Supplement Table 2. Performance measures of included studies. *95%CI inserted when reported

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Jauk 2020	Random Forest	Random Forest	Longitudinal EHR data .	Expert opinion	74.1	82.2	0.85	5.8	99.5	Low concordance on calibration plot
Jauk 2022	Random Forest	Random Forest	Longitudinal EHR data	Senior physician delirium risk estimation	100.0	90.6	0.92 (0.9155 - 0.9392)	NR	NR	Low concordance on calibration plot
Sun 2022	Natural Language processing	Natural Language Processing	Retrospective structured EHR data	Compared prospective performance with retrospective and cross-hospital evaluations	NR	NR	0.94	NR	NR	Used a calibration tool for cross-hospital evaluation but metric not clearly stated.
Kramer 2017	Random Forest; Artificial Neural Network; Support Vector Machine; Logistic Regression; k nearest neighbour; Other: Linear Discriminant Analysis, Elastic Net	Random Forest	Retrospective HER data	NR	69.0	90.0	0.91	NR	NR	NR

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Bishara 2022	Gradient Boosting (e.g., GBM, XGBoost); Artificial Neural Network	Gradient Boosting	Pragmatically collected Retrospective EHR data	Regression model using clinician-selected variables and a delirium risk stratification tool (AWOL-S)	80.6 (77.1%-84.1%)	73.7 (72.4%-74.9%)	0.852 (0.84 - 0.86)	14.4 (13.5%-15.3%)	98.6 (98.3%-98.8%)	High concordance on calibration plot
Cano-Escalera 2022	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); k nearest neighbour	Random Forest	Sociodemo-grahic data, personal antecedent, and clinical data extracted from EHRs, in addition to information from functional status and frailty tests.	NR	NR	NR	0.745	NR	NR	NR
Coombes 2021	Random Forest; Decision Tree; Support Vector Machine; Logistic Regression; Naive Bayes	Logistic Regression	Medical Information Mart for Intensive Care-III database (MIMIC-III)	Compared best model with two models previously proposed in the literature for goodness of fit, precision, and biological validation	79.4	71.5	0.83	19.7	97.6	NR
Corradi 2018	Random Forest	Random Forest	Retrospective EHR data	NR	69.8	92.7	0.91 (0.90 - 0.92	NR	NR	Low concordance on calibration plot

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Davoudi 2017	Random Forest; Support Vector Machine; Linear Regression; Other: Generalised additive models	Generalised Additive Model	Preoperative EHR data collected on admission	NR	75.0 (0.69 - 0.83)	81.0 (0.76 - 0.83)	0.86 (0.84 - 0.88)	NR	NR	NR
Gutheil 2022	Gradient Boosting (e.g., GBM, XGBoost); Artificial Neural Network	Self-Attention and Intersample Attention Transformer	Data from a benchmarking and reporting system	NR	NR	NR	0.82 (0.76 - 0.87)	NR	NR	NR
Halladay 2018	Random Forest	Random Forest	Data from the Veteran Affairs (VA) External Peer Review Program (EPRP).	Compared random forest generated consolidated NICE rule with previously confirmed scoring algorithms (electronic NICE and Pendlebury NICE)	NR	NR	0.91 (0.91 - 0.92)	47.0	96.0	NR
Hu 2022	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Support Vector Machine; Logistic Regression	Logistic Regression	Clinical data from observational study	NR	89.1 (79.1 - 95.0)	44.2 (31.1 - 58.3)	0.80 (0.72 - 0.89)	NR	NR	High concordance on calibration plot and satisfactory Brier score (0.151)

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Jauk 2018	Random Forest; Artificial Neural Network; Linear Regression	Artificial Neural Network	Retrospective EHR data	Compared prevalence of diabetes in patients with and without delirium, standardised by age	~82.5 (~80.0-85.0)	~73.0 (~71.0 – 75.0)	0.46 (0.836 - 0.857)	NR	NR	NR
Ji 2018	Artificial Neural Network	Artificial Neural Network	Secondary data from an observational study, including sociodemographic, clinical, laboratory, and pharmacological information	NR	NR	NR	0.89	NR	NR	NR
Kurisu 2022	Decision Tree	Decision Tree	Secondary data from a multicenter prospective observational study	NR	60.5	82.2	0.72 (0.63 - 0.81)	NR	NR	NR

