Published in final edited form as:

Acta Psychiatr Scand. 2023 May ; 147(5): 493–505. doi:10.1111/acps.13551.

Evaluating the Performance of Machine Learning Methods for Risk Estimation of Delirium in Patients Hospitalized from the Emergency Department

Brianna Mueller [PhD candidate],
Tippie College of Business, The University of Iowa

W. Nick Street, PhD [Professor],
Tippie College of Business, The University of Iowa

Ryan M. Carnahan, PharmD, MS [Professor],
The University of Iowa College of Public Health, Department of Epidemiology

Sangil Lee, MD, MS [Clinical Associate Professor]
Department of Emergency Medicine, The University of Iowa

Abstract

Introduction: Delirium is a cerebral dysfunction seen commonly in the acute care setting. It is associated with increased mortality and morbidity and is frequently missed in the emergency department (ED) and inpatient care by clinical gestalt alone. Identifying those at risk of delirium may help prioritize screening and interventions in the hospital setting.

Objective: Our objective was to leverage electronic health records to identify a clinically valuable risk estimation model for prevalent delirium in patients being transferred from the ED to inpatient units.

Methods: This was a retrospective cohort study to develop and validate a risk model to detect delirium using patient data available from prior visits and ED encounter. Electronic health records were extracted for patients hospitalized from the ED between January 1, 2014, and December 31, 2020. Eligible patients were aged 65 or older, admitted to an inpatient unit from the emergency department, and had at least one DOSS assessment or CAM-ICU recorded within 72 hours of hospitalization. Six machine learning models were developed to estimate the risk of delirium using clinical variables including demographic features, physiological measurements, medications administered, lab results, and diagnoses.

Results: A total of 28,531 patients met the inclusion criteria with 8,057 (28.4%) having a positive delirium screening within the outcome observation period. Machine learning models were compared using the area under the receiver operating curve (AUC). The gradient boosted machine achieved the best performance with an AUC of 0.839 (95% CI, 0.837 - 0.841). At a 90% sensitivity threshold, this model achieved a specificity of 53.5% (95% CI 53.0% - 54.0%) a positive predictive value of 43.5% (95% CI 43.2% - 43.9%), and a negative predictive value of 93.1% (95% CI 93.1%-93.2%). A random forest model and L1-penalized logistic regression also demonstrated notable performance with AUCs of 0.837 (95% CI, 0.835-0.838) and 0.831 (95% CI, 0.830-0.833) respectively.

Conclusion: This study demonstrated the use of machine learning algorithms to identify a combination of variables that enables an estimation of risk of positive delirium screens early in hospitalization to develop prevention or management protocols.

INTRODUCTION

Delirium is a global cerebral dysfunction seen in 8% up to 64% of patients in the acute care setting. The prevalence of delirium in the emergency department (ED) and inpatient populations is surprisingly high.[1-3] The presence of delirium is associated with a prolonged hospital stay, a higher likelihood of skilled nursing facility placement, and a 2- to 4-fold increase in mortality.[4] Despite the mortality rate being comparable to myocardial infarction, the fluctuating nature of symptoms, uncertainty of baseline cognitive function, and limited diagnostic modality lead to diagnostic dilemmas. By clinical gestalt alone, providers miss up to 80% of patients experiencing delirium upon presentation to the ED, and missed delirium cases often went undetected even after hospitalization.[5, 6]

Unfortunately, delirium continues to be underdiagnosed and undertreated.[5] Although several cognitive assessment tools exist, they require additional training and dissemination. [7-9] Early screening and interventional options are emerging and seem promising, as reported by several recent studies.[10-12] There is a need to identify an optimal screening strategy for delirium beyond cognitive assessment because, until we have it, delirium will likely remain an elusive diagnosis. In recent years, numerous studies have developed models in the hospital setting for estimating the risk of delirium in postoperative patients[13-22] and ICU patients ,[23-28] but limited work has focused on the ED patient population.[3, 29, 30]

Our objective was to leverage electronic health records to identify a clinically valuable risk estimation model for positive delirium screens in patients admitted from the ED to inpatient units. An accurate risk estimation model derived from variables available around the time of the ED visit could be a solution to identifying patients who are at risk for or already experiencing delirium and may benefit the most from screening, preventive, and proactive management measures to improve prevention and management of delirium during hospital stays.

METHODS

Study Design

This was a retrospective cohort study using patient data from a tertiary care medical center. Electronic health record data for patients hospitalized from the ED between January 1, 2014, and December 31, 2020, was extracted by The Institute for Clinical Translational Science Data Warehouse. The study was approved by the local institutional review board (IRB), and we adhered to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement recommendations.[31]

Study setting

As a tertiary care academic medical center, the study site receives referrals and transfers from the entire state of Iowa, in addition to other states. The facility has a trauma and stroke centers and provides specialty care. Regular medical care is also provided.

Participants

The study population comprised patients hospitalized from an academic center ED with approximately 60,000 visits a year. At this institution, a nursing protocol was in place to screen hospitalized patients aged 65 and older for delirium twice daily from the time of hospitalization until discharge. Non-intubated patients were screened using the Delirium Observation Screening Scale (DOSS), a 13-item scale based on nurse observations.[32, 33] If a patient was ventilated, the Confusion Assessment Method for the ICU (CAM-ICU) was used.[34]

Eligible patients were aged 65 or older, admitted to an inpatient unit from the ED, and had at least one DOSS assessment or CAM-ICU recorded within the first 72 hours of hospitalization. If a patient was hospitalized more than once between January 2014 and December 2020, one encounter was randomly selected.

Outcome

The outcome measure of this study was a positive delirium screen within 72 hours of hospitalization determined by a DOSS score of 3 or greater or a positive CAM-ICU assessment. Both of these assessments are done by the trained nursing staff twice daily in the hospital units, and the sensitivity and specificity of these tests are reported to be above 90% in the literature, including a validation study of DOSS assessments done by nurses at the current study site.[33, 34] Observations within 72 hours were selected to optimize the model's ability to identify delirium cases most related to factors observed around the time of the ED visit.

