

# 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



Annie M. Racine, PhD<sup>1,2</sup>, Douglas Tommet, MS<sup>3</sup>, Madeline L. D'Aquila, BS<sup>1</sup>, Tamara G. Fong, MD, PhD<sup>1,2,4</sup>, Yun Gou, MA<sup>1</sup>, Patricia A. Tabloski, GNP-BC, PhD<sup>5</sup>, Eran D. Metzger, MD<sup>2,6</sup>, Tammy T. Hshieh, MD, MPH<sup>2,7</sup>, Eva M. Schmitt, PhD<sup>1</sup>, Sarinnapha M. Vasunilashorn, PhD<sup>2,7</sup>, Lisa Kunze, MD<sup>2,8</sup>, Kamen Vlassakov, MD<sup>2,5</sup>, Ayesha Abdeen, MD<sup>2,9</sup>, Jeffrey Lange, MD<sup>2,10</sup>, Brandon Earp, MD<sup>2,11</sup>, Bradford C. Dickerson, MD<sup>12</sup>, Edward R. Marcantonio, MD, SM<sup>1,2,7</sup>, Jon Steingrimsson, PhD<sup>13</sup>, Thomas G. Trivison, PhD<sup>1,2</sup>, Sharon K. Inouye, MD, MPH<sup>1,2,7</sup>, and Richard N. Jones, ScD<sup>3</sup> , the RISE Study Group Steven Arnold, Bradford Dickerson, Tamara Fong, Richard Jones, Towia Libermann, Thomas Trivison, Simon T. Dillon, Jacob Hooker, Tammy Hshieh, Long Ngo, Hasan Otu, Annie Racine, Alexandra Touroutoglou, Sarinnapha Vasunilashorn, Douglas Ayres, Gregory Brick, Antonia Chen, Robert Davis, Jacob Drew, Richard Iorio, Fulton Kornack, Michael Weaver, Anthony Webber, Richard Wilk, David Shaff, Brett Armstrong, Angelee Banda, Sylvie Bertrand, Madeline D'Aquila, Jacqueline Gallagher, Baileigh Hightower, Shannon Malloy, Jacqueline Nee, Chloe Nobuhara, Abigail Overstreet, Bianca Trombetta, Baileigh Hightower, David Urick, Guoquan Xu, Grae Arabasz, Michael Brickhouse, Regan Butterfield, Shirley Hsu, Sara Makaretz, Judith Sore, Fan Chen, Sabrina Carretie, Ted Gruen and Katherine Tasker

<sup>1</sup> Aging Brain Center, Institute for Aging Research, Boston, MA, USA; <sup>2</sup> Harvard Medical School, Boston, MA, USA; <sup>3</sup> Department of Psychiatry & Human Behavior, and Neurology, Brown University Warren Alpert Medical School, Providence, RI, USA; <sup>4</sup> Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA; <sup>5</sup> William F Connell School of Nursing at Boston College, Boston, MA, USA; <sup>6</sup> Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA; <sup>7</sup> Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA; <sup>8</sup> Department of Anesthesia, Beth Israel Deaconess Medical Center, Boston, MA, USA; <sup>9</sup> Department of Orthopedic Surgery, Beth Israel Deaconess Medical Center, Boston, MA, USA; <sup>10</sup> Department of Orthopedic Surgery, Brigham and Women's Hospital, Boston, MA, USA; <sup>11</sup> Department of Orthopedics, Brigham and Women's Faulkner Hospital, Boston, MA, USA; <sup>12</sup> Department of Neurology and Massachusetts Alzheimer's Disease Research Center, Massachusetts General Hospital, Boston, MA, USA; <sup>13</sup> Department of Biostatistics, Brown University, Providence, RI, USA.

**BACKGROUND:** Our objective was to assess the performance of machine learning methods to predict post-operative delirium using a prospective clinical cohort.

**METHODS:** We analyzed data from an observational cohort study of 560 older adults ( $\geq 70$  years) without dementia undergoing major elective non-cardiac surgery. Post-operative delirium was determined by the Confusion Assessment Method supplemented by a medical chart review ( $N = 134$ , 24%). Five machine learning algorithms and a standard stepwise logistic regression model were

developed in a training sample (80% of participants) and evaluated in the remaining hold-out testing sample. We evaluated three overlapping feature sets, restricted to variables that are readily available or minimally burdensome to collect in clinical settings, including interview and medical record data. A large feature set included 71 potential predictors. A smaller set of 18 features was selected by an expert panel using a consensus process, and this smaller feature set was considered with and without a measure of pre-operative mental status.

**RESULTS:** The area under the receiver operating characteristic curve (AUC) was higher in the large feature set conditions (range of AUC, 0.62–0.71 across algorithms) versus the selected feature set conditions (AUC range, 0.53–0.57). The restricted feature set with mental status had intermediate AUC values (range, 0.53–0.68). In the full feature set condition, algorithms such as gradient boosting, cross-validated logistic regression, and neural network (AUC = 0.71, 95% CI 0.58–0.83) were comparable

---

Sharon K. Inouye and Richard N. Jones contributed equally as co-senior authors

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11606-020-06238-7>) contains supplementary material, which is available to authorized users.

---

Received June 26, 2019

Accepted September 11, 2020

with a model developed using traditional stepwise logistic regression (AUC = 0.69, 95% CI 0.57–0.82). Calibration for all models and feature sets was poor.

**CONCLUSIONS:** We developed machine learning prediction models for post-operative delirium that performed better than chance and are comparable with traditional stepwise logistic regression. Delirium proved to be a phenotype that was difficult to predict with appreciable accuracy.

**KEY WORDS:** machine learning; statistical learning; model prediction; delirium; post-operative.

## Abbreviations

BIDMC	Beth Israel Deaconess Medical Center
BWH	Brigham and Women's Hospital
HMS	Harvard Medical School
HSL	Hebrew SeniorLife
MGH	Massachusetts General Hospital
PI	principal investigator
UNL	University of Nebraska-Lincoln

J Gen Intern Med

DOI: 10.1007/s11606-020-06238-7

© Society of General Internal Medicine 2020

## INTRODUCTION

Delirium, the most common post-operative complication in adults over the age of 65, has an incidence of 15–25% after major elective surgery.<sup>1</sup> Delirium is associated with both short- and long-term clinical and functional complications<sup>2–9</sup> and greater risk for developing dementia.<sup>10,11</sup> Delirium prediction algorithms could help pre-operatively stratify patients according to risk for delirium to improve patient care, reduce risk of adverse outcomes, and facilitate enrollment into clinical trials.

Multiple approaches have previously been proposed to predict delirium,<sup>12–17</sup> but few studies have employed machine learning (ML) algorithms.<sup>18–23</sup> ML methods are optimally applied when there is an abundance of data.<sup>24</sup> Nevertheless, it is important to consider if ML algorithms can be usefully applied to smaller datasets, which are more common in clinical cohort studies of disorders like delirium. Generally, situations that result in high-volume data rely upon administrative data or routinely collected clinical data that may suffer from measurement error to a greater degree than purposefully collected research data.<sup>25</sup> Our objective was to identify the optimal ML approach to predict delirium in a rigorous and well-characterized, prospective, observational cohort study of delirium, and to compare it with a traditional statistical prediction model.