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Li 2022	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Decision Tree; Logistic Regression	Decision Tree	Prospective cohort dataset with basic information, clinical signs and symptoms, laboratory findings, and scale variables.	NR	93.3	94.3	0.95	71.8	98.9	NR
Lucini 2020	Random Forest; Artificial Neural Network; Support Vector Machine; Logistic Regression; AdaBoost	Logistic Regression	Data obtained from a data repository specific to ICUs across Alberta, Canada	NR	84.0 (83.9 – 84.1)	86.5 (86.5 – 86.5)	0.85 (0.82 - 0.88)	NR	NR	NR
Menzenbach 2022	Gradient Boosting (e.g., GBM, XGBoost)	Gradient Boosting	Preoperative dataset with structured data and test results entered into REDCap electronic database.	Predictor selection by experts (Investigators vs external DELirium Prediction based on Hospital Information (DELPHI) score)	NR	NR	0.77 (0.65 - 0.85)	NR	NR	Satisfactory Brier score (0.142)

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Mufti 2019	Random Forest; Decision Tree; Artificial Neural Network; Support Vector Machine; Logistic Regression; Naive Bayes; Other: Bayesian belief networks	Artificial Neural Network	Prospective registry of all cardiac surgical cases.	Compared feature importance of input variables in the random forest model to univariate logistic regression analysis.	67.7	72.9	0.78 (0.72 - 0.84)	24.3	94.6	NR
Netzer 2020	Random Forest; Decision Tree; Support Vector Machine; Logistic Regression; Naive Bayes; k nearest neighbour; Other: Multilayer perceptron	Random Forest	Retrospective EHR data	Predictive ability (Kappa) of DOSS and CAM delirium screening methods.	NR	NR	0.87	NR	NR	NR

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Oh 2018	Support Vector Machine; Other: SVM with radial basis function (RBF) kernels, linear extreme learning machine (ELM), ELM with RBF kernels, linear discriminant analysis & quadratic discriminant analysis	SVM with RBF Kernels	Heart rate variability data obtained from electrocardiograms	NR	91.3	90.8	NR	91.6	90.55	NR
Oosterhoff 2021	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Artificial Neural Network; Support Vector Machine; Logistic Regression	Logistic Regression	Data from NSQIP database	Default strategies of decision change and preoperative delirium presence	NR	NR	0.79 (0.77 - 0.80)	NR	NR	High concordance on calibration plot and satisfactory Brier score (0.15)

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Racine 2021	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Artificial Neural Network; Logistic Regression	Artificial Neural Network	Secondary analysis of a prospective study (Successful Aging after Elective Surgery study)	Standard approach for delirium prediction, including backwards stepwise logistic regression and a previously published delirium risk prediction rule for hospitalised patients	50.0	82.0	0.71 (0.58 - 0.83)	46.0	84.0	Low concordance on calibration plot
Son 2022	Random Forest; Artificial Neural Network; Support Vector Machine; Logistic Regression; Other: Four rule-mining algorithms (C4.5, CBA, MCAR & LEM2)	LEM2 Rule Mining	Retrospective EHR data	NR	96.7	NR	NR	96.74	NR	NR
Sun 2021	Natural Language Processing	Natural Language Processing	Retrospective EHR data	Compared performance of delirium prediction model trained in the training site with the model generated with the calibration process	NR	NR	0.82	NR	NR	Used a 'calibration tool' in model development, but no calibration metrics reported.

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Veeranki 2019	Random Forest; Logistic Regression; Other: Random forest in combination with logistic regression	Combined logistic Regression and Random forest	Retrospective demographic data, diagnoses, procedures, laboratory results, nursing assessments obtained from hospital information system	NR	NR	NR	0.91	NR	NR	NR
Veeranki 2018	Random Forest	Random Forest	Data obtained from hospital information systems	NR	NR	NR	0.90	NR	NR	NR
Wang 2020	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Decision Tree; Logistic Regression	Random Forest	Retrospective EHR data	NR	0.712	NR	0.96	0.887	NR	NR

