured data we have used, and thus high-risk patients may be missed by the model. We hope to address this in future work by integrating natural language processing techniques into the model.

The dataset used for development and validation of the model is another potential source of bias in this study. Because the patients from this study are only those patients from a single academic medical center who received delirium screens, they may not be reflective of patients in other settings, and potentially not even representative to patients at the same institution who did not receive delirium screens. Although this could raise concerns that this would bias our dataset towards patients with delirium, this does not appear to have been the case. Only 4583 of the 23006 patients (see Table 1), or about 20%, of the patients in our dataset had one or more positive CAM screens. This is consistent with the 23% (CI 19-26%) incidence of delirium reported in the meta-analysis of estimates of delirium occurrence reported by Gibb et al[35]. Although this is reassuring, future studies will be needed to provide external validation of the model at other institutions and on other patient populations.

Conclusions

In this paper we have described a method for predicting delirium in hospitalized patients given information already present in the electronic medical record. A large dataset of over twenty-three thousand patients allowed us to consider a larger number of candidate features while still allowing for rigorous validation with a blinded test dataset. The resulting model provides both good accuracy and good calibration and can be run in an automated fashion on data in the electronic patient record without requiring additional human effort. We believe this model can be of use in guiding clinicians and researchers in focusing on patients at greatest risk of delirium in hopes of mitigating the morbidity and mortality associated with this disease.

Acknowledgments

KS was supported by NIH (T32-GM007592). MBW was supported by the Glenn Foundation for Medical Research and American Federation for Aging Research (Breakthroughs in Gerontology Grant); American Academy of Sleep Medicine (AASM Foundation Strategic Research Award); Football Players Health Study (FPHS) at Harvard University; Department of Defense through a subcontract from Moberg ICU Solutions, Inc; NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, R01AG062989); and NSF (award SCH-2014431).

Conflicts of Interest

Dr. Westover is co-founder of Beacon Biosignals, which had no role in this work. All authors report no potential conflicts of interest.

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operator characteristic curve
CAM: confusion assessment method
CI: confidence interval
ECE: expected calibration error
EMR: electronic medical record
HELP: hospital elder life program
IRB: institutional review board
MCE: maximum calibration error
PRC: probability-recall curve
RELU: rectified linear unit
ROC: receiver operator characteristic curve
SHAP: shapley additive explanations
TRIPOD: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

References

1. Siddiqi N, House AO, Holmes JD. Occurrence and outcome of delirium in medical in-patients: A systematic literature review. Age and Ageing 2006 Jul;35(4):350–364. doi: 10.1093/ageing/afl005X

2. Witlox J, Eurelings LSM, Jonghe JFM de, Kalisvaart KJ, Eikelenboom P, Gool WA van. Delirium in Elderly Patients and the Risk of Postdischarge Mortality, Institutionalization, and Dementia: A Meta-analysis. JAMA 2010 Jul;304(4):443–451. doi: 10.1001/jama.2010.1013X

3. Wilson JE, Mart MF, Cunningham C, Shehabi Y, Girard TD, MacLullich AMJ, Slooter AJC, Ely EW. Delirium. Nature Reviews Disease Primers 2020 Nov;6(1):1–26. doi: 10.1038/s41572-020-00223-4X

4. Leslie DL, Marcantonio ER, Zhang Y, Leo-Summers L, Inouye SK. One-Year Health Care Costs Associated With Delirium in the Elderly Population. Archives of Internal Medicine 2008 Jan;168(1):27–32. doi: 10.1001/archinternmed.2007.4X

5. Delirium: Prevention, diagnosis and management in hospital and long-term care. London: National Institute for Health; Care Excellence (NICE); 2023. PMID:31971702ISBN:978-1-4731-4953-3X

6. Barr J, Fraser GL, Puntillo K, Ely EW, Gélinas C, Dasta JF, Davidson JE, Devlin JW, Kress JP, Joffe AM, Coursin DB, Herr DL, Tung A, Robinson BRH, Fontaine DK, Ramsay MA, Riker RR, Sessler CN, Pun B, Skrobik Y, Jaeschke R. Clinical Practice Guidelines for the Management of Pain, Agitation, and Delirium in Adult Patients in the Intensive Care Unit. Critical Care Medicine 2013 Jan;41(1):263. doi: 10.1097/CCM.0b013e3182783b72X

7. Patel R, Gambrell M, Speroff T, Scott T, Pun B, Okahashi J, Strength C, Pandharipande P, Girard T, Burgess H, Dittus R, Bernard G, Ely E. Delirium and Sedation in the Intensive Care Unit (ICU): Survey of behaviors and attitudes of 1,384 healthcare professionals. Critical care medicine 2009 Mar;37(3):825–832. PMID:19237884X

8. Hshieh TT, Yang T, Gartaganis SL, Yue J, Inouye SK. Hospital Elder Life Program: Systematic Review and Meta-analysis of Effectiveness. The American Journal of Geriatric Psychiatry 2018 Oct;26(10):1015–1033. doi: 10.1016/j.jagp.2018.06.007X

9. Inouye SK, Viscoli CM, Horwitz RI, Hurst LD, Tinetti ME. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. Annals of Internal Medicine 1993 Sep;119(6):474–481. PMID:8357112X

10. Boogaard M van den, Schoonhoven L, Maseda E, Plowright C, Jones C, Luetz A, Sackey PV, Jorens PG, Aitken LM, Haren FMP van, Donders R, Hoeven JG van der, Pickkers P. Recalibration of the delirium prediction model for ICU patients (PRE-DELIRIC): A multinational observational study. Intensive Care Medicine 2014 Mar;40(3):361–369. PMID:24441670X