We analyzed data from the Successful Aging after Elective Surgery (SAGES) study, which used reference standard approaches to assess both pre-operative cognitive function and post-operative delirium,<sup>26–28</sup> which are often not available in large datasets that rely primarily on electronic health records (EHR). Based on prior work,<sup>17,23,29</sup> we hypothesized that we could identify a ML model to predict delirium with an area under the receiver operating characteristic curve (AUC)

greater than 0.70, indicating good diagnostic accuracy,<sup>30</sup> and that this ML model would have a higher AUC than a model derived using stepwise logistic regression. Given the known importance of cognitive function to delirium prediction,<sup>31–33</sup> we further sought to determine the extent that prediction can be improved with the inclusion of a measure of pre-operative cognitive function (a variable that may not always be available in the pre-operative setting) in the feature set (i.e., list of predictors).

## METHODS

### Study Population

The SAGES study design and methods have been described in detail previously.<sup>26,27</sup> In brief, eligible participants were aged 70 years and older, English speaking, scheduled to undergo elective surgery at one of two Harvard-affiliated academic medical centers, and with an anticipated length of stay of at least 3 days. Eligible surgical procedures included the following: total hip or knee replacement; lumbar, cervical, or sacral laminectomy; lower extremity arterial bypass; open abdominal aortic aneurysm repair; and open or laparoscopic colectomy. Exclusion criteria were dementia, delirium, hospitalization within the past 3 months, terminal condition, legal blindness, severe deafness, history of schizophrenia or psychosis, and history of alcohol abuse or withdrawal. A total of 560 patients met all eligibility criteria and were enrolled between June 18, 2010, and August 8, 2013. Written informed consent for study participation was obtained from all participants according to procedures approved by the institutional review boards of Beth Israel Deaconess Medical Center and Brigham and Women's Hospital, the two study hospitals, and Hebrew SeniorLife, the coordinating center for the study.

### Data Collection

Participants underwent baseline assessment in their homes approximately 2 weeks (mean [standard deviation] 13 [15] days) prior to surgery.<sup>26</sup> All study interviews were conducted by experienced interviewers who underwent 2–4 weeks of intensive training and standardization. Inter-rater reliability assessment and standardization on all key study variables, including delirium assessment, was conducted every 6 months throughout the study and coding questions were addressed in weekly meetings of all study staff. Medical records were reviewed by study clinicians to collect information on surgical procedure, anesthesia type and duration, abnormal laboratory results, baseline diagnoses, development of delirium, precipitating factors for delirium (e.g., medications, iatrogenic events, or catheters), post-operative complications, and death.<sup>26</sup> Chart abstraction data were randomly checked for illogical values and against data collected as part of the screening process (e.g., surgery type). In addition, a 10% subset of charts underwent re-abstraction for reliability checks.<sup>26</sup>

## Assessment of Delirium

The delirium assessment, which took 10–15 min, included daily brief cognitive testing,<sup>27,34</sup> Delirium Symptom Interview (DSI),<sup>35</sup> and family and nurse interviews conducted from the first postoperative day until discharge. Delirium was rated using the Confusion Assessment Method (CAM).<sup>36</sup> The CAM is a standardized approach with high sensitivity (94–100%) and specificity (90–95%) in prior studies.<sup>37,38</sup> Inter-rater reliability was high in SAGES (kappa statistic = 0.92 in 71 paired ratings).<sup>26</sup> The DSI was used to rate CAM symptoms. An established chart review method was used to capture delirium symptoms between interviews.<sup>28,39</sup> Patients were classified as delirious if either the CAM or chart review criteria were met. The procedure results in approximately 83% of identified cases as resulting from the patient assessment (31% of which are also identified by the chart review), and 17% of cases of delirium are identified through chart review and not detected by patient assessment.<sup>28</sup> Given an overall incidence of delirium of 24%, this implies an incidence of CAM delirium of 20% and an incidence of chart delirium of about 10%.

## Identification and Formalization of the Predictor Variable Set

Medical records were reviewed with a comprehensive medical record abstraction tool to collect information on the surgical procedure, anesthesia type and duration, baseline diagnoses and comorbidity, abnormal laboratory results, development of delirium, precipitating factors for delirium (e.g., medications, iatrogenic events, catheters, or physical restraints), post-operative complications, and intercurrent illnesses.<sup>27</sup> From this information set, we identified features for use in our predictive models. Potential predictors were required either to be readily available in a clinical setting through existing sources (e.g., medical record or standard laboratory data) or through quick screening tests that would be feasible in a busy clinical setting. We decided that although pre-surgical medication use could be predictive of post-operative delirium risk, the process of identifying predictors from among the multitude of medications in various formulations and dosages would require extensive pre-processing and would not satisfy the criterion of being readily available in a clinical setting. Of the remaining potential features, 71 pre-operative variables were selected and included demographic characteristics, lifestyle factors, cognitive function, physical function, psychosocial factors, frailty, sensory function, medical conditions, and laboratory values (Appendix Table 1). We will refer to this set of 71 variables as the *full feature set*. Missing data in the feature set were multiply imputed by chained equations.

In addition to using the *full feature set*, we identified a *selected feature set* with a more accessible number of predictors. Features were selected based on a clinician expert review process consisting of two iterative rounds of review by clinicians (S.K.I., T.G.F., T.T.H., E.D.M., E.R.M., P.A.T.) with expertise in delirium, neurology, geriatrics, geriatric

psychiatry, general medicine, and nursing. The final set of 18 predictors is described in Table 1. Because of the known importance of cognitive function in delirium prediction,<sup>31</sup> we performed analyses of the selected feature set with and without a summary score from a brief mental status test, the modified mini-mental state examination (3MS).<sup>40</sup> Thus, all analyses were performed using three overlapping feature sets: (1) a *selected feature set* ( $q = 18$  features) selected by an expert panel using an iterative process; (2) the *selected feature set plus 3MS* ( $q = 19$  features); and (3) the *full feature set* ( $q = 71$  features).

