Journal of Medical Systems (2018) 42:261
https://doi.org/10.1007/s10916-018-1109-0

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Prediction of Incident Delirium Using a Random Forest classifier

John P. Corradi[1] · Stephen Thompson[1] · Jeffrey F. Mather[1] · Christine M. Waszynski[2] · Robert S. Dicks[2]

Received: 18 December 2017 / Accepted: 24 October 2018

Abstract

Delirium is a serious medical complication associated with poor outcomes. Given the complexity of the syndrome, prevention and early detection are critical in mitigating its effects. We used Confusion Assessment Method (CAM) screening and Electronic Health Record (EHR) data for 64,038 inpatient visits to train and test a model predicting delirium arising in hospital. Incident delirium was defined as the first instance of a positive CAM occurring at least 48 h into a hospital stay. A Random Forest machine learning algorithm was used with demographic data, comorbidities, medications, procedures, and physiological measures. The data set was randomly partitioned 80% / 20% for training and validating the predictive model, respectively. Of the 51,240 patients in the training set, 2774 (5.4%) experienced delirium during their hospital stay; and of the 12,798 patients in the validation set, 701 (5.5%) experienced delirium. Under-sampling of the delirium negative population was used to address the class imbalance. The Random Forest predictive model yielded an area under the receiver operating characteristic curve (ROC AUC) of 0.909 (95% CI 0.898 to 0.921). Important variables in the model included previously identified predisposing and precipitating risk factors. This machine learning approach displayed a high degree of accuracy and has the potential to provide a clinically useful predictive model for earlier intervention in those patients at greatest risk of developing delirium.

Keywords Delirium · Prediction · Decision support · Machine learning · Random forest

Introduction

Delirium is a potentially lethal condition of altered mental status, attention, and level of consciousness with an acute onset and fluctuating course. Reported rates of incident delirium have ranged from 11 to 14% in general medicine wards, 20–29% in geriatric units, and 19–82% in intensive care [1]. Multiple studies have found that delirium is associated with poor outcomes (e.g. length of stay, in-hospital mortality, discharge disposition, and readmission), even after adjusting for additional factors such as age and severity of illness [2–4]. Recent studies provide evidence that delirium is also associated with long term cognitive decline and an increase in depressive symptoms [5–8].

There are multiple known predisposing and precipitating risk factors for delirium, but the pathophysiology of the syndrome is still poorly understood [9]. Given the significant impact of delirium on patient outcomes, much emphasis has been placed on risk reduction and early detection [10, 11]. Bedside testing methods have been developed, such as the Confusion Assessment Method[1] (CAM), that allow for the rapid detection of delirium [12]. In late 2012, Hartford Hospital began implementing regular assessments, one per nursing shift, of patients using CAM in non-critical care areas and the CAM-ICU in critical care areas [13, 14].

Delirium is a direct consequence of a general medical condition (e.g. infection, organ insufficiency), intoxicating substance use or withdrawal, exposure to medications or toxins, or a combination of these factors. As such, the priority in treatment is to identify and address the underlying cause(s)

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10916-018-1109-0) contains supplementary material, which is available to authorized users.

✉ John P. Corradi
john.corradi@hhchealth.org

[1] Research Department, Hartford Hospital, 80 Seymour Street, ERD-223W, Hartford, CT 06102, USA

[2] Division of Geriatric Medicine, Hartford Hospital, Hartford, CT, USA

[1] Confusion Assessment Method. © 1988, 2003, Hospital Elder Life Program. All rights reserved. Not to be reproduced without permission. Instructions for correct usage available at: http://www.hospitalelderlifeprogram.org

Published online: 14 November 2018

of the delirium. Evidence suggests that the severity and/or duration of the episode of delirium bears some relationship to poor in-hospital and long term outcomes [15]. Therefore, early detection is important in mitigating the effects of this syndrome. Accurate individual patient risk prediction would allow for evaluation and intervention even earlier in the pathologic process.

Several studies have sought to identify significant risk factors and/or produce models for delirium prediction [16–19]. In addition to a focus on specific patient populations, most published models have used small sample sizes, a limited set of predictor variables, and have not accounted for temporal changes in clinical measures. We hypothesized that the use of a flexible modeling approach and a large data set with many features would yield accurate predictions. We leveraged a retrospective data set of delirium assessments using the Random Forest (RF) machine learning algorithm [20] to generate a predictive model for incident delirium across all hospital inpatients. The model incorporated demographics, comorbidities, procedures, medications, and measures reflecting dynamic changes in acute physiology.

Material and methods

Data set creation

This retrospective study was approved by the institutional review board. During the timeframe in which patient records were analyzed, Hartford Hospital used Sunrise Clinical Manager (SCM) (Allscripts, Chicago, IL) as the primary EHR system. For the analyses described here, patient records with recorded CAMs were identified from September 1, 2012 to September 30, 2015. The end date was chosen so that almost all hospital inpatient visits used ICD-9 diagnostic and procedural codes. Demographic data, comorbidities, and time-ordered physiological measures (vital signs), assessments, medication orders, procedures and patient location were extracted from the EHR using the Structured Query Language (SQL) for all inpatient stays. In order to focus on cases of incident delirium, 'positive' patients were defined as those with one or more positive CAMs where the first positive was recorded at least 48 h from time of admission. 'Negative' patients were those with at least a 48 h length of stay for whom *all* CAM results were negative. The distributions of time to event (first positive CAM or discharge for Negative patients) are shown in Figure S1.

Feature set

Only physiological measurements, medications, procedures, and assessments occurring prior to the first positive CAM were considered for 'positive' patients; whereas data from the entire hospital stay was used for 'negative' patients. Basic demographic data included age, sex, race/ethnicity, and marital status.