Input variables

We collected variables recorded during the ED encounter including demographic features, physiological measurements, medications administered, lab results, and diagnoses given by ICD-10 codes, as well as a patient's history of diagnoses from prior visits if available. The physiological variables collected at the time of ED evaluation included heart rate, respiratory rate, body mass index (BMI), and temperature. Medications were obtained with drug flags for opioids, antibiotics, and benzodiazepines. We defined the anticholinergic variable as receipt of drugs classified as level 2 or 3 on an updated version of the Anticholinergic Drug Scale (Supplemental table 1 displays anticholinergics received by the sample).[35, 36]

We collected diagnoses given by ICD-10 codes from the record using the same approach. For dimensionality reduction, we applied the Clinical Classifications Software Refined (CSSR) to aggregate ICD-10 codes into clinically meaningful categories.[37] We also created a variable to distinguish ICD-10 codes corresponding to dementia from the broader CSSR category of neurocognitive disorders (Supplemental table 2 displays ICD-10 codes for dementia). ICD-10 codes corresponding to delirium and recorded during the

encounter of interest were excluded and CSSR variables observed in less than 0.5% of the study population were removed (Supplemental table 3 displays ICD-10 codes for delirium). Diagnoses from the current hospitalization of interest and diagnoses from past hospitalizations or outpatient visits were treated separately during model development, with a lookback period of one year considered for past diagnoses.

Medications and diagnoses were transformed into binary variables and labs were categorized as follows: below the normal range, within the normal range, and above the normal range.

Analysis

Variables with missing values were temperature (0.05%), respiratory rate (0.94%), heart rate (1.05%), and BMI (2.39%). Missing values were imputed with KNN-imputation which has been shown to outperform other widely used imputation methods.[38] Possible outliers were identified with the IQR Extreme Value analysis and treated as missing values. Variance inflation factors were observed to detect multicollinearity, and continuous variables were checked for linearity with the Box-Tidwell test. The continuous variables BMI, temperature, and respiration rate were discretized into 3 categories.

We compared the performance of six machine learning models using the Python Machine Learning library Scikit-Learn.[39] The algorithms included Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and K Nearest Neighbor (KNN). Cross-validation was implemented for both hyperparameter tuning and model evaluation with AUC as the evaluation metric. This re-sampling method was selected over repeated sub-sampling to prevent any loss of information about the positive class by ensuring every observation appears in both the training and test data.

To avoid an optimistic bias that can result from using the same cross-validation procedure for both hyperparameter tuning and model evaluation, we employed nested cross-validation. In nested cross-validation, k-fold cross-validation for hyperparameter tuning is nested inside the k-fold cross-validation for model evaluation. Using tenfold nested cross-validation, the data was randomly divided into 10 equally-sized subsets. Out of the 10 sets, 9 were used to train the classifier, and the 10th was used for testing. The training set was further partitioned into 5 folds for an inner cross-validation grid search to optimize hyperparameters. This process was repeated until each of the 10 subsets had served as the test set. Similar to a regular cross-validation procedure, evaluation metrics are obtained by averaging the test set scores of the 10 runs. By conducting model selection independently in each trial of the model fitting procedure, the risk of overfitting during hyperparameter tuning is reduced. The final models were selected using a 10-fold cross-validation grid search on all available data.

Sample Size

To examine the importance of sample size on model efficacy, the performance of a gradient boosted model was tested at various sample sizes. Subsets of the data were created with stratified random sampling. At each sample size, 10-fold cross validation with 10 repeats was employed to obtain an AUC estimate and standard error. AUC estimates and confidence intervals were plotted against sample size. Increasing the sample size from 11,341 to 28,351

resulted in a .01 AUC increase. This suggests a sample size larger than the sample used in the analysis would not result in a significant improvement in model performance. However, as the sample size increased, the standard error decreased which creates more confidence in the AUC estimate. (Supplemental figure 1)

RESULTS

Participants

We identified a total of 70,550 unique ED encounters for patients aged 65 and older during the study period and of these encounters; 44,722 resulted in hospitalization. An additional 1,104 encounters did not have records of at least one DOSS or CAM-ICU assessment recorded within 72 hours of a hospitalization, or death, reducing the study population to 43,618. After randomly selecting one encounter for patients with multiple encounters, the remaining 28,351 encounters met all study criteria (26,750 [94.4%] white; 14,430 [50.9% male]; mean age of 76.3 [SD 8.1]). Of the study participants, 8,057 (28.4%) had at least one positive delirium screen within 72 hours of hospitalization (Figure 1). We reported summary statistics for the population demographics and the length of hospital stay by positive delirium screen status (Table 1). The median length of hospitalization for the study participants was 3.8 days, further justifying the decision to limit the outcome measure to delirium screening within 72 hours. During the first 72 hours of hospitalization, subjects with positive screenings received 5.42 delirium screens on average while subjects with no positive screenings received 5.38 delirium screens on average.

Model Performance

Figure 2 illustrates the performance of each model with receiver operating characteristic (ROC) curves (Figure 2 and Table 2). AUC scores ranged from 0.767 to 0.839 with gradient boosted machine (GBM), random forest (RF), and logistic regression (LR) demonstrating the best performance with respective AUCs of .839, 0.837, and 0.831. There was no statistically significant difference in AUC between GBM and RF but both models outperformed the LR model at the 95% confidence level. We present model performance at various probability cutoffs for a positive diagnosis to assess the tradeoffs between sensitivity and specificity. In the discussion section, we consider different use cases where one performance metric may be prioritized over the other. Setting sensitivity at 90%, GBM achieves a specificity of 53.5% (95% CI 53.0% - 54.0%), a positive predictive value of 43.5% (95% CI 43.2% - 42.8%), and a negative predictive value of 93.1% (95% CI 93.1%-93.2%). At this threshold, 41.2% of patients would be classified as low risk and excluded from screening or preventive interventions. Of the 41.2%, 2.8% of patients would be missed cases of delirium. At the 90% specificity threshold, GBM achieves a sensitivity of 55.6% (95% CI 55.2% - 56.0%), a positive predictive value of 68.9% (95% CI 68.7% - 69.1%), and a negative predictive value of 83.6% (95% CI 83.5%-83.8%). At this threshold 77.0% of patients would be considered low risk, resulting in 12.6% of patients being missed cases of delirium. One might also consider the Youden index to identify an optimal threshold, at which GBM achieves a sensitivity of 75.9% (95% CI 75.8% - 76.1%), a specificity of 76.0% (95% CI 75.8% - 76.1%), a positive predictive value of 55.6% (95% CI 55.4% - 55.9%) and a negative predictive value of 88.8% (95% CI 88.7% - 88.9%).

Performance metrics for all classifiers at the various thresholds are presented in Table 3 and Table 4.

Model Specification

From the initial 641 variables, GMB selected 605 input variables, RF selected 598, and LR selected 250. Table 5 reports the odd ratios and confidence intervals for the LR variables with the highest coefficients. The most relevant risk factors identified by GBM and RF are provided in Figure 3 and Figure 4. Instead of calculating coefficients, tree-based models such as GBM report the relative importance of each feature during model development. Feature importance is a measure of the average decrease in node impurity (or class heterogeneity) observed when the data is split based on the values of that feature. In Table 6, the top 30 variables selected by GBM are summarized with variable frequencies by delirium screening outcome. We also note which variables were also selected as the top 30 variables for RF and LR. There were many variables highly ranked across all 3 models including dementia, dysphagia, urinary tract infections, acute hemorrhagic cerebrovascular disease, age, epilepsy, and traumatic brain injuries.