## Machine Learning Algorithms and Comparison Statistical Prediction Model

ML algorithms for prediction of delirium included cross-validated logistic, gradient boosting, neural network, random forest, and regularized regression (least absolute shrinkage and selection (LASSO) and ridge regularization).<sup>41–44</sup> In addition, we assessed model performance with two ensemble approaches. Ensemble methods combine multiple ML

Table 1 Pre-operative Patient Characteristics

Patient characteristic	Dataset		
	Total sample ( $N = 560$ )	Training dataset ( $n = 449$ )	Test/validation dataset ( $n = 111$ )
Delirium ( $N$ (%))	134 (24%)	108 (24%)	26 (23%)
Age (years) (mean (SD))	76.6 (5.2)	76.7 (5.1)	76.4 (5.3)
Female sex ( $N$ (%))	326 (58%)	255 (57%)	71 (64%)
Nonwhite race ( $N$ (%))	42 (8%)	32 (7%)	10 (9%)
Education (years) (mean (SD))	15.0 (2.9)	14.9 (2.9)	15.2 (3.1)
Past smoker ( $N$ (%))	312 (56%)	251 (56%)	61 (55%)
Current smoker ( $N$ (%))	26 (5%)	21 (5%)	5 (5%)
Alcohol $\geq 5$ times/week ( $N$ (%))	106 (19%)	83 (19%)	23 (21%)
Hearing impairment ( $N$ (%))	182 (33%)	143 (32%)	39 (35%)
Charlson Comorbidity Index score $\geq 2$ ( $N$ (%))	164 (29%)	130 (29%)	34 (31%)
ASA class $\geq 3$ ( $N$ (%))	353 (63%)	287 (64%)	66 (60%)
3MS score (mean (SD))*	93.5 (5.4)	93.4 (5.4)	93.9 (5.1)
Surgery type ( $N$ (%))			
Orthopedic	454 (81%)	360 (80%)	94 (85%)
Vascular	35 (6%)	30 (7%)	5 (5%)
Gastrointestinal	71 (13%)	59 (13%)	12 (11%)
White blood count (mean (SD))	7.2 (2.1)	7.2 (2.2)	7.1 (1.9)
Hematocrit (mean (SD))	39.5 (3.7)	39.6 (3.8)	39.1 (3.4)
Creatinine (mean (SD))	1.0 (0.3)	1.0 (0.3)	1.0 (0.3)
Sodium (mean (SD))	139.5 (2.6)	139.5 (2.6)	139.2 (2.6)
Oxygen saturation (mean (SD))	98.0 (1.4)	98.0 (1.4)	97.9 (1.6)
BUN/creatinine ratio (mean (SD))	22.5 (6.3)	22.4 (6.6)	22.9 (5.2)

ASA, American Society of Anesthesiologists Physical Status Classification; 3MS, Modified Mini-Mental State Examination; BUN, blood urea nitrogen

\*Range 0–100. Lower scores indicate greater cognitive impairment

algorithms to obtain better predictive performance.<sup>45</sup> We considered two relatively straightforward ensemble methods. The first ensemble method results in a positive test (i.e., predicts delirium) if any of the 5 individual algorithms are positive (i.e., ensemble-union). The second ensemble method results in a positive test if a majority of tests ( $\geq 3$ ) are positive (i.e., ensemble-majority).

We used two strategies to compare our ML algorithms to standard approaches for delirium prediction. First, we evaluated prediction using standard backwards stepwise logistic regression. The starting model included all predictors, and each subsequent step eliminated the predictor that resulted in the largest improvement to the Akaike information criterion (AIC), and terminated when the maximum AIC value was reached. Second, we used a previously published delirium risk prediction rule for hospitalized medical patients<sup>16</sup> to obtain predictions of delirium in the SAGES sample used for model testing. This predictive model uses vision impairment, severe illness, cognitive impairment, and high blood urea nitrogen/creatinine ratio to stratify patients according to risk for delirium.

## Analysis and Comparison of Models

To implement the learning algorithms, we split the SAGES sample into a training set (80%) to be used for model derivation and a testing set (20%) to be used for model validation. Random assignment to training/testing set was stratified on delirium status. For ML models, we performed repeated  $k$ -fold cross-validation ( $k=4$ , 10 repeats) to identify the optimal model parameters based on optimization of the AUC in the training set.<sup>46</sup> We compared models based on performance in the test set on the following criteria: AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), detection rate, and balanced accuracy. In order to compare these performance characteristics across models, we set the incidence rate at 25%, which is similar to the incidence in our sample (24%) rounded to the nearest 5%.

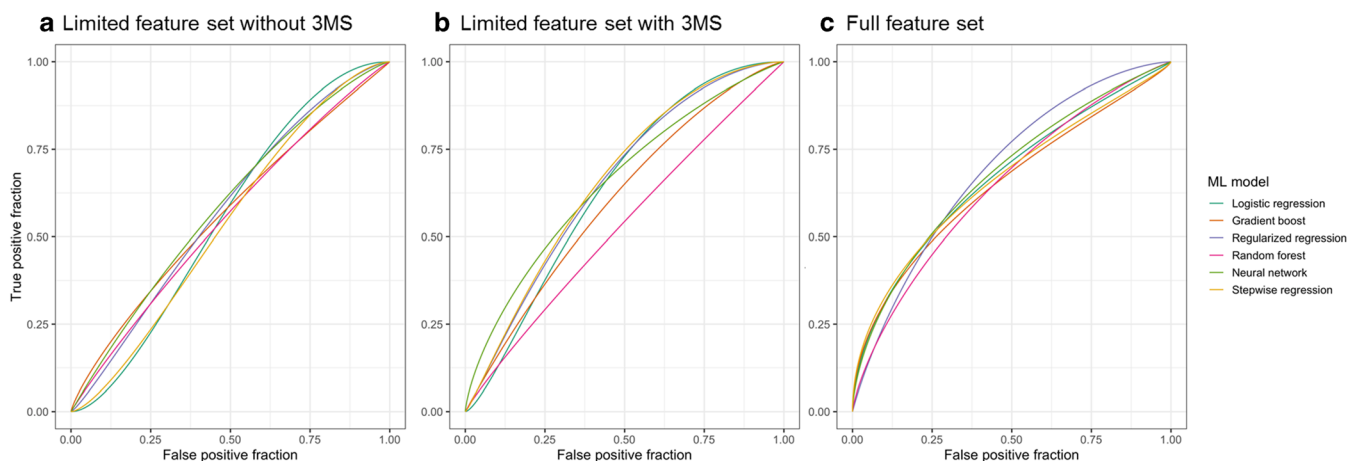
We examined overall calibration by plotting the distribution of predicted probabilities of delirium as a function of observed delirium for all algorithms using violin plots. We also generated calibration curves, which are plots of the observed proportion classified as delirious against the model-implied proportion with delirium-given predictions derived from a model. These figures are displayed in the [supplementary material](#). All analyses were conducted within the R computing programming environment (version 3.6.1, R Core Development Team, Vienna, Austria) using several different packages, including *caret*,<sup>47</sup> *nnet*,<sup>48</sup> *earth*,<sup>49</sup> *glmnet*,<sup>42</sup> *randomForest*,<sup>50</sup> *kernlab*,<sup>51</sup> and *gbm*.<sup>52</sup> Analysis code is available upon request.