A Python script was written to parse time-ordered patient visit data and compute summary statistics for clinical vital signs - blood pressure, heart rate, respiratory rate, body temperature, and oxygen saturation - with the conditions that there were at least three longitudinal measures available and the last value had to have been recorded 8–24 h prior to the first positive CAM result. In addition to the physiological measures, the Richmond Agitation-Sedation Scale (RASS) was available for most patients. The RASS is an assessment of level of consciousness, with an integer scale ranging from −5 (unarousable) to +4 (combative). The minimum, maximum, and range were calculated for each physiological variable. The Pearson correlation between all physiological variables is shown in Figure S2. To avoid skewing of the descriptive statistics for physiological measures, filters were put in place to remove clearly erroneous values that fell outside the extremes of clinical observation. Only patient visits with more than 50% of the physiological values present were used in the predictive models. The fraction of missing values for each variable are listed in Table S1.

Comorbidities were accounted for by calculating the Quan modification of the Charlson Comorbidity Index (CCI) [21], using ICD-9 and ICD-10 diagnostic codes and the 'icd' R package [22]. In addition, the ICD codes were also used to identify the following specific comorbidities: dementia, alcohol and/or substance abuse, mood disorders, malignancies, fractures, and impaired vision.

Surgical procedures, intensive care, post-anesthesia care, and mechanical ventilation were captured for negative patients, while for positive patients only if they had occurred prior to the first positive CAM. Binary variables for surgery were elective, emergent, cardiac, neurological, trauma, and head/neck.

Medication orders in hospital were captured using a binary flag (absent/present), with the condition for positive patients of having the first order at least 8 h prior to the first positive CAM. Individual drugs were mapped to the categories available in Epic Clarity (Epic Systems Corp., Verona, WI), as Hartford Hospital has since migrated to an Epic EHR system. The longest matching substring between the SCM medication orders and those in Epic Clarity were mapped to pharmaceutical classes in Clarity. The mappings were then manually inspected and curated to reduce erroneous classifications. In addition, the total number of unique medications ordered for each patient was included as a predictor variable.

Predictive modeling

Predictive models were generated using the Distributed Random Forest (DRF) algorithm implemented in the H_2O

machine learning library [23], and used from within the R statistical computing environment. This implementation handles missing values by treating them as information, sending them down a branch based on minimizing a loss function. The DRF algorithm was run with a randomization seed using one core for reproducibility. The data set was randomly split into a training set containing 80% of the samples, and a validation set containing the remaining 20%. A grid search was employed on the training data with cross-validation to tune the RF parameters for variable selection size (mtries = $\sqrt{p}$, where p is the number of predictor variables), number of trees to generate (ntrees = 200), and maximum tree depth (max_depth = 30). To address the degree of imbalance in the classes (positives were 5.5% of the sample set), the majority (negative) class was under-sampled at each bootstrapping step [24, 25]. Specifically, a bootstrap sample of the negative class equivalent in size to the positive class was selected for tree building. To derive more accurate class probabilities from the Random Forest model, calibration by Platt scaling (a logistic transformation of the classification scores [26]) was implemented within the H_2O model using the validation data set. A calibration plot was generated to compare the original classification scores with the calibrated scores using the R package ‘caret’ [27] (Figure S3).

All reported ROC AUC values and confusion matrix-related metrics were calculated on predictions using the validation data and calibrated classification scores. The 95% confidence interval of the ROC AUC from the validation set was computed with 2000 bootstrapped stratified replicates using the R package ‘pROC’ [28]. The R package ‘PRROC’ was used to generate the ROC and precision-recall (PR) curves and area under the curve calculations [29]. Accuracy, precision, recall, and specificity were reported at operating points chosen to maximize commonly used performance metrics. Mean per

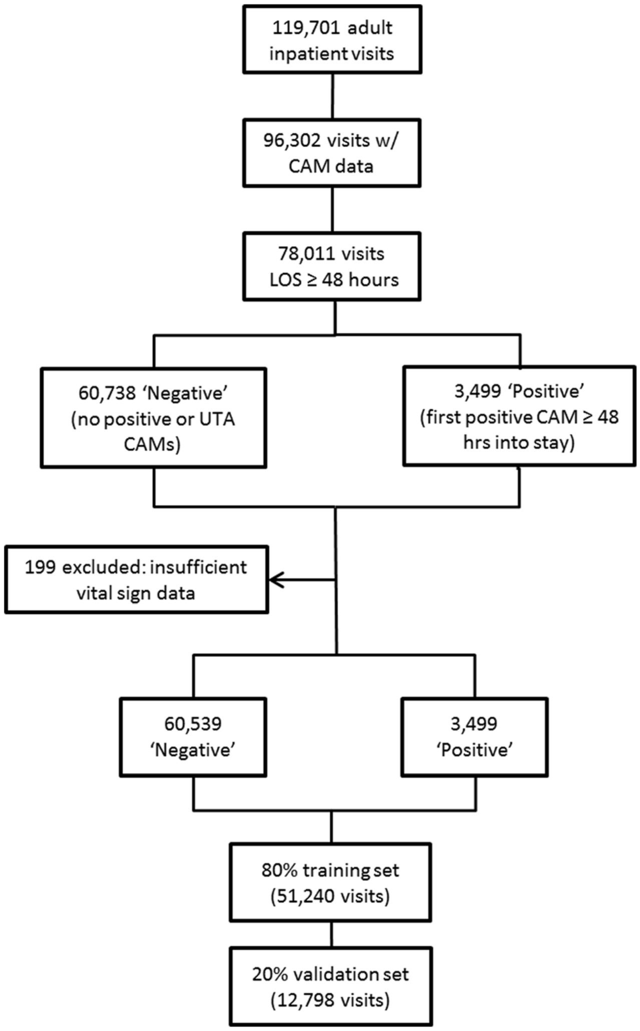

119,701 adult inpatient visits
96,302 visits w/ CAM data
78,011 visits LOS ≥ 48 hours
60,738 ‘Negative’ (no positive or UTA CAMs)
3,499 ‘Positive’ (first positive CAM ≥ 48 hrs into stay)
199 excluded: insufficient vital sign data
60,539 ‘Negative’
3,499 ‘Positive’
80% training set (51,240 visits)
20% validation set (12,798 visits)