DISCUSSION

This study leveraged electronic health records to develop and validate a machine learning model to identify patients at increased risk of a positive delirium screen during the first three days of hospitalization using data from ED visits and early in hospitalization. The prevalence of delirium in the emergency department (ED) is estimated to be 8%–17%, but it is difficult to identify without an effective screening process which leads to underdiagnosed and undertreated delirium. [1-3, 5] In the literature, the prevalence of delirium in inpatient care units is 18%–64%, which further underscores the need for prevention and treatment strategies prior to hospitalization.[40] Our study of patients hospitalized via the ED was consistent with the literature, with the rate of positive delirium screens reaching 28.4% within the first 72 hours of hospitalization.

Although several cognitive assessment tools exist, they require training and dissemination, and compliance can be limited without strong merit and support from leadership and stakeholders.[7-9] Emerging early screening and interventional options seem promising, as reported by several recent studies, but these programs depend on effective screening.[10-12] If resources are not available for screening or implementing preventive or proactive management interventions for delirium in all patients, our machine learning approach might facilitate the selection of appropriate candidates. A 90% sensitivity threshold may be ideal if screening is for a low intensity, low risk preventative program that is not resource intensive. However, if the primary goal is to make screening more efficient by excluding more patients from resource intensive interventions, a higher specificity threshold may be desirable. A high specificity threshold might be used to trigger intervention without screening, but that threshold would miss many patients with delirium. Those with at least intermediate risk who did not meet this high specificity threshold may benefit the most from screening.

The fluctuating nature of delirium poses a challenge to clinicians, as evidenced by Lewis et al., who reported that the estimated misdiagnosis rate of delirium in the ED was up to 80%

in the 1990s, and the rate of misdiagnosis has remained high.[5, 29] Clinical gestalt is limited without any additional diagnostic modality or assessment tools. The inpatient unit in this study routinely uses the DOSS and CAM-ICU for ventilated patients, and validation studies have shown sensitivity and specificity to be above 90%.[33, 41] Because of their superior accuracy over clinical gestalt in the ED, we used DOSS and CAM-ICU as an approximation of the delirium outcome in this study. This approach likely included both patients who presented with delirium in the ED and who developed delirium after the ED visit. A future study should investigate both the prevalence and the incidence of delirium by screening for delirium in the ED and inpatient unit to explore the distinctive features of those who come to the ED with delirium and those who develop delirium during a hospital stay.

Our machine learning models identified a combination of clinical variables relevant for delirium risk estimation that included patient demographic information, current and past diagnoses, medications, and labs. We previously analyzed the diagnostic characteristics of three pre-existing delirium risk estimation models in the literature: delirium risk score, risk prediction model, and susceptibility score.[3, 29, 42] These models were examined in our retrospective hospital-wide data, and AUCs ranged from 0.71 to 0.8.[43] Several models from this study demonstrated an improvement in model performance with the GBM, RF, and LR models achieving AUCs of 0.839, 0.837, 0.831 respectively. Although the AUCs of RF and GBM were significantly better than LR at the 95% confidence level, the difference in performance may not be clinically meaningful. We prioritized model performance at a higher sensitivity threshold, as it could be used as a screening tool potentially triggered by electronic health records and then requiring confirmatory tests for delirium.

Our study reported the importance of variables such as age, dementia, and other neurocognitive disorders which cannot be modified, and also variables such as heart rate, BMI, and urinary tract infections which can be modified but may reflect underlying illness which may or may not be modifiable. The importance of benzodiazepines and anticholinergics was not as high, but these drugs can be decreased with a deprescribing program.[44, 45] These findings highlight several modifiable variables that we can approach in the ED. Among conditions particularly relevant to psychiatry, neurodevelopmental disorders, schizophrenia spectrum and other psychotic disorders, epilepsy or convulsions, and depressive disorders all appeared to increase risk of delirium. Nervous system signs and symptoms, other nervous system disorders (including in previous encounters), and symptoms of mental and substance use conditions also were risk factors for delirium, though these may have represented symptoms of delirium in some cases.

Our study used the ML models to estimate the risk of delirium by predicting positive delirium screening in the hospital setting using data available from prior visits, the ED, and early in hospitalization. Several hospital-based delirium prediction studies are in the literature, including a recent study by Wong et al., who evaluated the models LR, GBM, RF, SVM and an artificial neural network with a single hidden layer for the risk prediction of postoperative delirium in aged 18 or older using electronic health records. They report the best performance of a GBM with an AUC of 0.855. However, when models are stratified by ages 18 to 64 vs ages greater than 64, the best performing model for patients older than 64 achieves an AUC of 0.807. In comparison to our study with a patient population also

aged 64 or older, our top three models outperform these results, but the patient population is different.[30] . Spiller et al. employed the prediction models for positive DOSS screening in the hospital setting, using the nursing documentation with excellent accuracy.[46] We are adding knowledge related to predicting delirium in the hospital setting from ED data during the transition of care, and emphasize that rigorous prediction models are needed to improve delirium detection in the hospital setting.

Strengths and Limitations

This study included DOSS or CAM-ICU assessments on all hospitalized older adults which provided a large dataset to derive and validate a delirium risk estimation model with a relatively high AUC, which could be used to focus screening and preventive interventions on higher risk patients. By leveraging ED data, our model is able to predict risk of positive delirium screens to identify prevalent and impending delirium early in the hospital stay. During model development, we also employed cross-validation which is a robust method for mitigating overfitting and reducing bias.

There are several limitations that are important to recognize. First, these results may not generalize to other institutions with different delirium prevention practices. Second, delirium screenings are imperfect measures of delirium, though prior work at our institution found that the DOSS performed by nursing staff in clinical settings was 90% sensitive and 91% specific for delirium as classified by the Delirium Rating Scale-Revised-98, with false positives displaying meaningful delirium symptoms.[33] Nevertheless, it is possible that screening performance differed from that observed in the validation study. Third, we did not evaluate older adults who were discharged home after ED evaluation. Fourth, this study was conducted in a primarily white population and would benefit from further validation in a more diverse population. Fifth, the availability of diagnoses from prior encounters to predict delirium during hospitalization may vary across health systems. Lastly, we acknowledge that the screening tool performance may differ in patients with dementia from those without.