## RESULTS

Patient characteristics, including the selected feature set, are described in Table 1. The training and test sets were selected at random, so any differences in the training and test datasets were due to chance. None of the differences across training and testing sets exceeds an effect size (Cohen's  $h$  for proportions,  $d$  for continuous variables<sup>53</sup>) exceeded 0.15, well below the conventional threshold for small effects. The mean effect size was 0.07 across all features. By design, the incidence of delirium was constant (minor variation due to rounding) in the full sample (24%), training sample (24%), and test sample (23%).

Figure 1 panels a, b, and c illustrate receiver operating characteristic (ROC) curves for predictive models under the selected, selected +3MS, and full feature sets, respectively. Under the selected feature set, all models perform comparably and near the main diagonal, indicating poor prediction relative to chance. The models improve with the addition of a measure of cognitive performance, 3MS (panel b), with the highest AUC observed using the full feature set (panel c).

Detailed results of the prediction modeling are summarized in Table 2. In this table, we summarize a number of statistics



**Fig. 1** Comparison of receiver operator curves (ROCs) for prediction of delirium by the various machine learning (ML) algorithms examined. **a** ROC curves when a measure of pre-operative cognitive function (3MS) was not included in the selected feature set; **b** ROC curves when 3MS was included in the selected feature set; **c** ROC curves for the full feature set.



Table 2 Comparison of Machine Learning Algorithms for Prediction of Delirium in Three Overlapping Feature Sets

Method	AUC	AUC 95% CI	SEN	SPEC	PPV	NPV	Detection prevalence*
Selected feature set ( $q = 18$ )							
Cross-validated logistic	0.55	(0.41, 0.67)	0.27	0.75	0.25	0.77	0.25
Gradient boosting	0.57	(0.44, 0.70)	0.38	0.79	0.36	0.81	0.25
Neural network	0.55	(0.40, 0.66)	0.31	0.76	0.29	0.78	0.25
Random forest	0.55	(0.42, 0.68)	0.27	0.75	0.25	0.77	0.25
Regularized regression	0.57	(0.44, 0.70)	0.31	0.76	0.29	0.78	0.25
Ensemble (union)	0.57	(0.45, 0.70)	0.62	0.52	0.28	0.81	0.51
Ensemble (majority)	0.53	(0.41, 0.67)	0.27	0.79	0.28	0.78	0.23
Stepwise logistic regression†	0.54	(0.39, 0.65)	0.31	0.75	0.28	0.78	0.26
Selected feature set + 3MS ( $q = 19$ )							
Cross-validated logistic	0.66	(0.50, 0.76)	0.35	0.78	0.32	0.80	0.25
Gradient boosting	0.57	(0.44, 0.70)	0.35	0.78	0.32	0.80	0.25
Neural network	0.65	(0.50, 0.75)	0.38	0.79	0.36	0.81	0.25
Random forest	0.53	(0.40, 0.66)	0.27	0.75	0.25	0.77	0.25
Regularized regression	0.66	(0.53, 0.78)	0.38	0.79	0.36	0.81	0.25
Ensemble (union)	0.61	(0.50, 0.75)	0.65	0.56	0.31	0.84	0.49
Ensemble (majority)	0.57	(0.46, 0.72)	0.35	0.80	0.35	0.80	0.23
Stepwise logistic regression†	0.68	(0.54, 0.79)	0.42	0.80	0.39	0.82	0.25
Full feature set ( $q = 71$ )							
Cross-validated logistic	0.70	(0.58, 0.82)	0.46	0.81	0.43	0.83	0.25
Gradient boosting	0.64	(0.54, 0.79)	0.46	0.81	0.43	0.83	0.25
Neural network	0.71	(0.58, 0.83)	0.50	0.82	0.46	0.84	0.25
Random forest	0.65	(0.51, 0.77)	0.42	0.80	0.39	0.82	0.25
Regularized regression	0.70	(0.56, 0.81)	0.50	0.82	0.46	0.84	0.25
Ensemble (union)	0.62	(0.51, 0.76)	0.62	0.62	0.33	0.84	0.43
Ensemble (majority)	0.67	(0.58, 0.83)	0.50	0.84	0.48	0.85	0.24
Stepwise logistic regression†	0.69	(0.57, 0.82)	0.46	0.81	0.43	0.83	0.25

3MS, modified mini-mental status examination; AUC, area under the receiver operating characteristic curve; CI, confidence interval; SEN, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value;  $q$ , number of features

All statistics reflect models estimated in training data and validated in a hold-out sample not included in the derivation

\*Performance characteristics are based on a set detection prevalence of 0.25. However, because the two Ensemble methods use a combination of approaches, the detection prevalence cannot be pre-defined for those approaches

†Backwards stepwise logistic regression is a statistical model (not a ML algorithm) that was evaluated for comparison