Fig. 1 Schematic representing the sample selection process. In order to exclude cases of prevalent delirium, we removed patients with a positive CAM within the first 48 h of admission. Clinical data for the delirium ‘Positive’ groups were considered only if performed at least 8 h prior to the first positive CAM. *LOS* = length of stay, *UTA* = unable to assess

Table 1 Demographic composition and frequency of comorbidities and procedures in the training group and the validation group. Random partitioning resulted in similar composition

	Training (*N* = 51,240)	Validation (*N* = 12,798)
Age mean ± sd	63.2 ± 17.7	63.4 ± 17.6
Sex (female)	26,407 (51.5%)	6611 (51.7%)
Race		
White	36,361 (71%)	9061 (70.8%)
Black	5.097 (9.9%)	1268 (9.9%)
Hispanic	6600 (12.9%)	1674 (13.1%)
Asian	240 (0.5%)	55 (0.4%)
Other/Unknown	2942 (5.7%)	740 (5.8%)
Marital Status		
Married	22,320 (43.6%)	5558 (43.4%)
Single	14,350 (28%)	3618 (28.3%)
Divorced	5804 (11.3%)	1409 (11%)
Widowed	8065 (15.7%)	2028 (15.8%)
Separated	701 (1.4%)	185 (1.4%)
CCI median (IQR)	1 (3)	1 (3)
Dementia	3144 (6.1%)	769 (6%)
Substance abuse	9070 (17.7%)	2293 (17.9%)
Mood disorder	12,800 (25%)	3110 (24.3%)
Impaired vision	321 (0.6%)	85 (0.7%)
Fracture	3427 (6.7%)	857 (6.7%)
Malignancy	6602 (12.9%)	1652 (12.9%)
Previous delirium	1323 (2.6%)	312 (2.4%)
Surgery		
Elective	11,394 (22.2%)	2790 (21.8%)
Emergent	2073 (4%)	514 (4%)
ICU	10,886 (21.2%)	2790 (21.8%)
Mechanical ventilation	1012 (2%)	237 (1.9%)
Delirium	2774 (5.4%)	701 (5.5%)

CCI Charlson comorbidity index

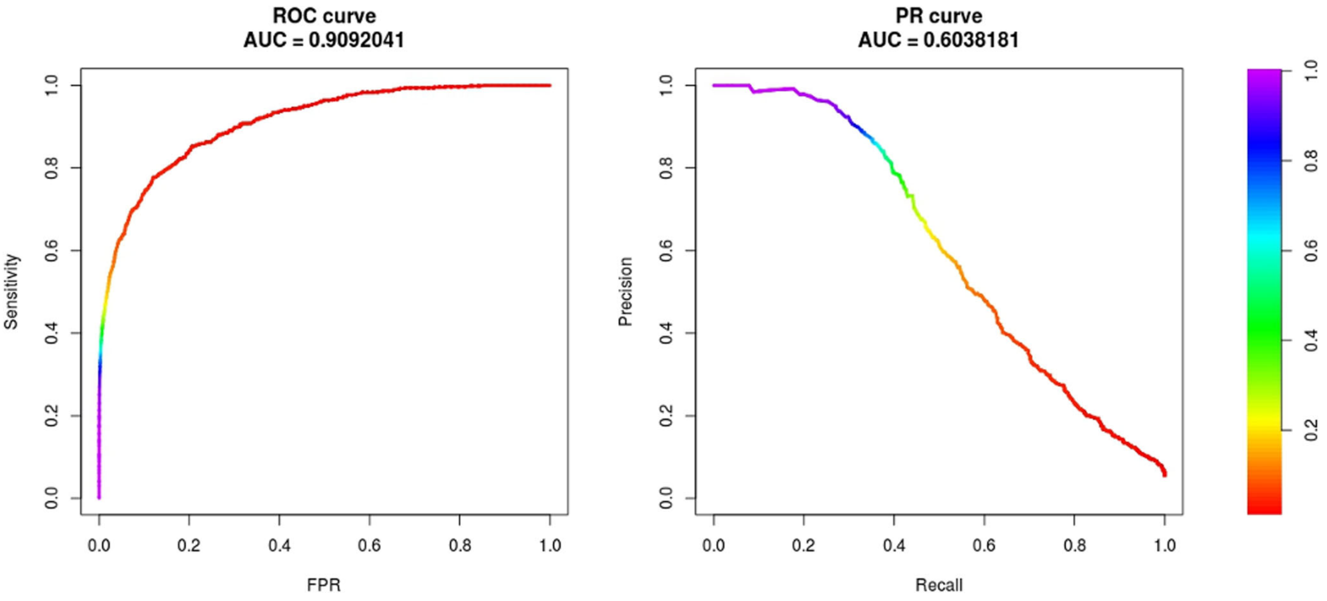


Fig. 2 Performance of the full predictive model. ROC and Precision-Recall (PR) curves illustrating performance of the model across all prediction thresholds. The coloring of sections of the curves corresponds to the calibrated classification scores. *AUC* = area under the curve, *FPR* = false positive rate, *Sensitivity* = *Recall*

class accuracy is simply the average of recall and specificity. The 'F-measures' are weighted harmonic means of precision and recall. They are calculated using this formula:

$$F_\beta = \left(1+\beta^2\right)\cdot\frac{precision\cdot recall}{\left(\beta^2\cdot precision\right)+recall}$$

The Matthews correlation coefficient (MCC) is a balanced measure considering true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It is defined as:

$$MCC = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The H_2O DRF algorithm scores variable importance by the improvement in the squared error when a variable is selected during a split, averaged over all trees. The variable lists reported here are shown ranked by the relative importance scaled to the most important variable, with a minimum threshold of 0.05.