CONCLUSION

This study demonstrated the use of machine learning algorithms to identify a combination of variables that enables the estimation of delirium risk in patients hospitalized from the ED. The discovery of a risk estimation model that clinicians can use as a clinical decision aid could lead to improved detection of delirium and identification of a high-risk group that most benefits from preventive interventions. Our future objective will be to develop a clinical decision aid integrated into electronic health records to estimate the risk of delirium in real-time, so ED providers and the inpatient team can focus on delirium screening for high-risk individuals and implementing a delirium prevention program.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This study was supported by an award to Sangil Lee by the NIDUS delirium network (No. NIA R24AG054259). Research reported in this publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002537. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Francis J, Martin D, Kapoor WN. A prospective study of delirium in hospitalized elderly. JAMA. Feb 23 1990;263(8):1097–101. [PubMed: 2299782]
2. O'Keeffe S, Lavan J. The prognostic significance of delirium in older hospital patients. J Am Geriatr Soc. Feb 1997;45(2):174–8. [PubMed: 9033515]
3. Kennedy M, Enander RA, Tadiri SP, Wolfe RE, Shapiro NI, Marcantonio ER. Delirium risk prediction, healthcare use and mortality of elderly adults in the emergency department. J Am Geriatr Soc. Mar 2014;62(3):462–9. doi:10.1111/jgs.12692 [PubMed: 24512171]
4. Ely EW, Gautam S, Margolin R, et al. The impact of delirium in the intensive care unit on hospital length of stay. Intensive Care Med. Dec 2001;27(12):1892–900. doi:10.1007/s00134-001-1132-2 [PubMed: 11797025]
5. Lewis LM, Miller DK, Morley JE, Nork MJ, Lasater LC. Unrecognized delirium in ED geriatric patients. Am J Emerg Med. Mar 1995;13(2):142–5. doi:10.1016/0735-6757(95)90080-2 [PubMed: 7893295]
6. Lee S, Angel C, Han JH. Succinct Approach to Delirium in the Emergency Department. Curr Emerg Hosp Med Rep. Mar 18 2021:1–8. doi:10.1007/s40138-021-00226-9
7. Trzepacz PT, Mittal D, Torres R, Kanary K, Norton J, Jimerson N. Validation of the Delirium Rating Scale-revised-98: comparison with the delirium rating scale and the cognitive test for delirium. J Neuropsychiatry Clin Neurosci. Spring 2001;13(2):229–42. doi:10.1176/jnp.13.2.229 [PubMed: 11449030]
8. Han JH, Wilson A, Graves AJ, et al. Validation of the Confusion Assessment Method for the Intensive Care Unit in older emergency department patients. Acad Emerg Med. Feb 2014;21(2):180–7. doi:10.1111/acem.12309 [PubMed: 24673674]
9. Han JH, Wilson A, Vasilevskis EE, et al. Diagnosing delirium in older emergency department patients: validity and reliability of the delirium triage screen and the brief confusion assessment method. Ann Emerg Med. Nov 2013;62(5):457–65. doi:10.1016/j.annemergmed.2013.05.003 [PubMed: 23916018]
10. Luetz A, Weiss B, Boettcher S, Burmeister J, Wernecke KD, Spies C. Routine delirium monitoring is independently associated with a reduction of hospital mortality in critically ill surgical patients: A prospective, observational cohort study. J Crit Care. Oct 2016;35:168–73. doi:10.1016/j.jcrc.2016.05.028 [PubMed: 27481754]
11. Moon KJ, Lee SM. The effects of a tailored intensive care unit delirium prevention protocol: A randomized controlled trial. Int J Nurs Stud. Sep 2015;52(9):1423–32. doi:10.1016/j.ijnurstu.2015.04.021 [PubMed: 26032729]
12. Pasin L, Landoni G, Nardelli P, et al. Dexmedetomidine reduces the risk of delirium, agitation and confusion in critically Ill patients: a meta-analysis of randomized controlled trials. J Cardiothorac Vasc Anesth. Dec 2014;28(6):1459–66. doi:10.1053/j.jvca.2014.03.010 [PubMed: 25034724]
13. Oosterhoff JHF, Karhade AV, Oberai T, Franco-Garcia E, Doornberg JN, Schwab JH. Prediction of Postoperative Delirium in Geriatric Hip Fracture Patients: A Clinical Prediction Model Using Machine Learning Algorithms. Geriatr Orthop Surg Rehabil. 2021;12:21514593211062277. doi:10.1177/21514593211062277