Calculation of summary statistics

SEN = Sensitivity =  $TP/(TP + FN)$

SPEC = Specificity =  $TN/(TN + FP)$

PPV = Positive predictive value =  $TP/(TP + FP)$

NPV = Negative predictive value =  $TN/(TN + FN)$

TP = true positive

TN = true negative

FP = false positive

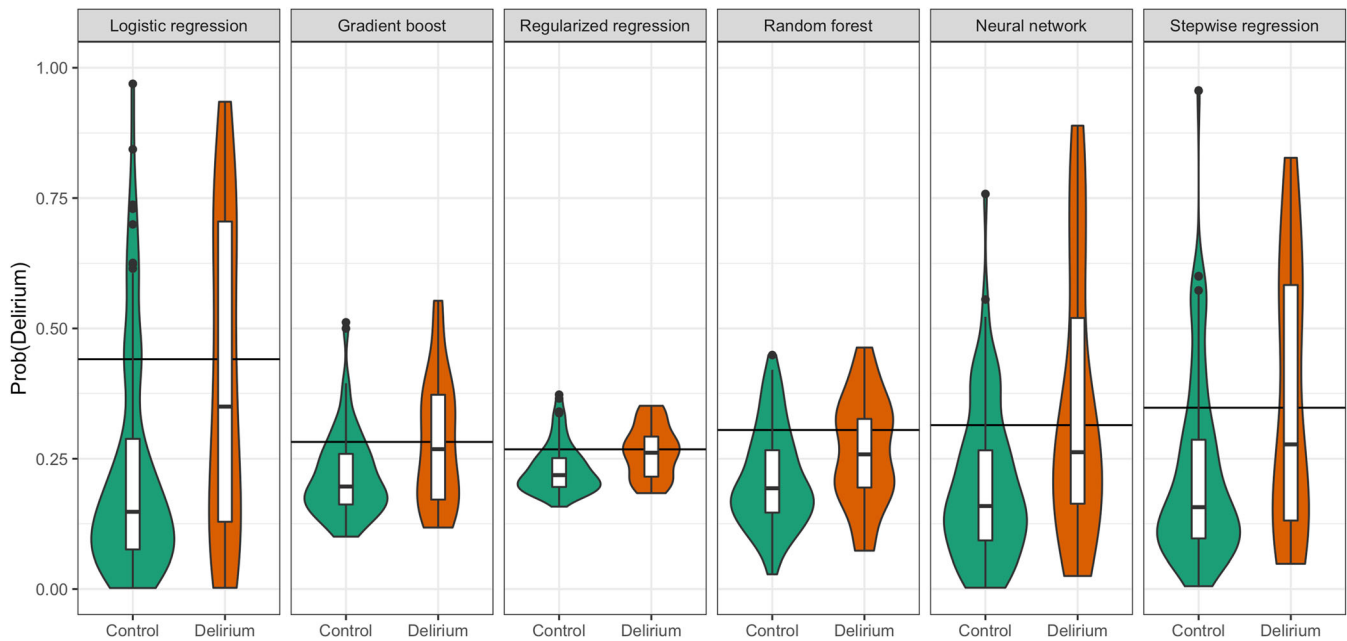
FN = false negative

that describe the algorithm and success in predicting delirium in the training data. These include the area under the receiver operating characteristic curve (AUC, and 95% confidence interval) which can be interpreted as the probability that if presented with a random case and random control, the case would have the higher predicted value. We also present standard confusion matrix results, including the sensitivity (proportion of cases that are predicted to be cases given the model), specificity (proportion of controls predicted to be controls), positive predictive value (proportion of those predicted to be cases that are actually cases), and negative predictive value (proportion of those predicted to be controls that are actually controls). The detection prevalence is the proportion of persons in the testing set classified as delirious with the algorithm, and was fixed to 25% for all algorithms other than the ensemble methods. With the *selected feature set* (3MS not included among the predictors), all models performed similarly and close to chance, with AUCs derived from the testing/validation data ranging from 0.53 to 0.57. With the *selected feature set + 3MS*, prediction improved for most algorithms, indicated by higher AUCs relative to the selected feature set

(Table 2). Using this set, stepwise logistic regression had the highest AUC (0.68) among all the predictive modeling approaches. Among the ML models, regularized regression and cross-validated logistic regression shared the highest AUC (0.66). Regularized regression had a positive predictive value (PPV) of 0.36, negative predictive value (NPV) of 0.81, sensitivity of 0.38, and specificity of 0.79.

Prediction was strongest when we used the 71 predictors in the *full feature set* (Table 2). The highest AUC (0.71) was observed for the neural network algorithm. When the detection prevalence was set to 25% for comparison across models, the neural network model had among the highest PPV (0.46), NPV (0.84), sensitivity (0.50), and specificity (0.82). The ensemble-union approach achieved the highest sensitivity (0.62). Note the ensemble-union approach is the only approach where the detection prevalence (the proportion of the sample identified as a probable case of delirium) deviates from 0.25. For all other algorithms, the detection prevalence was constrained to 0.25 by design, but this is not possible with the ensemble-union.

Similar to the ML models, delirium prediction using stepwise logistic regression was poor (AUC = 0.54; Fig. 1a;



**Fig. 2** Violin plots showing the distribution of the probability of delirium across ML models and stepwise logistic regression for the full feature set. In addition to a marker for the median of the data and a box indicating the interquartile range (as in standard box plots), these violin plots also show the kernel probability density of the data at different values for non-delirious patients (green) and delirious patients (salmon). The horizontal bar indicates a detection prevalence of 25%.

Table 2) without the inclusion of the 3MS. After 3MS was added to the potential predictor set, the stepwise logistic regression showed improved model performance (AUC = 0.68, sensitivity = 0.42; specificity = 0.80; PPV = 0.39; NPV = 0.82; Fig. 1b; Table 2).

Figure 2 illustrates the range and distribution of predicted probabilities of post-operative delirium among patients identified as having delirium and those without delirium for each of the predictive modeling approaches in the validation dataset (hold-out sample) and using the full feature set. The violin-like shape of the distributions among those who did not have delirium illustrates the relatively high levels of specificity of the predictive models; the more rectangular shape of the distributions among those who did have delirium illustrates the relatively low sensitivity of all predictive modeling approaches (Table 2). Model calibration is described more completely in the [Supplemental Appendix](#). Overall, model calibration was poor. This is important for situations in which the actual predicted value of the prediction model is to be used as an indicator of the predicted probability of delirium. Our evaluation models addressed poor calibration by forcing the detected prevalence to match the sample prevalence.

Finally, we used a published delirium prediction model<sup>16</sup> to classify hospitalized patients according to risk for delirium. This predictive model uses vision impairment, severe illness, cognitive impairment, and high blood urea nitrogen/creatinine ratio to stratify patients according to risk for delirium, and all but vision impairment were among the features in our full feature set. Vision impairment was considered a plausible risk factor by our clinical experts, but was dropped from the feature set due to a lack of variance. The published algorithm resulted

in identifying 59/111 (53%) of the testing set as delirious, and a positive predictive value of 25%, a sensitivity of 58%, and specificity of 48%. The AUC was 0.55.

## CONCLUSIONS

Using multiple ML approaches and a standard statistical technique, we were able to predict delirium with moderate accuracy based on variables that are readily available or minimally burdensome to collect in a clinical setting (e.g., on hospital admission). We demonstrated that ML methods can be used to develop prediction algorithms that perform better than chance but fail to demonstrate superior performance relative to models developed using stepwise logistic regression in hold-out validation data from a single clinical cohort study. Predictive accuracy of derived models had higher AUC than a single previously published predictive model derived in a different population. Nevertheless, predictive performance was modest, model calibration was poor, and the general pattern of results (e.g., Fig. 2) suggests that persons at low risk for delirium may tend to be alike (good pre-operative cognition, few vulnerability factors) but that persons who do develop delirium do so for widely varying reasons that are difficult to identify robustly. In agreement with a recently published systematic review, we find no strong evidence for benefit of ML over traditional logistic regression in developing a prediction rule.<sup>54</sup>

Strong prediction requires strong predictors. Delirium is by nature a heterogeneous, multifactorial condition, and

predictive models have been relatively limited in their overall performance and their ability to generalize across populations. In our sample, predictive performance was better when pre-operative cognitive function (3MS) was included in the feature set. The role of cognitive impairment in delirium prediction is well established.<sup>31,32</sup> Further improvements in performance were observed when the feature set included a large number of clinical variables, demonstrating the advantage of using ML for cohort studies with a large number of predictors and high-quality data. Modestly improved performance over stepwise regression supports the notion that ML approaches may help improve prediction, with the understanding that the full realization of these advantages may require more data than we had available.