Results

Hartford Hospital patients are evaluated for delirium using the CAM during every eight hour shift and with a change in condition. For the 37 month period from September 2012 through September 2015, there were 96,302 adult inpatient visits comprised of 58,973 unique patients in Hartford Hospital's delirium registry. Of those visits, 7613 (7.9%) had at least one positive CAM recorded. In order to focus only on incident delirium for predictive modeling, the study population was restricted to those patients whose first positive CAM occurred at least 48 h into their stay (the 'Positive' group), as well as those patients with only negative CAM recordings (i.e. no 'positive' or 'unable to assess' results) who were in the hospital for at least 48 h (the 'Negative' group). This selection yielded 64,237 patient visits, 3499 (5.4%) of which were in the Positive group. One additional restriction placed on the cohort was the availability of longitudinal physiological data. After excluding patients with insufficient vital sign data, the

Table 2 Accuracy of the model at original and calibrated thresholds selected to maximize metrics that weight performance measures differently. A description of the metrics and their calculation are provided in the methods section

Metric	Metric maximum	Threshold	Calibrated threshold	Accuracy	Precision	Recall	Specificity
Mean per-class accuracy	0.828	0.063	0.039	0.875	0.273	0.776	0.880
F2	0.586	0.091	0.057	0.915	0.357	0.698	0.927
F1	0.554	0.166	0.144	0.953	0.572	0.536	0.977
Absolute MCC	0.551	0.270	0.421	0.961	0.781	0.412	0.993
F0.5	0.675	0.310	0.557	0.962	0.841	0.377	0.996

MCC Mathews correlation coefficient

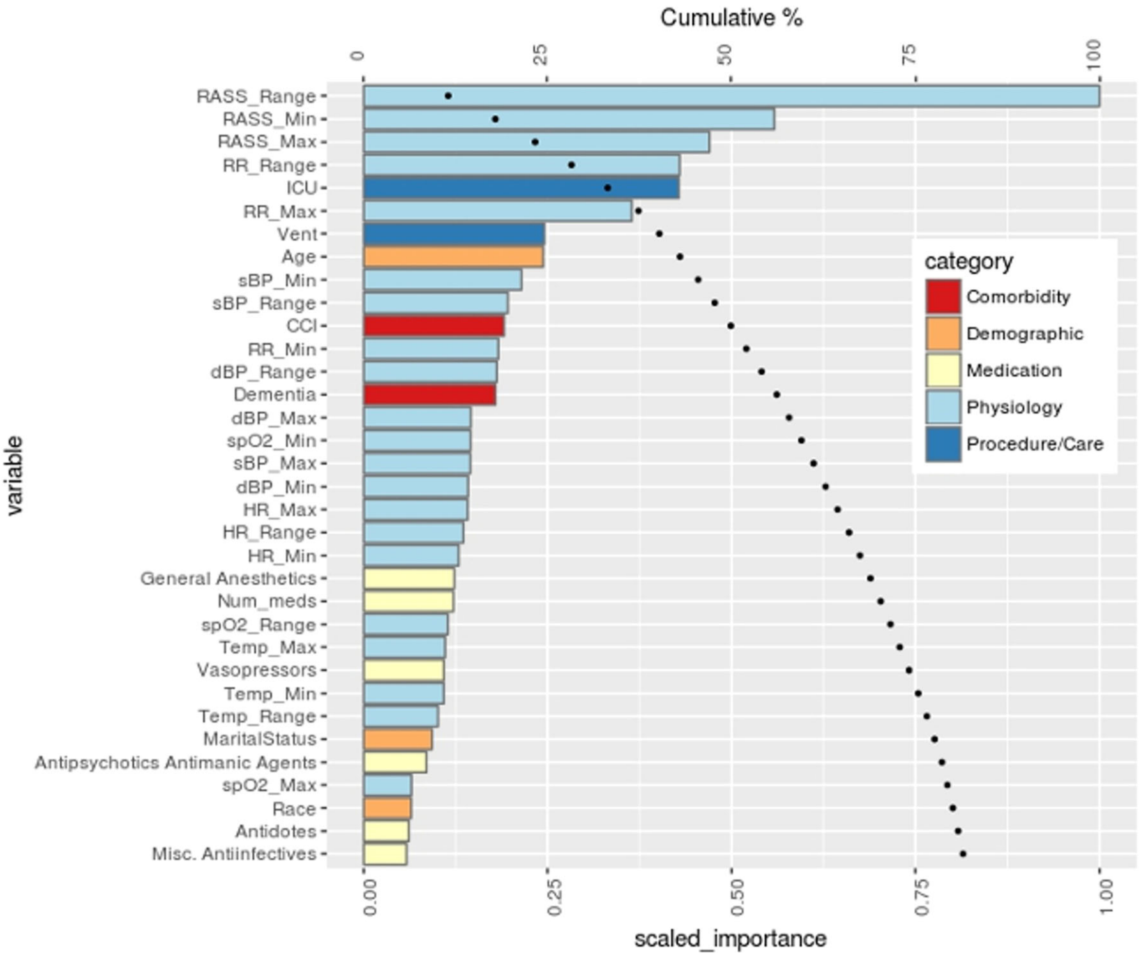


Fig. 3 Scaled relative variable importance for the full predictive model. Variable importance (bars) is scaled relative to the most important variable in the model. The minimum cutoff for reporting in the graph was a scaled importance of 0.05. Descriptions of the variables are found in supplemental data. The variables are colored by general category. Also plotted is the cumulative percentage of the model accounted for by the variables (points). *RASS* = Richmond Agitation Sedation Scale, *RR* = respiratory rate, *sBP* = systolic blood pressure, *dBP* = diastolic blood pressure, *HR* = heart rate, *spO2* = blood oxygen saturation, *Temp* = body temperature, *CCI* = Charlson Comorbidity Index

remaining set consisted of 64,038 patient visits (41,826 unique patients), 3499 (5.5%) of which were in the Positive group (Fig. 1).