14. Bishara A, Chiu C, Whitlock EL, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. BMC Anesthesiol. Jan 3 2022;22(1):8. doi:10.1186/s12871-021-01543-y [PubMed: 34979919]
15. Wang Y, Lei L, Ji M, Tong J, Zhou CM, Yang JJ. Predicting postoperative delirium after microvascular decompression surgery with machine learning. J Clin Anesth. Nov 2020;66:109896. doi:10.1016/j.jclinane.2020.109896 [PubMed: 32504969]
16. Jung JW, Hwang S, Ko S, et al. A machine-learning model to predict postoperative delirium following knee arthroplasty using electronic health records. BMC Psychiatry. Jun 27 2022;22(1):436. doi:10.1186/s12888-022-04067-y [PubMed: 35761274]
17. Zhang Y, Wan DH, Chen M, et al. Automated machine learning-based model for the prediction of delirium in patients after surgery for degenerative spinal disease. CNS Neurosci Ther. Jan 2023;29(1):282–295. doi:10.1111/cns.14002 [PubMed: 36258311]
18. Liu Y, Shen W, Tian Z. Using Machine Learning Algorithms to Predict High-Risk Factors for Postoperative Delirium in Elderly Patients. Clin Interv Aging. 2023;18:157–168. doi:10.2147/CIA.S398314 [PubMed: 36789284]
19. Hu XY, Liu H, Zhao X, et al. Automated machine learning-based model predicts postoperative delirium using readily extractable perioperative collected electronic data. CNS Neurosci Ther. Apr 2022;28(4):608–618. doi:10.1111/cns.13758 [PubMed: 34792857]
20. Mufti HN, Hirsch GM, Abidi SR, Abidi SSR. Exploiting Machine Learning Algorithms and Methods for the Prediction of Agitated Delirium After Cardiac Surgery: Models Development and Validation Study. JMIR Med Inform. Oct 23 2019;7(4):e14993. doi:10.2196/14993 [PubMed: 31558433]
21. Oberai T, Oosterhoff JHF, Woodman R, Doornberg JN, Kerkhoffs G, Jaarsma R. Development of a postoperative delirium risk scoring tool using data from the Australian and New Zealand Hip Fracture Registry: an analysis of 6672 patients 2017-2018. Arch Gerontol Geriatr. May-Jun 2021;94:104368. doi:10.1016/j.archger.2021.104368 [PubMed: 33556634]
22. Racine AM, Tommet D, D'Aquila ML, et al. Machine Learning to Develop and Internally Validate a Predictive Model for Post-operative Delirium in a Prospective, Observational Clinical Cohort Study of Older Surgical Patients. J Gen Intern Med. Feb 2021;36(2):265–273. doi:10.1007/s11606-020-06238-7 [PubMed: 33078300]
23. Hur S, Ko RE, Yoo J, Ha J, Cha WC, Chung CR. A Machine Learning-Based Algorithm for the Prediction of Intensive Care Unit Delirium (PRIDE): Retrospective Study. JMIR Med Inform. Jul 26 2021;9(7):e23401. doi:10.2196/23401 [PubMed: 34309567]
24. Bhattacharyya A, Sheikhalishahi S, Torbic H, et al. Delirium prediction in the ICU: designing a screening tool for preventive interventions. JAMIA Open. Jul 2022;5(2):ooac048. doi:10.1093/jamiaopen/ooac048 [PubMed: 35702626]
25. Gong KD, Lu R, Bergamaschi TS, et al. Predicting Intensive Care Delirium with Machine Learning: Model Development and External Validation. Anesthesiology. Mar 1 2023;138(3):299–311. doi:10.1097/ALN.0000000000004478 [PubMed: 36538354]
26. Coombes CE, Coombes KR, Fareed N. A novel model to label delirium in an intensive care unit from clinician actions. BMC Med Inform Decis Mak. Mar 9 2021;21(1):97. doi:10.1186/s12911-021-01461-6 [PubMed: 33750375]
27. Green C, Bonavia W, Toh C, Tiruvoipati R. Prediction of ICU Delirium: Validation of Current Delirium Predictive Models in Routine Clinical Practice. Crit Care Med. Mar 2019;47(3):428–435. doi:10.1097/CCM.0000000000003577 [PubMed: 30507844]
28. Oh J, Cho D, Park J, et al. Prediction and early detection of delirium in the intensive care unit by using heart rate variability and machine learning. Physiol Meas. Mar 27 2018;39(3):035004. doi:10.1088/1361-6579/aaab07 [PubMed: 29376502]
29. Han JH, Zimmerman EE, Cutler N, et al. Delirium in older emergency department patients: recognition, risk factors, and psychomotor subtypes. Acad Emerg Med. Mar 2009;16(3):193–200. doi:10.1111/j.1553-2712.2008.00339.x [PubMed: 19154565]
30. Wong BA YA, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk

in Newly Hospitalized Patients Without Known Cognitive Impairment. JAMA Network Open. 2018;1(4):e181018. doi:10.1001

31. Collins GS, Reitsma JB, Altman DG, Moons KGM, members of the Tg. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. Eur Urol. Jun 2015;67(6):1142–1151. doi:10.1016/j.eururo.2014.11.025 [PubMed: 25572824]
32. Schuurmans MJ, Shortridge-Baggett LM, Duursma SA. The Delirium Observation Screening Scale: a screening instrument for delirium. Res Theory Nurs Pract. Spring 2003;17(1):31–50. doi:10.1891/rtnp.17.1.31.53169 [PubMed: 12751884]
33. Gavinski K, Carnahan R, Weckmann M. Validation of the delirium observation screening scale in a hospitalized older population. J Hosp Med. Jul 2016;11(7):494–7. doi:10.1002/jhm.2580 [PubMed: 26970312]
34. Ely EW, Inouye SK, Bernard GR, et al. Delirium in mechanically ventilated patients: validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU). JAMA. Dec 5 2001;286(21):2703–10. doi:10.1001/jama.286.21.2703 [PubMed: 11730446]
35. Kersten H, Molden E, Willumsen T, Engedal K, Bruun Wyller T. Higher anticholinergic drug scale (ADS) scores are associated with peripheral but not cognitive markers of cholinergic blockade. Cross sectional data from 21 Norwegian nursing homes. Br J Clin Pharmacol. Mar 2013;75(3):842–9. doi:10.1111/j.1365-2125.2012.04411.x [PubMed: 22924454]
36. Carnahan RM, Lund BC, Perry PJ, Culp KR, Pollock BG. The relationship of an anticholinergic rating scale with serum anticholinergic activity in elderly nursing home residents. Psychopharmacol Bull. Autumn 2002;36(4):14–9. [PubMed: 12858139]
37. Salsabili M, Kiogou S, Adam TJ. The Evaluation of Clinical Classifications Software Using the National Inpatient Sample Database. AMIA Jt Summits Transl Sci Proc. 2020;2020:542–551. [PubMed: 32477676]
38. Jadhav A, Pramod D, Ramanathan K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence. 2019/08/24 2019;33(10):913–933. doi:10.1080/08839514.2019.1637138
39. Pedregosa F VG, Gramfort A et al. . Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011) 2825–2830.
40. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. Lancet. Mar 8 2014;383(9920):911–22. doi:10.1016/S0140-6736(13)60688-1 [PubMed: 23992774]
41. Jorgensen SM, Carnahan RM, Weckmann MT. Validity of the Delirium Observation Screening Scale in Identifying Delirium in Home Hospice Patients. Am J Hosp Palliat Care. Sep 2017;34(8):744–747. doi:10.1177/1049909116658468 [PubMed: 27413013]
42. Pendlebury ST, Lovett N, Smith SC, Cornish E, Mehta Z, Rothwell PM. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of externally derived risk scores. Age Ageing. Jan 2016;45(1):60–5. doi:10.1093/ageing/afv177 [PubMed: 26764396]
43. Lee S, Harland K, Mohr NM, et al. Evaluation of emergency department derived delirium prediction models using a hospital-wide cohort. J Psychosom Res. Dec 2019;127:109850. doi:10.1016/j.jpsychores.2019.109850 [PubMed: 31678811]
44. Wilson MG, Lee TC, Hass A, Tannenbaum C, McDonald EG. EMPOWERing Hospitalized Older Adults to Deprescribe Sedative Hypnotics: A Pilot Study. J Am Geriatr Soc. Jul 2018;66(6):1186–1189. doi:10.1111/jgs.15300 [PubMed: 29492957]
45. Martin P, Tamblyn R, Benedetti A, Ahmed S, Tannenbaum C. Effect of a Pharmacist-Led Educational Intervention on Inappropriate Medication Prescriptions in Older Adults: The D-PRESCRIBE Randomized Clinical Trial. JAMA. Nov 13 2018;320(18):1889–1898. doi:10.1001/jama.2018.16131 [PubMed: 30422193]
46. Spiller TR, Tufan E, Petry H, et al. Delirium screening in an acute care setting with a machine learning classifier based on routinely collected nursing data: A model development study. J Psychiatr Res. Dec 2022;156:194–199. doi:10.1016/j.jpsychires.2022.10.018 [PubMed: 36252349]