11. Wassenaar A, Boogaard M van den, Achterberg T van, Slooter AJC, Kuiper MA, Hoogendoorn ME, Simons KS, Maseda E, Pinto N, Jones C, Luetz A, Schandl A, Verbrugghe W, Aitken LM, Haren FMP van, Donders ART, Schoonhoven L, Pickkers P. Multinational development and validation of an early prediction model for delirium in ICU patients. Intensive Care Medicine 2015 Jun;41(6):1048–1056. doi: 10.1007/s00134-015-3777-2X

12. Chen Y, Du H, Wei B-H, Chang X-N, Dong C-M. Development and validation of risk-stratification delirium prediction model for critically ill patients: A prospective, observational, single-center study. Medicine 2017 Jul;96(29):e7543. PMID:28723773X

13. Chua SJ, Wrigley S, Hair C, Sahathevan R. Prediction of delirium using data mining: A systematic review. Journal of Clinical Neuroscience 2021 Sep;91:288–298. doi: 10.1016/j.jocn.2021.07.029X

14. Inouye SK, Dyck CH van, Alessi CA, Balkin S, Siegal AP, Horwitz RI. Clarifying Confusion: The Confusion Assessment Method. Annals of Internal Medicine 1990 Dec;113(12):941–948. doi:

10.7326/0003-4819-113-12-941X
15. Ely EW, Inouye SK, Bernard GR, Gordon S, Francis J, May L, Truman B, Speroff T, Gautam S, Margolin R, Hart RP, Dittus R. Delirium in mechanically ventilated patients: Validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU). JAMA 2001 Dec;286(21):2703–2710. PMID:11730446X
16. Jain R, Chlamtac I. The P2 algorithm for dynamic calculation of quantiles and histograms without storing observations. Communications of the ACM 1985 Oct;28(10):1076–1085. doi: 10.1145/4372.4378X
17. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–794. doi: 10.1145/2939672.2939785X
18. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 2001 Oct;29(5):1189–1232. doi: 10.1214/aos/1013203451X
19. Breiman L. Random Forests. Machine Learning 2001 Oct;45(1):5–32. doi: 10.1023/A:1010933404324X
20. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 1996;58(1):267–288. doi: 10.1111/j.2517-6161.1996.tb02080.xX
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 2011 Nov;12(null):2825–2830. doi: 10.5555/1953048.2078195X
22. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. TensorFlow: A System for Large-scale Machine Learning. Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation Berkeley, CA, USA: USENIX Association; 2016. p. 265–283. doi: 10.5555/3026877.3026899X
23. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Commun ACM 2012;60:84–90. doi: 10.1145/3065386X
24. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014 Jun;15:1929–1958. doi: 10.5555/2627435.2670313X
25. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society: Series B (Methodological) 1974;36(2):111–133. doi: 10.1111/j.2517-6161.1974.tb00994.xX
26. Rossum G van, Boer J de. Linking a stub generator (AIL) to a prototyping language (Python). Proceedings of the Spring 1991 EurOpen Conference, Troms, Norway 1991. p. 229–247.
27. Bezanson J, Edelman A, Karpinski S, Shah V. Julia: A Fresh Approach to Numerical Computing. SIAM Review 2017 Jan;59(1):65–98. doi: 10.1137/141000671X
28. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems Curran Associates, Inc.; 2017.
29. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2020 Jan;2(1):56–67. doi: 10.1038/s42256-019-0138-9X
30. Jones CM, Athanasiou T. Summary Receiver Operating Characteristic Curve Analysis Techniques in the Evaluation of Diagnostic Tests. The Annals of Thoracic Surgery 2005 Jan;79(1):16–20. doi: 10.1016/j.athoracsur.2004.09.040X
31. Fu Z, Zhuang X, He Y, Huang H, Guo W. The diagnostic value of D-dimer with simplified Geneva score (SGS) pre-test in the diagnosis of pulmonary embolism (PE). Journal of Cardiothoracic Surgery 2020 Jul;15:176. PMID:32690039X

32. Boogaard M van den, Pickkers P, Slooter AJC, Kuiper MA, Spronk PE, Voort PHJ van der, Hoeven JG van der, Donders R, Achterberg T van, Schoonhoven L. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: Observational multicentre study. BMJ (Clinical research ed) 2012 Feb;344:e420. PMID:22323509X

33. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of Incident Delirium Using a Random Forest classifier. Journal of Medical Systems 2018 Dec;42(12):261. doi: 10.1007/s10916-018-1109-0X

34. Marwaha JS, Kvedar JC. Crossing the chasm from model performance to clinical impact: The need to improve implementation and evaluation of AI. npj Digital Medicine 2022 Mar;5(1):1–2. doi: 10.1038/s41746-022-00572-2X

35. Gibb K, Seeley A, Quinn T, Siddiqi N, Shenkin S, Rockwood K, Davis D. The consistent burden in published estimates of delirium occurrence in medical inpatients over four decades: A systematic review and meta-analysis study. Age and Ageing 2020 May;49(3):352–360. doi: 10.1093/ageing/afaa040X

Supplementary Files

Multimedia Appendixes

Details of the model development.
URL: http://asset.jmir.pub/assets/08ba2cfbb8b04bbeb033f75b8f103894.docx

Details of the analysis.
URL: http://asset.jmir.pub/assets/d987152f76ffea85281ab0e25f411fc7.docx

Example report for a given patient.
URL: http://asset.jmir.pub/assets/be13773aad9f1266e077050468bc594e.docx

Supplemental figure 1: flowchart showing assignment of delirium status.
URL: http://asset.jmir.pub/assets/9d995040801508a44dba57a513992a5c.docx