A Random Forest (RF) model was trained on 80% of the patient visit data using all 128 variables covering demographics, physiological signs, comorbidities, procedures, and medications (Table S1). Model performance was assessed on the 20% held out validation set (see Table 1 for composition of training and test sets). The model was able to predict incident delirium in the validation set with an area under the receiver operating characteristic curve (ROC AUC) of 0.909 (95% CI 0.898 to 0.921) (Fig. 2). Training and testing error rates were comparable, indicating that the model did not overfit the training data. For prediction of unbalanced classes, it is useful to inspect the precision-recall (PR) curve as well. The PR AUC for the model was 0.604 (Fig. 2), and at the maximum F2 score - a weighted average of precision and recall favoring recall - the precision was 0.357, the recall (sensitivity) was 0.698, and the specificity was 0.927 (Table 2).

Table 3 Performance of sub-population models as measured by area under the ROC and Precision-Recall (PR) curves

	N	Delirium incidence	ROC AUC	PR AUC
All patients	64,038	5.5%	0.909 (95% CI 0.898 to 0.921)	0.604
ICU	13,676	17.5%	0.930 (95% CI 0.918 to 0.943)	0.826
Non-ICU	50,362	2.2%	0.861 (95% CI 0.839 to 0.884)	0.143

Random forest models can provide measures of variable importance, allowing some insight into the factors that most influence the predictions. Highly ranked variables in this model were level of consciousness (as measured by the RASS) and vital sign changes, critical care, mechanical ventilation, and other known predisposing risk factors such as age, dementia, comorbidity burden, and polypharmacy (Fig. 3).

Upon inspection of the prediction results, it was evident that the overall model performed somewhat better for ICU patients. To provide a more detailed view of the utility of an RF model for delirium within different patient populations, the sample set was split into critical care and non-critical care subgroups. The critical care sample set contained 13,676 patient visits, 2388 (17.5%) of which were positive for delirium. There were 50,362 non-critical care visits with 1087 (2.2%) in the Positive group. New RF models were generated for each cohort using the same process outlined for the full model. A model trained on ICU patients achieved a ROC AUC of 0.930 (95% CI 0.918 to 0.943), whereas a model trained on non-ICU patients yielded a ROC AUC of 0.861 (95% CI 0.839 to 0.884) (Table 3).

Important variables in the ICU population were dominated by acute physiological changes, especially in level of consciousness (RASS), while predisposing factors such as age, dementia, and comorbidities figured more prominently in prediction of delirium outside of critical care (Fig. 4). A greater number of variables reached the same importance threshold in general medical wards.

Discussion

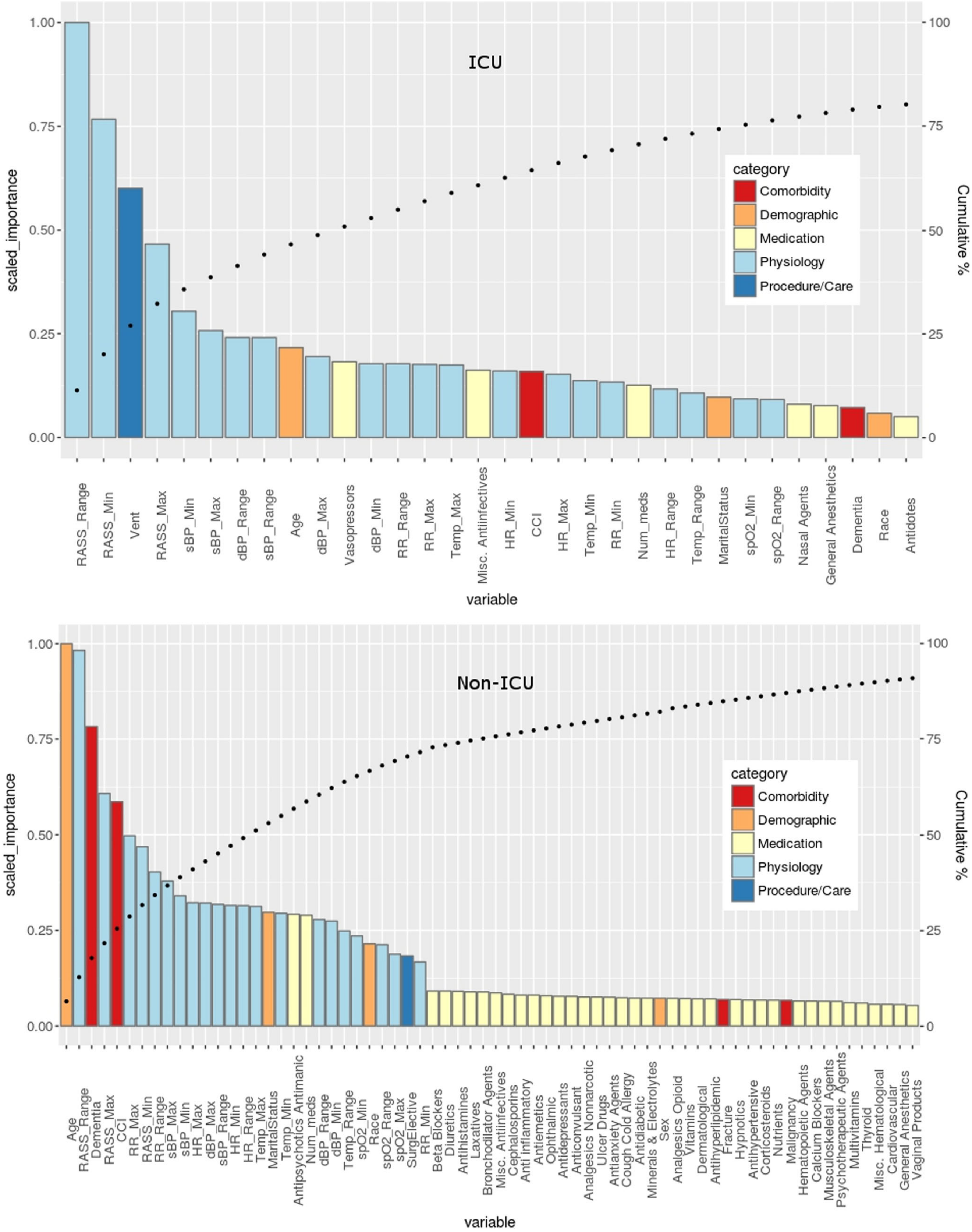
This proof-of-concept study demonstrates the value of machine learning applied to large, complex clinical data sets. Generally applicable and accurate prediction of patients at high risk of developing delirium would be of great value in identifying underlying medical conditions and modifiable risk factors. Reduction of hospital delirium rates and time spent in delirium would improve outcomes for patients, as well as decrease overall care burden and associated costs for the healthcare system. The factors incorporated in this model may be predictive, even if they are not precipitating (e.g. changes in vital signs or level of consciousness). Unfortunately, not all precipitating factors are universally available for patients when they would be predictive. Often incident delirium is the clinical syndrome that prompts a search for the underlying cause, such as infection, organ dysfunction, toxicity, etc. The predictive model described here could be used to initiate those actions, rather than waiting until there is frank delirium.