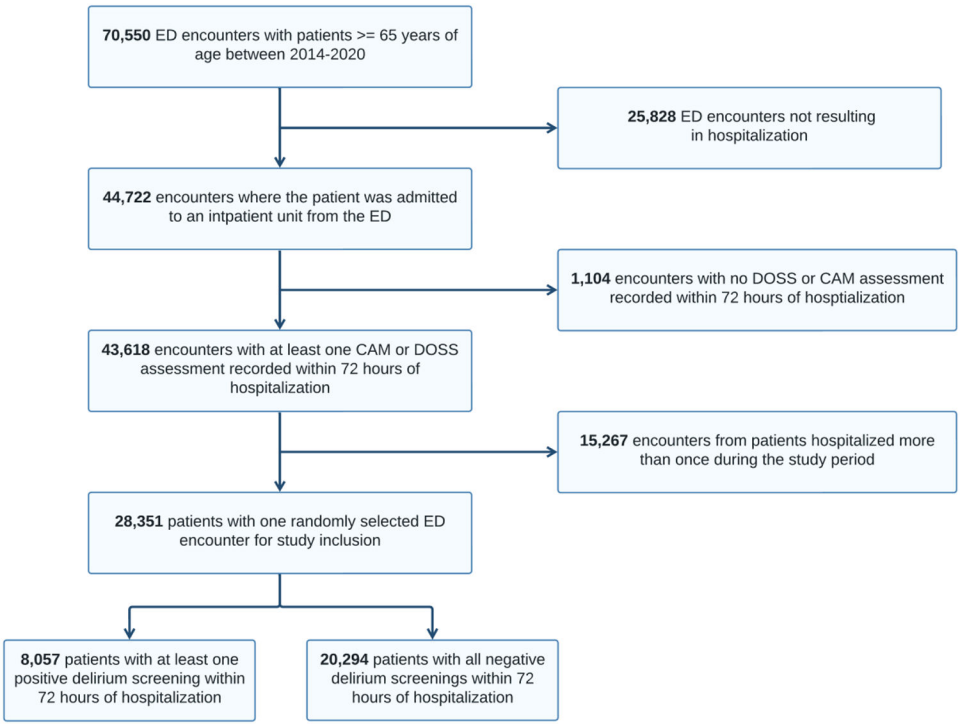

Figure 1.
Flow chart of patient selection

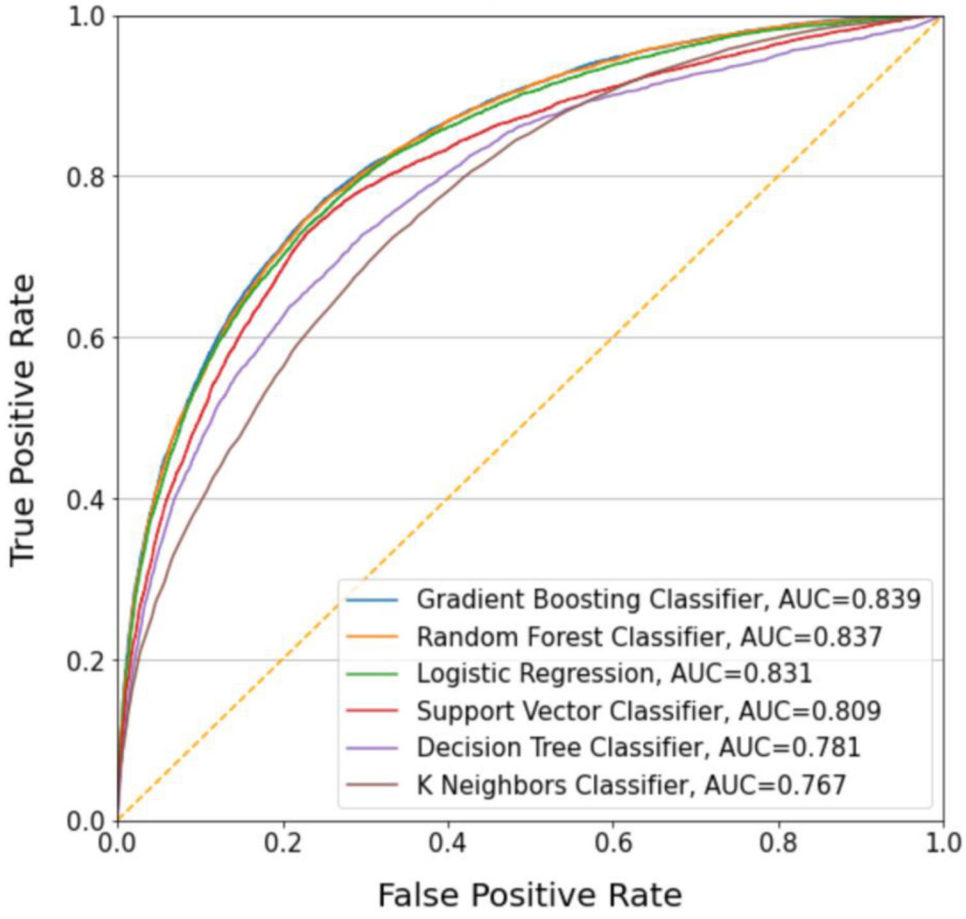

Figure 2.
Receiver Operating Curves

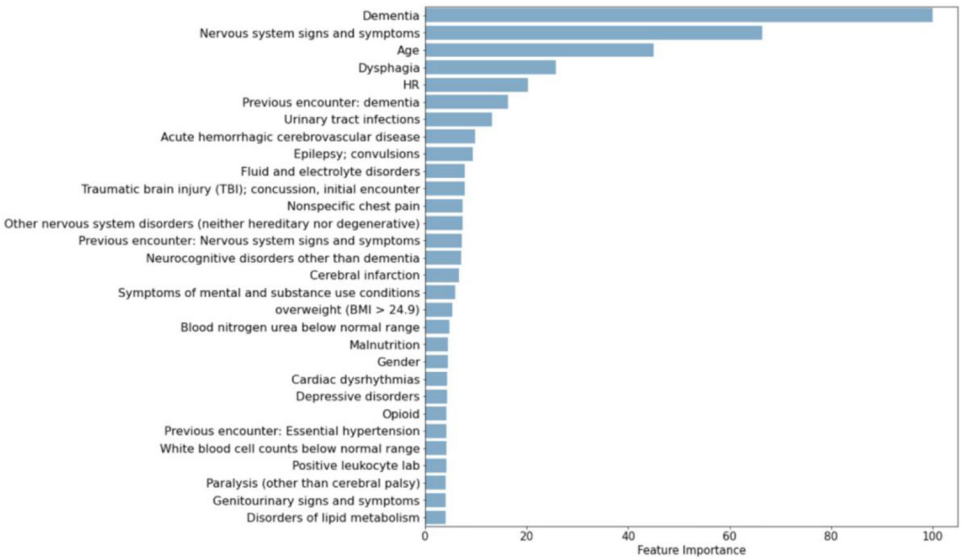

Figure 3.
Scaled feature performance for the top 30 variables selected by RF

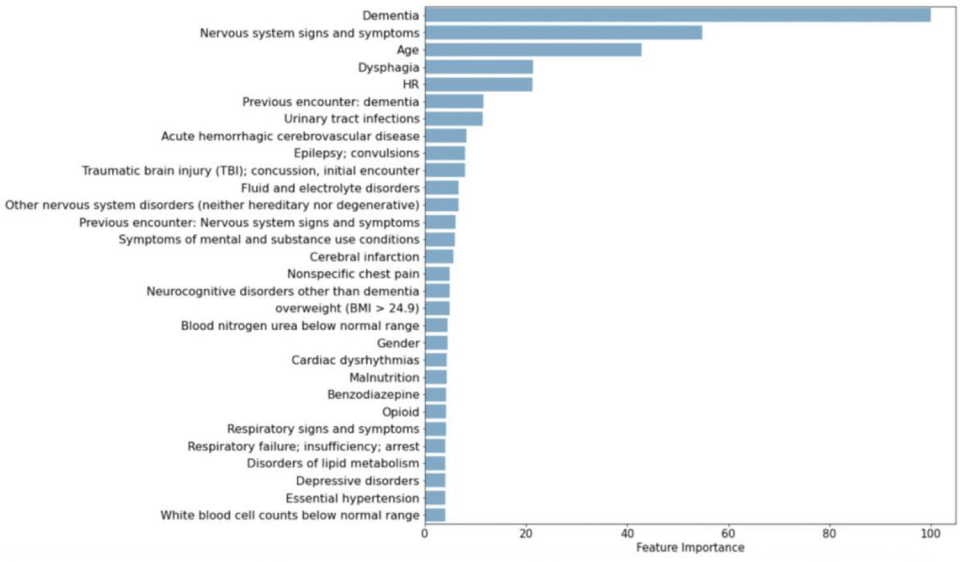

Figure 4.

Scaled feature performance for the top 30 variables selected by GBM

Table 1.

Patient Characteristics

	All	Negative Delirium Screening	Positive Delirium Screening	
	N = 28,351 No. (%)	n = 20,294 No. (%)	n = 8,057 No. (%)	Standardized Difference
Gender				
Female	13,921 (49.1)	9,687 (47.7)	4,234 (52.6)	−0.098
Male	14,430 (50.9)	10,607 (52.3)	3,823 (47.4)	0.098
Age				
Mean (SD)	76.3 (8.1)	75.2 (7.6)	79.1 (8.6)	0.494
Race				
African American/Black	573 (2.0)	412 (2.0)	161 (2.0)	0.0
Asian	219 (0.8)	168 (0.8)	51 (0.6)	0.024
Hispanic/Latino of any race	455 (1.6)	353 (1.7)	102 (1.3)	0.033
Multi-racial or other race	354 (1.2)	238 (1.2)	116 (1.4)	−0.018
White	26,750 (94.4)	19,123 (94.2)	7,627 (94.7)	−0.022
Hospital stay (days)				
Median (Q1-Q3)	3.7 (1.8 – 6.7)	3.1 (1.7 – 5.8)	5.0 (3.0 – 8.5)	

SD: standard deviation

Table 2.

Diagnostic characteristics for machine learning models

Machine Learning Model	Accuracy	AUC (95% CI)
Gradient Boosted Machine	0.80	0.839 (0.837,0.841)
Random Forest	0.80	0.837 (0.835, 0.838)
Logistic Regression	0.80	0.831 (0.829, 0.833)