Our study adds to a small but growing body of literature using ML to predict delirium. Among published studies, there is substantial variability in terms of sample size, identification of delirium, patient population, and types of ML algorithms evaluated. For instance, compared with prior studies, our sample size was much smaller ( $N=560$  compared with  $N=9221-64,237$ ). Prior studies variably defined delirium using ICD-10 codes,<sup>21,22</sup> CAM alone or CAM and some other instrument,<sup>19,20</sup> or DSM criteria.<sup>18</sup> The method for detection of delirium is of critical importance, as certain methods may be more or less sensitive and often disproportionately detect the subtypes of delirium, such as hyperactive delirium. Model development and performance will also vary substantially depending on the population studied and factors included to develop and test delirium prediction models. Although model prediction may be improved by considering post-baseline (but pre-delirium) precipitating factors occurring during hospitalization, these were not examined in the current study since information collected during hospitalization would not be available during the window to recruit patients into a clinical trial, which is the long-term goal of this research. Model performance has also varied across studies using either ML or statistical models to predict delirium, with AUCs ranging from 0.56 to 0.94; the results of the present study ( $AUC = 0.70$ ) fall within this range.<sup>17</sup> While the AUC is the most commonly used metric to quantify the quality of a predictive model,<sup>54</sup> the interpretation can be somewhat challenging. It is worth considering what the implications of a given AUC are under different conditions of disease and screen positive proportions. For instance, if both the prevalence and screen positive proportion of delirium are 25%, we estimate that the required AUC should be about 0.75 in order to achieve a PPV greater than 50%, but if the prevalence and screen positive proportion of delirium are 5%, an AUC of about 0.92 is needed to achieve a PPV of greater than 50%. The delirium prevalence was considerably higher in our study (24%) compared with rates observed in other ML studies of delirium (3–9%). Reasons for this are likely that ML is optimally applied in very large datasets, very large datasets typically derived from administrative or routine clinical assessment data rather than from highly controlled research protocols with extensive

training and data quality controls.<sup>25</sup> There are very likely many more false negatives in administrative and routine clinical datasets than what are observed in data derived from rigorous field studies.

This study has several strengths including evaluating multiple ML approaches in a well-characterized cohort with little missing data, reference standard determination of delirium, and inclusion of a measure of pre-operative cognitive function in the feature set, which sets our study apart from those relying solely on EHR data. This study also has a number of limitations. First, although the SAGES study is one of the largest studies of surgical patients with detailed pre-operative assessment, it is smaller than most datasets used for ML and a larger dataset would result in more stable parameter estimates, and better replication within model training and testing. Second, it is important to note that delirium prediction may have been improved by inclusion of additional variables including neuropsychological test scores, genetic information (e.g., APOE4), biomarkers that are not commonly evaluated clinically (e.g., C-reactive protein), or post-baseline and precipitating factors, especially medications. The absence of an external dataset for model validation, and our reliance on an internal hold-out sample for model testing, is not ideal. We elected not to include variables that would be difficult or time-consuming to collect prior to surgery in order to increase the clinical applicability of the results, including the potential to use these models for recruitment from clinical settings into large, multi-site clinical trials. Third, we compared our machine learning models with only one statistical algorithm (backwards stepwise logistic regression); it is possible that other statistical algorithms or different parameters could have improved its performance compared with the machine learning models.

In conclusion, we developed prediction models for post-operative delirium that performed better than chance. This study supports the notion that using available or minimally burdensome clinical data, machine learning, or more traditional stepwise logistic regression methods can be used to identify patients at high risk of developing delirium after surgery. These models could be used to identify high-risk persons for known delirium prevention interventions, or to optimize recruitment into clinical trials aimed at improving post-operative outcomes.

---

**Acknowledgments:** The authors gratefully acknowledge the contributions of the patients, family members, nurses, physicians, staff members, and members of the Executive Committee who participated in the Successful Aging after Elective Surgery (SAGES) Study and the Role of Inflammation after Surgery in Elders (RISE) study. Members of the RISE study group are listed in the [supplementary material](#). This work is dedicated to the memory of Joshua Bryan Inouye Helfand and Jane Ann McDowell. The authors would like to acknowledge all authors who fulfilled authorship criteria, as well as all members of the larger RISE Study group.

RISE Study Group

[Presented in alphabetical order; individuals listed may be part of multiple groups, but are listed only once under major activity, listed in parentheses].

Principal Investigator: Sharon K. Inouye, MD, MPH (Overall PI; HSL, BIDMC, HMS)



**Executive Committee:** Steven Arnold, MD (MGH); Bradford Dickerson, MD (MGH Site PI, HMS); Tamara Fong, MD, PhD (HMS, HSL, BIDMC); Richard Jones, ScD (Brown University); sTovia Libermann, PhD (HMS, BIDMC); Edward R. Marcantonio, MD, SM (BIDMC Site PI, HMS); Thomas Trivison, PhD (HSL, HMS)

**Co-Investigators:** Simon T. Dillon, PhD (HMS, BIDMC); Jacob Hooker, PhD, (MGH, HMS) Tammy Hsieh, MD (BWH); Long Ngo, PhD (HMS, BIDMC); Hasan Otu, PhD (UNL); Annie Racine, PhD (HMS, HSL); Eva M. Schmitt, PhD (Overall Project Director, HSL); Alexandra Touroutoglou, PhD (HMS, MGH); Sarinnapha Vasunilashorn, PhD (BIDMC)

**Surgical Leaders:** Ayesha Abdeen, MD (HMS, BIDMC); Douglas Ayres, MD (HMS, BIDMC); Brandon Earp, MD (HMS, BWH); Jeffrey Lange, MD (HMS, BWH)

**Surgeons:** Gregory Brick, MBChB (HMS, BWH); Antonia Chen, MD (HMS, BWH); Robert Davis, MD (HMS, BIDMC); Jacob Drew, MD (HMS, BIDMC); Richard Iorio, MD (HMS, BWH); Fulton Kornack, MD (HMS, BWH); Michael Weaver, MD (HMS, BWH); Anthony Webber, MD (HMS, BWH); Richard Wilk, MD (HMS, BWH)

**Anesthesiology Leaders:** Lisa Kunze, MD (BIDMC, HMS); David Shaff, MD (BWH, HMS); Kamen Vlassakov, MD (BWH, HMS)

**Epidemiology Core:** Brett Armstrong, MPH (BIDMC); Angelee Banda, MA (BIDMC); Sylvie Bertrand, BS (HSL); Madeline D'Aquila (HSL); Jacqueline Gallagher, MS (BIDMC); Baileigh Hightower, BA (MGH); Shannon Malloy, MA (BIDMC); Jacqueline Nee, BA (HSL); Chloe Nobuhara (MGH); Abigail Overstreet, MA (BIDMC); Bianca Trombetta (MGH); David Urlick (MGH); Guoquan Xu, MD, PhD (HSL)

**Biomedical Imaging Core:** Grae Arabasz (MGH); Michael Brickhouse (MGH); Regan Butterfield (MGH); Baileigh Hightower, BA (MGH); Shirley Hsu (MGH); Sara Makaretz (MGH); Judit Sore (MGH)

**Data Management and Statistical Analysis Core:** Fan Chen, MPH, MS (HSL); Yun Gou, MA (HSL); Douglas Tommet, MS (Brown University)

**Fiscal Management Committee:** Sabrina Carretie (HSL); Ted Gruen (HSL); Katherine Tasker (Chair, HSL)

**Funding** This work was supported by the Alzheimer's Drug Discovery Foundation, the National Institutes of Health (P01AG031720 (SKI), K07AG041835 (SKI), R24AG054259 (SKI), R01AG044518 (SKI/RNJ), R01AG030618 (ERM/DMF), K24AG035075 (ERM), T32AG023480 (AMR), K01AG057836 (SMV), and R03AG061582 (SMV)), and the Alzheimer's Association (AARF-18-560786 (SMV)). Dr. Inouye holds the Milton and Shirley F. Levy Family Chair.