Fig. 4 Performance of the individual predictive models for ICU (top graph) and non-ICU (bottom graph) patient populations. Variable importance (bars) is scaled relative to the most important variable in the model. The minimum cutoff for reporting in the graph was a scaled importance of 0.05. Descriptions of the variables are found in supplemental data. The variables are colored by general category. Also plotted is the cumulative percentage of the model accounted for by the variables (points). *RASS* = Richmond Agitation Sedation Scale, *RR* = respiratory rate, *sBP* = systolic blood pressure, *dBP* = diastolic blood pressure, *HR* = heart rate, *spO2* = blood oxygen saturation, *Temp* = body temperature, *CCI* = Charlson Comorbidity Index

The overall accuracy achieved by the RF model described here is comparable to or exceeds those of other published predictive models for delirium, despite being applied to a diverse patient population. Two of the better performing reported models, one for general surgery patients and one for critical care patients, used either larger numbers of predictor variables or patients than most other studies [17, 18]. Our model incorporates a much larger set of features and tens of thousands of patient visits. While some of these variables may not have been relevant to the final model, there are several reasons for their inclusion. First, Random Forests are generally unaffected by the addition of irrelevant variables, and are somewhat robust to collinearity due to the use of random variable subsets for tree splits [30]. Second, all of the data included in this model were readily available in the EHR. Finally, given the complex multifactorial etiology of delirium, there may be unexpected or underappreciated predictors. That being said, a simplified model can be generated using a reduced variable set based on variable importance. In fact, using the top 48 predictor variables ranked by relative importance (Table S1) yielded a model with comparable performance (not shown).

Hartford Hospital's delirium assessment program ('ADAPT' [31]) possesses two critical strengths that facilitated this predictive modeling effort. Having the CAM performed three times daily provides the temporal resolution needed to detect changes in a patient's condition on the scale of hours. Delirium is a dynamic syndrome, and the precipitating factors that contribute to risk are changing on the same time scale. In addition, the RASS provides an independent measure of the level of arousal, while also being used as a component of the CAM. Our model suggests that longitudinal RASS is a sensitive indicator of delirium risk, and this observation is consistent with published studies [32–34].

Predictive modeling approaches offer different strengths, with a trade-off between complexity and interpretability. In the case of medical prediction, especially of an urgent condition associated with poor outcomes, accuracy should be paramount. However, in the clinical setting trust in diagnostic or prognostic tools is also important. For that reason, Random Forests provide a good balance of accuracy and understanding. While they do not provide simple rules or risk



categories, they do allow a 'peek under the hood' to see which variables are influencing prediction. If the important factors identified are consistent with clinical knowledge and intuition, there will be more confidence in the model as a decision support tool. In addition, Random Forests can accommodate many continuous and categorical variables without the need for variable transformation or selection, can model complex interactions, and have few parameters to tune.

As a retrospective study, there were limitations imposed by the availability of structured data for inclusion. One salient omission was that of laboratory test values, especially given the identification of specific measures associated with delirium risk (e.g. BUN/creatinine ratio, electrolytes). While most inpatients had test values available, many did not have sufficient longitudinal data prior to delirium onset to capture the dynamic changes that may precede the condition. We found that exclusion of laboratory test values significantly increased the patient sample size available for modeling, and there was no subsequent loss in prediction accuracy.

An additional limitation of using available data is the reliability of the outcome measure – i.e. the delirium assessments. The CAM results used in this study were performed by many clinical staff members in the routine care of hospital patients with no independent validation. As such, there will be some positive assessments that may have been made in error; either due to confounding symptoms of dementia or psychoses, or to an incomplete match to the criteria for a positive CAM. Conversely, there will undoubtedly be unrecognized delirium among the negative CAM group. One result of the CAM is ‘unable to assess’ (UTA), a rating that should be reserved for cases in which the patient is not able to acknowledge the presence of the examiner. However, we previously found instances of the use of UTA outside of critical care suggesting that it may mask cases of delirium, based on additional patient assessments and outcomes [35]. To reduce the impact of these ambiguous results, we confined our Negative delirium group to those patients who had exclusively negative CAM results (i.e. no CAM UTA).