Support Vector Machine	0.78	0.809 (0.807, 0.810)
Decision Tree	0.78	0.781 (0.779, 0.783)
K-nearest neighbors	0.73	0.767 (0.766, 0.769)

Table 3.

Performance metrics at a 90% sensitivity threshold: specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (PLR), negative likelihood ratio (NLR).

Classifier	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	PLR (95% CI)	NLR (95% CI)
Gradient Boosted Machine at 90% sensitivity	53.5% (53.0%, 54.0%)	43.5% (43.2%, 43.8%)	93.1% (93.1%, 93.2%)	1.943 (1.921, 1.966)	0.186 (0.184, 0.188)
Random Forest at 90% sensitivity	52.8% (52.3%, 53.4%)	43.2% (42.9%, 43.4%)	93.1% (93.0%, 93.1%)	1.916 (1.895, 1.937)	0.188 (0.186, 0.19)
Logistic Regression at 90% sensitivity	51.0% (50.4%, 51.6%)	42.3% (42.0%, 42.6%)	92.8% (92.7%, 92.9%)	1.846 (1.823, 1.869)	0.195 (0.193, 0.197)
Support Vector Machine at 90% sensitivity	43.3% (42.7%, 43.9%)	38.7% (38.5%, 39.0%)	91.6% (91.5%, 91.7%)	1.594 (1.577, 1.612)	0.23 (0.227, 0.233)
Decision Tree at 90% sensitivity	36.4% (35.3%, 37.4%)	36.3% (36.0%, 36.6%)	90.9% (90.7%, 91.1%)	1.438 (1.417, 1.459)	0.253 (0.247, 0.259)
K-Nearest Neighbors at 90% sensitivity	39.0% (38.5%, 39.5%)	37.3% (37.1%, 37.5%)	91.8% (91.7%, 91.9%)	1.498 (1.486, 1.511)	0.226 (0.223, 0.229)

Table 4.

Performance metrics at a 90% specificity threshold: sensitivity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (PLR), negative likelihood ratio (NLR).