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Wong 2018	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Artificial Neural Network; Support Vector Machine; Logistic Regression	Gradient Boosting	Retrospective EHR data	Compared against the 4-pont scoring system AWOL (age >79, failure to spell world backwards, disorientation to place, and higher nurse-rated illness severity)	59.7 (52.4-66.7)	90.0 (89.0-90.9)	0.85	23.1 (20.5-25.9)	97.8 (97.4-98.1)	NR
Xue 2021	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Artificial Neural Network; Support Vector Machine; Logistic Regression	Gradient Boosting	Retrospective data from electronic anasthesia and preoperative assessment records	NR	31.1 (30.3-32.0)	95.0 (94.9-95.0)	0.76 (0.76 - 0.76)	NR	NR	NR
Xue 2022	Multilayer perceptron	Multilayer Perceptron	Retrospective EHR data	NR	NR	NR	0.90	NR	NR	NR

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
H Zhao 2021	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Support Vector Machine; Logistic Regression; Other: Multilayer Perceptron	Logistic Regression	Data obtained from Anesthesia Information Management System (AIMS)	NR	NR	NR	0.78 (0.70, 0.86)	NR	NR	NR
Y Zhao 2021	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Logistic Regression	Logistic Regression	MIMIC-IV open-access dataset	NR	NR	NR	0.69	NR	NR	NR
Amador 2022	Gradient Boosting (e.g., GBM, XGBoost)	Gradient Boosting	Retrospectively collected data obtained from patient administrative data, laboratory results and vital signs available within the first hour after ICU admission.	Compared lightGBM regularised algorithm with XGBoost, Random Forest, and Logistic Regression in terms of gains for robustness and stability of explanations.	NR	NR	0.86	NR	NR	NR

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Castro 2022	L1-penalized regression.	L1 Penalised Regression	Retrospective EHR data, including diagnostic, medication, laboratory, and other clinical features available at time of hospital admission	Compared model calibration with initial COVID-9 surge of previous paper	62 (58 – 65)	75 (73 - 76)	0.75 (0.73 - 0.77)	28.0 (91-93)	92.0 (26-31)	Low concordance on calibration tests (evaluated using both quantile-by-quintile comparison and Spiegelhalter's Z test)
Castro 2021	L1-penalised regression	L1 Penalised Regression	Retrospective EHR data	NR	73 (65 – 80)	69 (65-73)	0.75 (0.70 - 0.79)	35 (30-41)	92 (89-94)	High concordance (evaluated using both quantile-by-quantile comparison and Hosmer-Lemeshow test)

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Hur 2021	Random Forest; Gradient Boosting (e.g., GBM, XGBoost); Artificial Neural Network; Logistic Regression	Random Forest	Data from the Clinical Data Warehouse Darwin-C database of the Samsung Medical Centre and the Medical Information Mart for Intensive Care III (MIMIC-III) database (v1.4). The Samsung Medical Centre data set was used for the derivation and validation cohort, and the MIMIC-III data set was used for the external validation cohort.	NR	91.0 (90.4 – 90.5)	27.0 (26.6– 27.3)	0.72 (0.720 - 0.721)	15.9 (15.9- 16.0)	95.2 (95.1- 95.3)	Satisfactory Brier Score (0.168)

Authors	Model(s)	Best-performing Model (by AU-ROC)	Training data source	Comparator	Sensitivity (%)	Specificity (%)	AU-ROC*	PPV (%)	NPV (%)	Calibration
Oosterhoff 2022	Elastic-net Penalised Logistic Regression	Elastic-net Penalised Regression	Postoperative EHR data from the NSQIP targeted files for hip fractures	NR	NR	NR	0.74 (0.73 - 0.76)	NR	NR	High concordance on calibration plot and satisfactory Brier score (0.22)
Jauk 2019	Random Forest	Random forest	Data obtained from hospital information system implemented on German software platforms	Compared model specifically trained on data with NR values to a currently-implemented delirium prediction model	NR	NR	0.83 (0.818 - 0.841)	NR	NR	NR

Abbreviations: LR, Logistic Regression; RF, Random Forest; SAINTENS, Self-Attention and Intersample Attention Transformer; SVM, Support Vector Machine; RBF, Radial basis function; AU-ROC, Area under the receiving operating curve; PPV, Positive predictive value; NPV, Negative predictive value; AWOL-S, Age, WORLD backwards, Orientation, Illness severity, Surgery-specific risk; DOSS, Delirium Observation Screening Scale. NR=not reported; NSQIP=National Surgical Quality Improvement Program