**Data Availability** Analysis code is available upon request.

## Compliance with Ethical Standards:

The funding sources had no role in the design, conduct, or reporting of this study.

**Conflict of Interest:** The authors declare that they do not have a conflict of interest.

**Corresponding Author:** Richard N. Jones, ScD; Department of Psychiatry and Human Behavior and Neurology, Brown University Warren Alpert Medical School, Providence, RI, USA (e-mail: Richard\_Jones@Brown.edu).

## REFERENCES

- Marcantonio ER. Delirium in Hospitalized Older Adults. *N Engl J Med*. 2017;377(15):1456-66. <https://doi.org/10.1056/NEJMc1605501>
- van den Boogaard M, Schoonhoven L, van der Hoeven JG, van Achterberg T, Pickkers P. Incidence and short-term consequences of delirium in critically ill patients: A prospective observational cohort study. *Int J Nurs Stud*. 2012;49(7):775-83. <https://doi.org/10.1016/j.ijnurstu.2011.11.016>
- Gleason LJ, Schmitt EM, Kosar CM, Tabloski P, Saczynski JS, Robinson T, et al. Effect of Delirium and Other Major Complications on Outcomes After Elective Surgery in Older Adults. *JAMA Surg*. 2015;150(12):1134-40. <https://doi.org/10.1001/jamasurg.2015.2606>
- Racine AM, Fong TG, Gou Y, Trivison TG, Tommet D, Erickson K, et al. Clinical outcomes in older surgical patients with mild cognitive impairment. *Alzheimers Dement*. 2017; doi:<https://doi.org/10.1016/j.jalz.2017.10.010>
- Hsieh TT, Saczynski J, Gou RY, Marcantonio E, Jones RN, Schmitt E, et al. Trajectory of Functional Recovery After Postoperative Delirium in Elective Surgery. *Ann Surg*. 2017;265(4):647-53. <https://doi.org/10.1097/SLA.0000000000001952>
- Saczynski JS, Marcantonio ER, Quach L, Fong TG, Gross A, Inouye SK, et al. Cognitive trajectories after postoperative delirium. *N Engl J Med*. 2012;367(1):30-9. <https://doi.org/10.1056/NEJMoa1112923>
- Inouye SK, Marcantonio ER, Kosar CM, Tommet D, Schmitt EM, Trivison TG, et al. The short-term and long-term relationship between delirium and cognitive trajectory in older surgical patients. *Alzheimers Dement*. 2016;12(7):766-75. <https://doi.org/10.1016/j.jalz.2016.03.005>
- Davis DH, Muniz-Terrera G, Keage HA, Stephan BC, Fleming J, Ince PG, et al. Association of Delirium With Cognitive Decline in Late Life: A Neuropathologic Study of 3 Population-Based Cohort Studies. *JAMA Psychiatry*. 2017;74(3):244-51. <https://doi.org/10.1001/jamapsychiatry.2016.3423>
- Vasunilashorn SM, Fong TG, Albuquerque A, Marcantonio ER, Schmitt EM, Tommet D, et al. Delirium Severity Post-Surgery and its Relationship with Long-Term Cognitive Decline in a Cohort of Patients without Dementia. *J Alzheimers Dis*. 2018;61(1):347-58. <https://doi.org/10.3233/JAD-170288>
- Davis DH, Muniz-Terrera G, Keage H, Rahkonen T, Oinas M, Matthews FE, et al. Delirium is a strong risk factor for dementia in the oldest-old: a population-based cohort study. *Brain*. 2012;135(Pt 9):2809-16. <https://doi.org/10.1093/brain/awt190>
- Fong TG, Davis D, Growdon ME, Albuquerque A, Inouye SK. The interface between delirium and dementia in elderly adults. *Lancet Neurol*. 2015;14(8):823-32. [https://doi.org/10.1016/S1474-4422\(15\)00101-5](https://doi.org/10.1016/S1474-4422(15)00101-5)
- van den Boogaard M, Pickkers P, Slooter AJ, Kuiper MA, Spronk PE, van der Voort PH, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICU patients) delirium prediction model for intensive care patients: observational multicentre study. *BMJ*. 2012;344:e420. <https://doi.org/10.1136/bmj.e420>
- de Wit HA, Winkens B, Mestres Gonzalvo C, Hurkens KP, Mulder WJ, Janknegt R, et al. The development of an automated ward independent delirium risk prediction model. *Int J Clin Pharm*. 2016;38(4):915-23. <https://doi.org/10.1007/s10996-016-0312-7>
- Rudolph JL, Jones RN, Levkoff SE, Rockett C, Inouye SK, Sellke FW, et al. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. *Circulation*. 2009;119(2):229-36. <https://doi.org/10.1161/CIRCULATIONAHA.108.795260>
- Marcantonio ER, Goldman L, Mangione CM, Ludwig LE, Muraca B, Haslauer CM, et al. A clinical prediction rule for delirium after elective noncardiac surgery. *Jama*. 1994;271(2):134-9.
- Inouye SK, Viscoli CM, Horwitz RI, Hurst LD, Tinetti ME. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. *Ann Intern Med*. 1993;119(6):474-81.
- Lindroth H, Bratzke L, Purvis S, Brown R, Coburn M, Mrkobrada M, et al. Systematic review of prediction models for delirium in the older adult inpatient. *BMJ Open*. 2018;8(4):e019223. <https://doi.org/10.1136/bmjopen-2017-019223>
- Halladay CW, Sillner A, Rudolph JL. Performance of electronic prediction rules for prevalent delirium at hospital admission. *JAMA Netw Open*. 2018;1(4):e181405. <https://doi.org/10.1001/jamanetworkopen.2018.1405>
- Wong A, Young AT, 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 Netw Open*. 2018;1(4):e181018. <https://doi.org/10.1001/jamanetworkopen.2018.1018>
- Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of Incident Delirium Using a Random Forest classifier. *J Med Syst*. 2018;42(12):261. <https://doi.org/10.1007/s10916-018-1109-0>
- Davoudi A, Ebadi A, Rashidi P, Ozzazgat-Baslanti T, Bihorac A, Bursian AC. Delirium Prediction using Machine Learning Models on Preoperative Electronic Health Records Data. *Proc IEEE Int Symp Bioinformatics Bioeng*. 2017;2017:568-73. <https://doi.org/10.1109/BIBE.2017.00014>
- Kramer D, Veeranki S, Hayn D, Quehenberger F, Leodolter W, Jagsch C, et al. Development and Validation of a Multivariable Prediction Model for the Occurrence of Delirium in Hospitalized Gerontopsychiatry and Internal Medicine Patients. *Stud Health Technol Inform*. 2017;236:32-9.
- Oh J, Cho D, Park J, Na SH, Kim J, Heo J, et al. Prediction and early detection of delirium in the intensive care unit by using heart rate