Recent applications of machine learning (including but not limited to Random Forests) to medical data have shown great promise, from imaging diagnostics [36, 37] to prediction of poor outcomes associated with a variety of conditions [38–42]. During the preparation of this report, Kramer et al. published a preliminary comparison of multivariate models to predict delirium in a large patient population where Random Forest performed well [43]; however, there were insufficient data available for a comparison with our model. Two additional recent delirium studies employing machine learning were published while this manuscript was under review. Wong et al. compared several algorithms (including Random Forest) to predict risk for incident delirium using only information available within the first 24 h of admission [44]. Unlike our model, the authors excluded high risk populations in critical care or with known cognitive impairment. We believe that our model allows for more dynamic prediction, using updated clinical information, and that individual patient risk prediction is still very valuable even in known high risk populations. In the study by Halladay et al., the authors used Random Forest to identify the most important factors useful as a diagnostic aid in recognizing prevalent delirium, rather than as a predictive model for incident delirium [45].

Our machine learning approach, while displaying excellent internal validation with the largest training cohort that we are aware of, will require prospective validation to be considered for a real-time implementation. Random Forests do not yield simple scoring systems or clinical decision rules; and therefore, we expect that it is the approach itself that will be generalizable. Machine learning models can be trained and optimized using local data, accounting for differences in assessment measures, reporting formats, and institutional protocols. This generally requires more automation and processing to function as effective decision support tools. However, the accuracy afforded by such models could justify the additional implementation effort.

Compliance with Ethical Standards

Conflicts of Interest The authors do not have any conflicts of interest to declare. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent For this type of study formal consent is not required.

References

1. Inouye, S. K., Westendorp, R. G. J., and Saczynski, J. S., Delirium in elderly people. Lancet 383:911–922, 2014. https://doi.org/10.1016/S0140-6736(13)60688-1.
2. Inouye, S. K., Rushing, J. T., Foreman, M. D. et al., Does delirium contribute to poor hospital outcomes?: A three-site epidemiologic study. J. Gen. Intern. Med. 13:234–242, 1998. https://doi.org/10.1046/j.1525-1497.1998.00073.x.
3. Ely, E. W., Gautam, S., Margolin, R. et al., The impact of delirium in the intensive care unit on hospital length of stay. Intensive Care Med. 27:1892–1900, 2001. https://doi.org/10.1007/s00134-001-1132-2.
4. Witlox, J., Eurelings, L. S. M., De Jonghe, J. F. M. et al., Delirium in Elderly Patients and the Risk of Postdischarge Mortality. JAMA 304:443–451, 2010. https://doi.org/10.1001/jama.2010.1013.
5. Rudolph, J. L., and Marcantonio, E. R., Postoperative Delirium: Acute change with long-term implications. Anesth. Analg. 112:1202–1211, 2011. https://doi.org/10.1213/ANE.0b013e3182147f6d.
6. Girard, T. D., Jackson, J. C., Pandharipande, P. P. et al., Delirium as a predictor of long-term cognitive impairment in survivors of critical illness. Crit. Care Med. 38:1513–1520, 2010. https://doi.org/10.1097/CCM.0b013e3181e47be1.