Classifier	Sensitivity (95% CI)	PPV (95% CI)	NPV (95% CI)	PLR (95% CI)	NLR (95% CI)
Gradient Boosted Machine at 90% specificity	55.6% (55.2%, 56.0%)	68.9% (68.7%, 69.1%)	83.6% (83.5%, 83.8%)	5.587 (5.542, 5.631)	0.493 (0.488, 0.498)
Random Forest at 90% specificity	54.9% (54.5%, 55.3%)	68.7% (68.5%, 68.8%)	83.4% (83.3%, 83.5%)	5.528 (5.487, 5.569)	0.501 (0.496, 0.505)
Logistic Regression at 90% specificity	54.7% (54.3%, 55.2%)	68.6% (68.4%, 68.7%)	83.4% (83.2%, 83.5%)	5.497 (5.454, 5.541)	0.503 (0.498, 0.508)
Support Vector Machine at 90% specificity	50.5% (50.0%, 50.9%)	66.8% (66.6%, 67.0%)	82.1% (81.9%, 82.2%)	5.075 (5.030, 5.120)	0.550 (0.545, 0.555)
Decision Tree at 90% specificity	47.3% (46.8%, 47.7%)	64.5% (64.3%, 64.7%)	81.1% (80.9%, 81.2%)	4.58 (4.540, 4.621)	0.588 (0.583, 0.593)
K-Nearest Neighbors at 90% specificity	40.8% (40.4%, 41.2%)	60.2% (60.0%, 60.4%)	79.2% (79.0%, 79.3%)	3.816 (3.778, 3.855)	0.663 (0.659, 0.668)

Table 5.

Top 30 variables selected by LR

Variable	Odds Ratio	95% CI
Dementia	5.03	(4.53, 5.58)
Nervous system signs and symptoms	2.75	(2.55, 2.96)
Symptoms of mental and substance use conditions	2.29	(1.79, 2.94)
Dysphagia	2.02	(1.84, 2.23)
Neurodevelopmental disorders	1.97	(1.33, 2.91)
Hepatic failure	1.92	(1.48, 2.51)
Schizophrenia spectrum and other psychotic disorders	1.9	(1.31, 2.75)
Nervous system cancers - brain	1.87	(1.32, 2.67)
Previous encounter: Schizophrenia spectrum and other psychotic disorders	1.83	(1.18, 2.84)
Previous encounter: dementia	1.73	(1.51, 1.99)
Fracture of the neck of the femur (hip), initial encounter	1.66	(1.38, 1.99)
Previous encounter: Other nervous system disorders (neither hereditary nor degenerative)	1.55	(1.39, 1.74)
Traumatic brain injury (TBI); concussion, initial encounter	1.50	(1.32, 1.71)
Epilepsy; convulsions	1.47	(1.28, 1.69)
Age (years)	1.45	(1.40, 1.50)
Urinary tract infections	1.44	(1.30, 1.60)
Alcohol-related disorders	1.41	(1.18, 1.70)
Benzodiazepine	1.39	(1.26, 1.55)
Previous encounter: Nervous system signs and symptoms	1.37	(1.24, 1.50)
Secondary malignancies	1.35	(1.16, 1.58)
Parkinson’s disease	1.35	(1.12, 1.63)
Pressure ulcer of skin	1.34	(1.11, 1.62)
CNS abscess	1.34	(0.85, 2.10)
Neurocognitive disorders	1.34	(1.05, 1.71)
Previous encounter: Symptoms of mental and substance use conditions	1.26	(1.00, 1.59)
Previous encounter: Neurocognitive disorders	1.23	(1.03, 1.48)
Malnutrition	1.22	(1.10, 1.35)
Hypertension with complications and secondary hypertension	1.21	(1.00, 1.42)
Anticholinergic	1.20	(1.08, 1.34)
Depressive disorders	1.20	(1.09-1.33)

Table 6.

Variables with the greatest importance selected by GBM

	Variable Frequencies, No.(%)			
Categorical variables	**Negative Delirium Screening n = 20294**	**Positive Delirium Screening n = 8057**	**Selection by RF**	**Selection by LR**
Dementia	814 (4.0)	2466 (30.6)	✓	✓
Nervous system signs and symptoms	3928 (19.4)	4040 (50.1)	✓	✓
Dysphagia	1567 (7.7)	1953 (24.2)	✓	✓
Previous encounter: dementia	741 (3.70)	1294 (16.1)	✓	✓
Urinary tract infections	1818 (9.0)	1585 (19.7)	✓	✓
Fluid and electrolyte disorders	6192 (30.5)	3291 (40.8)	✓	
Epilepsy; convulsions	771 (3.8)	793 (9.8)	✓	✓
Traumatic brain injury (TBI); concussion, initial encounter	1554 (7.7)	1177 (14.6)	✓	✓
Previous encounter: Other nervous system disorders (neither hereditary nor degenerative)	1961 (9.7)	1323 (16.4)	✓	✓
Symptoms of mental and substance use conditions	128 (0.6)	302 (3.7)	✓	✓
Neurocognitive disorders	147 (0.7)	367 (4.6)	✓	✓
Acute hemorrhagic cerebrovascular disease	1616 (8.0)	1271 (15.8)	✓	
overweight (BMI > 24.9)	14286 (70.4)	4895 (60.8)	✓	
Other nervous system disorders (neither hereditary nor degenerative)	1418 (7.0)	1095 (13.6)	✓	
Previous encounter: Nervous system signs and symptoms	4045 (19.9)	2228 (27.7)	✓	✓
Malnutrition	2174 (10.7)	1412 (17.5)	✓	✓
Cerebral infarction	2092 (10.3)	1418 (17.6)	✓	
Respiratory failure; insufficiency; arrest	2871 (14.1)	1476 (18.3)	✓	
Previous encounter: Neurocognitive disorders	436 (2.1)	534 (6.6)	✓	
Blood nitrogen urea below normal range	7464 (36.8)	3497 (43.4)	✓	
Previous encounter: Essential hypertension	9543 (47.0)	3043 (37.8)		
Nonspecific chest pain	3213 (15.8)	695 (8.6)	✓	
Gender	10607 (52.3)	3823 (47.4)	✓	
Benzodiazepine	1730 (8.5)	1066 (13.2)		✓
Cardiac dysrhythmias	5788 (28.5)	2789 (34.6)	✓	
Previous encounter: Disorders of lipid metabolism	7245 (35.7)	2095 (26.0)	✓	
Previous encounter: Abdominal pain and other digestive/abdomen signs and symptoms	5168 (25.5)	1333 (16.5)	✓	
Previous encounter: Respiratory signs and symptoms	9535 (49.4)	3504 (47.4)	✓	
Continuous Variables	**Mean (SD)**			
Age	75.2 (7.6)	79.1 (8.6)	✓	✓
Heart rate, beats/min	77.6 (15.5)	80.1 (16.9)	✓	