- variability and machine learning. *Physiol Meas*. 2018;39(3):035004. <https://doi.org/10.1088/1361-6579/aaab07>
24. **Obermeyer Z, Emanuel EJ.** Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-9. <https://doi.org/10.1056/NEJMp1606181>
  25. **Odden MC, Melzer D.** Machine Learning in Aging Research. *J Gerontol A*. 2019;74(12):1901-2. <https://doi.org/10.1093/gerona/gz074>
  26. **Schmitt EM, Saczynski JS, Kosar CM, Jones RN, Alsop DC, Fong TG, et al.** The Successful Aging After Elective Surgery Study: Cohort Description and Data Quality Procedures. *J Am Geriatr Soc*. 2015;63(12):2463-71. <https://doi.org/10.1111/jgs.13793>
  27. **Schmitt EM, Marcantonio ER, Alsop DC, Jones RN, Rogers SO, Jr, Fong TG, et al.** Novel risk markers and long-term outcomes of delirium: the successful aging after elective surgery (SAGES) study design and methods. *J Am Med Dir Assoc*. 2012;13(9):818 e1-10. <https://doi.org/10.1016/j.jamda.2012.08.004>
  28. **Saczynski JS, Kosar CM, Xu G, Puella MR, Schmitt E, Jones RN, et al.** A tale of two methods: chart and interview methods for identifying delirium. *J Am Geriatr Soc*. 2014;62(3):518-24. <https://doi.org/10.1111/jgs.12684>
  29. **Devore EE, Fong TG, Marcantonio ER, Schmitt EM, Trivison TG, Jones RN, et al.** Prediction of Long-term Cognitive Decline Following Postoperative Delirium in Older Adults. *J Gerontol A Biol Sci Med Sci*. 2017; doi:<https://doi.org/10.1093/gerona/glx030>
  30. **Rice ME, Harris GT.** Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law Hum Behav*. 2005;29(5):615-20. <https://doi.org/10.1007/s10979-005-6832-7>
  31. **Jones RN, Marcantonio ER, Saczynski JS, Tommet D, Gross AL, Trivison TG, et al.** Preoperative Cognitive Performance Dominates Risk for Delirium Among Older Adults. *J Geriatr Psychiatry Neurol*. 2016; doi:<https://doi.org/10.1177/0891988716666380>
  32. **Racine AM, Fong TG, Gou Y, Trivison TG, Tommet D, Erickson K, et al.** Clinical outcomes in older surgical patients with mild cognitive impairment. *Alzheimers Dement*. 2018;14(5):590-600. <https://doi.org/10.1016/j.jalz.2017.10.010>
  33. **van Meenen LC, van Meenen DM, de Rooij SE, ter Riet G.** Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *J Am Geriatr Soc*. 2014;62(12):2383-90. <https://doi.org/10.1111/jgs.13138>
  34. **Simon SE, Bergmann MA, Jones RN, Murphy KM, Orav EJ, Marcantonio ER.** Reliability of a structured assessment for nonclinicians to detect delirium among new admissions to postacute care. *JAMDA*. 2006;7(7):412-5. <https://doi.org/10.1016/j.jamda.2006.02.006>
  35. **Albert MS, Levkoff SE, Reilly C, Liptzin B, Pilgrim D, Cleary PD, et al.** The Delirium Symptom Interview: An interview for the detection of delirium symptoms in hospitalized patients. *J Geriatr Psychiatry Neurol*. 1992;5(1):14-21.
  36. **Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegel AP, Horwitz RI.** Clarifying confusion: The Confusion Assessment Method. A new method for detection of delirium. *Ann Intern Med*. 1990;113(12):941-8.
  37. **Wei LA, Fearing MA, Sternberg EJ, Inouye SK.** The Confusion Assessment Method: a systematic review of current usage. *J Am Geriatr Soc*. 2008;56(5):823-30.
  38. **Wong CL, Holroyd-Leduc J, Simel DL, Straus SE.** Does This Patient Have Delirium? *JAMA*. 2010;304(7):779-86.
  39. **Inouye SK, Leo-Summers L, Zhang Y, Bogardus ST, Jr, Leslie DL, Agostini JV.** A chart-based method for identification of delirium: validation compared with interviewer ratings using the confusion assessment method. *J Am Geriatr Soc*. 2005;53(2):312-8.
  40. **Teng EL, Chui HC.** The Modified Mini-Mental State (3MS) examination. *J Clin Psychol*. 1987;48(8):314-8.
  41. **James G, Witten D, Hastie T, Tibshirani R.** An introduction to statistical learning. New York: Springer; 2013.
  42. **Friedman J, Hastie T, Tibshirani R.** Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.
  43. **Zou J, Han Y, So S-S.** Overview of artificial neural networks. *Methods Mol Biol*. 2008;458:15-23. [https://doi.org/10.1007/978-1-60327-101-1\\_2](https://doi.org/10.1007/978-1-60327-101-1_2)
  44. **Zou H, Hastie T.** Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301-20.
  45. **Bannick MS, McGaughey M, Flaxman AD.** Ensemble modelling in descriptive epidemiology: burden of disease estimation. *Int J Epidemiol*. 2019; doi:<https://doi.org/10.1093/ije/dyz223>
  46. **James G, Witten D, Hastie T, Tibshirani R.** An introduction to statistical learning. Springer; 2013.
  47. **Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al.** caret: Classification and Regression Training (R package version 6.0-86). 2020. <https://CRAN.R-project.org/package=caret>
  48. **Venables W, Ripley B.** Modern Applied Statistics with S. New York: Springer; 2002.
  49. **Mulborrow S.** Earth: Multivariate Adaptive Regression Splines (R package version 5.2.0). 2020. <https://CRAN.Rproject.org/package=earth>
  50. **Liaw A, Winer M.** Classification and regression by randomForest. *R News*. 2002;2(3):18-22.
  51. **Karatzoglou A, Smola A, Hornik K, Zeileis A.** kernlab-an S4 package for kernel methods in R. *J Stat Softw*. 2004;11(9):1-20.
  52. **Greenwell B, Boehmke B, Cunningham J.** GBM Developers. gbm: Generalized Boosted Regression Models (R package version 2.1.8). 2020. <https://CRAN.R-project.org/package=gbm>
  53. **Cohen J.** Statistical power analysis for the behavioral sciences. New York: Academic Press; 1969.
  54. **Christodoulou E, Jie M, Collins GS, Steyerberg EW, Verbakel JY, van Calster B.** A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.