7. Langan, C., Sarode, D. P., Russ, T. C. et al., Psychiatric symptomatology after delirium: a systematic review. Psychogeriatrics, 2017. https://doi.org/10.1111/psyg.12240.
8. Davis, D. H. J., Muniz-Terrera, G., Keage, H. A. D. et al., Association of Delirium With Cognitive Decline in Late Life. JAMA Psychiatry 74:244, 2017. https://doi.org/10.1001/jamapsychiatry.2016.3423.
9. Brown, E., and Douglas, V., Moving Beyond Metabolic Encephalopathy: An Update on Delirium Prevention, Workup, and Management. Semin. Neurol. 35:646–655, 2015. https://doi.org/10.1055/s-0035-1564685.
10. Hshieh, T. T., Yue, J., Oh, E. et al., Effectiveness of Multicomponent Nonpharmacological Delirium Interventions. JAMA Intern. Med. 175:512, 2015. https://doi.org/10.1001/jamainternmed.2014.7779.
11. Lawlor, P. G., and Bush, S. H., Delirium diagnosis, screening and management. Curr Opin Support Palliat Care 8:286–295, 2014. https://doi.org/10.1097/SPC.0000000000000062.
12. Inouye, S. K., Van Dyck, C. H., Alessi, C. A. et al., Clarifying confusion: The confusion assessment method: A new method for detection of delirium. Ann. Intern. Med. 113:941–948, 1990. https://doi.org/10.7326/0003-4819-113-12-941.
13. Waszynski, C., and Petrovic, K., Nurses' evaluation of the Confusion Assessment Method: a pilot study. J. Gerontol. Nurs. 34:49–56, 2008. https://doi.org/10.3928/00989134-20080401-06.
14. Ely, E. W., Margolin, R., Francis, J. et al., Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). Crit. Care Med. 29:1370–1379, 2001.
15. Inouye, S. K., Kosar, C. M., Tommet, D. et al., The CAM-S: Development and validation of a new scoring system for delirium severity in 2 cohorts. Ann. Intern. Med. 160:526–533, 2014. https://doi.org/10.7326/M13-1927.
16. Van Meenen, L. C. C., van Meenen, D. M. P., de Rooij, S. E., and ter Riet, G., Risk prediction models for postoperative delirium: A systematic review and meta-analysis. J. Am. Geriatr. Soc. 62:2383–2390, 2014. https://doi.org/10.1111/jgs.13138.
17. van den Boogaard, M., Pickkers, P., Slooter, A. J. C. et al., Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. BMJ 344:e420–e420, 2012. https://doi.org/10.1136/bmj.e420.
18. Kim MY, Park UJ, Kim HT, Cho WH (2016) DELirium Prediction Based on Hospital Information (Delphi) in General Surgery Patients. 95:1–7. doi: https://doi.org/10.1097/MD.0000000000003072
19. Kennedy, M., Enander, R. A., Tadiri, S. P. et al., Delirium risk prediction, healthcare use and mortality of elderly adults in the emergency department. J. Am. Geriatr. Soc. 62, 2014. https://doi.org/10.1111/jgs.12692.
20. Breiman, L., Random Forests. Mach. Learn. 45:5–32, 2001. https://doi.org/10.1023/A:1010933404324.
21. Quan, H., Li, B., Couris, C. M. et al., Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am. J. Epidemiol. 173:676–682, 2011. https://doi.org/10.1093/aje/kwq433.
22. Wasey JO (2016) ICD: Tools for Working with ICD-9 and ICD-10 Codes, and Finding Comorbidities
23. Team TH a. (2017) h2o: R Interface for H2O
24. Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data. Univ California, Berkeley 1–12. doi: ley.edu/sites/default/files/tech-reports/666.pdf
25. Khalilia, M., Chakraborty, S., and Popescu, M., Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11:51, 2011. https://doi.org/10.1186/1472-6947-11-51.
26. Platt, J., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classif 10:61–74, 1999. 10.1.1.41.1639.
27. Kuhn M (2008) Building Predictive Models in R Using the caret Package. J Stat Software, Artic 28:1–26. https://doi.org/10.18637/jss.v028.i05.
28. Robin, X., Turck, N., Hainard, A. et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77, 2011. https://doi.org/10.1186/1471-2105-12-77.
29. Grau, J., Grosse, I., and Keilwagen, J., PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics 31:2595–2597, 2015. https://doi.org/10.1093/bioinformatics/btv153.
30. Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R., Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2:493–507, 2012. https://doi.org/10.1002/widm.1072.
31. Semancik, L., Waszynski, C., and Udeh, E., Delirium in hospitalized patients: recognition, prevention, and management. Conn. Med. 78:105–109, 2014.
32. Tieges, Z., McGrath, A., Hall, R. J., and MacLullich, A. M. J., Abnormal Level of Arousal as a Predictor of Delirium and Inattention: An Exploratory Study. Am. J. Geriatr. Psychiatry 21:1244–1253, 2013. https://doi.org/10.1016/j.jagp.2013.05.003.
33. Han, J. H., Vasilevskis, E. E., Schnelle, J. F. et al., The Diagnostic Performance of the Richmond Agitation Sedation Scale for Detecting Delirium in Older Emergency Department Patients. Acad. Emerg. Med. 22:878–882, 2015. https://doi.org/10.1111/acem.12706.
34. Morandi, A., Han, J. H., Meagher, D. et al., Detecting Delirium Superimposed on Dementia: Evaluation of the Diagnostic Performance of the Richmond Agitation and Sedation Scale. J. Am. Med. Dir. Assoc. 17:828–833, 2016. https://doi.org/10.1016/j.jamda.2016.05.010.
35. Corradi, J. P., Chhabra, J., Mather, J. F. et al., Analysis of multidimensional contemporaneous EHR data to refine delirium assessments. Comput. Biol. Med. 5799:1–24, 2016. https://doi.org/10.1016/j.compbiomed.2016.06.013.
36. Gulshan, V., Peng, L., Coram, M. et al., Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama 304:649–656, 2016. https://doi.org/10.1001/jama.2016.17216.
37. Dawes, T. J. W., de Marvao, A., Shi, W. et al., Machine Learning of Three-dimensional Right Ventricular Motion Enables Outcome Prediction in Pulmonary Hypertension: A Cardiac MR Imaging Study. Radiology 161315, 2017. https://doi.org/10.1148/radiol.2016161315.
38. Ghose, S., Mitra, J., Khanna, S., and Dowling, J., An Improved Patient-Specific Mortality Risk Prediction in ICU in a Random Forest Classification Framework. Stud Health Technol Inform 214:56–61, 2015. https://doi.org/10.3233/978-1-61499-558-6-56.
39. Futoma, J., Morris, J., and Lucas, J., A comparison of models for predicting early hospital readmissions. J. Biomed. Inform. 56:229–238, 2015. https://doi.org/10.1016/j.jbi.2015.05.016.
40. Kessler, R. C., Warner, C. H., Ivany, C. et al., Predicting Suicides After Psychiatric Hospitalization in US Army Soldiers: The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). JAMA Psychiatry 72:1–9, 2014. https://doi.org/10.1001/jamapsychiatry.2014.1754.
41. Mortazavi, B. J., Downing, N. S., Bucholz, E. M. et al., Analysis of Machine Learning Techniques for Heart Failure Readmissions. Circ Cardiovasc Qual Outcomes 9:629–640, 2016. https://doi.org/10.1161/CIRCOUTCOMES.116.003039.

42. Taylor, R. A., Pare, J. R., Venkatesh, A. K. et al., Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad. Emerg. Med. 23:269–278, 2016. https://doi.org/10.1111/acem.12876.

43. Kramer, D., Veeranki, S., Hayn, D., and Quehenberger, F., Development and Validation of a Multivariable Prediction Model for the Occurrence of Delirium in Hospitalized Gerontopsychiatry and Internal Medicine Patients. Stud Health Technol Inform:32–39, 2017. https://doi.org/10.3233/978-1-61499-759-7-32.

44. Wong, A., Young, A. T., Liang, A. S. et al., Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. JAMA Netw. Open 1:e181018, 2018. https://doi.org/10.1001/jamanetworkopen.2018.1018.

45. Halladay, C. W., Sillner, A. Y., and Rudolph, J. L., Performance of Electronic Prediction Rules for Prevalent Delirium at Hospital Admission. JAMA Netw. Open 1:e181405, 2018. https://doi.org/10.1001/JAMANETWORKOPEN.2018.1405.