Delirium screening in an acute care setting with a machine learning classifier based on routinely collected nursing data: A model development study

Spiller, Tobias R ; Tufan, Ege ; Petry, Heidi ; Böttger, Sönke ; Fuchs, Simon ; Duek, Or ; Ben-Zion, Ziv ; Korem, Nachshon ; Harpaz-Rotem, Ilan ; von Känel, Roland ; Ernst, Jutta

DOI: https://doi.org/10.1016/j.jpsychires.2022.10.018

Journal of Psychiatric Research 156 (2022) 194–199

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/jpsychires

Delirium screening in an acute care setting with a machine learning classifier based on routinely collected nursing data: A model development study

Tobias R. Spiller [a,b,c,d,*], Ege Tufan [e], Heidi Petry [b,f], Sönke Böttger [b,g], Simon Fuchs [a,b,h], Or Duek [c,d], Ziv Ben-Zion [c,d], Nachshon Korem [c,d], Ilan Harpaz-Rotem [c,d,i,j], Roland von Känel [a,b], Jutta Ernst [b,f]

[a] Department of Consultation-Liaison Psychiatry and Psychosomatic Medicine, University Hospital Zurich (USZ), Zurich, Switzerland
[b] University of Zurich (UZH), Zurich, Switzerland
[c] Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA
[d] National Center for PTSD, Clinical Neurosciences Division, VA Connecticut Healthcare System, West Haven, CT, USA
[e] German Institute for Literature, Leipzig University, Leipzig, Germany
[f] Center of Clinical Nursing Science, University Hospital Zurich (USZ), Zurich, Switzerland
[g] Department of Gastroenterology, University Hospital Zurich (USZ), Zurich, Switzerland
[h] Psychiatric University Hospital Zurich (PUK), Zurich, Switzerland
[i] Northeast Program Evaluation Center, VA Connecticut Healthcare System, West Haven, USA
[j] Department of Psychology, Yale University, New Haven, CT, USA

ARTICLE INFO

Keywords:
Delirium
Machine learning
Prediction model
Screening

ABSTRACT

Delirium screening in acute care settings is a resource intensive process with frequent deviations from screening protocols. A predictive model relying only on daily collected nursing data for delirium screening could expand the populations covered by such screening programs. Here, we present the results of the development and validation of a series of machine-learning based delirium prediction models. For this purpose, we used data of all patients 18 years or older which were hospitalized for more than a day between January 1, 2014, and December 31, 2018, at a single tertiary teaching hospital in Zurich, Switzerland. A total of 48,840 patients met inclusion criteria. 18,873 (38.6%) were excluded due to missing data. Mean age (SD) of the included 29,967 patients was 71.1 (12.2) years and 12,231 (40.8%) were women. Delirium was assessed with the Delirium Observation Scale (DOS) with a total score of 3 or greater indicating that a patient is at risk for delirium. Additional measures included structured data collected for nursing process planning and demographic characteristics. The performance of the machine learning models was assessed using the area under the receiver operating characteristic curve (AUC). The training set consisted of 21,147 patients (mean age 71.1 (12.1) years; 8,630 (40.8%) women|) including 233,024 observations with 16,167 (6.9%) positive DOS screens. The test set comprised 8,820 patients (median age 71.1 (12.4) years; 3,601 (40.8%) women) with 91,026 observations with 5,445 (6.0%) positive DOS screens. Overall, the gradient boosting machine model performed best with an AUC of 0.933 (95% CI, 0.929 - 0.936). In conclusion, machine learning models based only on structured nursing data can reliably predict patients at risk for delirium in an acute care setting. Prediction models, using existing data collection processes, could reduce the resources required for delirium screening procedures in clinical practice.

1. Introduction

Delirium is a severe neuropsychiatric syndrome characterized by fluctuations in alertness and cognition with an acute onset. Up to one third of hospitalized patients develop delirium during their hospital stay (Inouye et al., 2014; Schubert et al., 2018). Although delirium has been

* Corresponding author. Department of Consultation-Liaison Psychiatry and Psychosomatic Medicine, University Hospital Zurich, Haldenbachstrasse 16/18, CH-8091, Zurich, Switzerland.
E-mail address: tobiasraphael.spiller@uzh.ch (T.R. Spiller).

https://doi.org/10.1016/j.jpsychires.2022.10.018
Received 2 June 2022; Received in revised form 20 September 2022; Accepted 3 October 2022
Available online 10 October 2022

under-diagnosed for a long time, its association with a broad range of negative outcomes (e.g., prolonged hospitalization and increased mortality) has been well documented (Siddiqi et al., 2006). To minimize the negative impact of delirium on these and additional outcomes, its early detection and non-pharmacological treatment are of crucial importance (Kim et al., 2020; Oh et al., 2017). Thus, daily screening of patients at risk for delirium are recommended by current guidelines (Young et al., 2010) and have been established in many acute care facilities (de Wit et al., 2016; Rudolph et al., 2016). Because most screening procedures need to be carried out by qualified health care providers (i.e., nursing staff or clinicians) the costs of such procedures are relatively high and thus, screening is often restricted to high-risk populations (De and Wand, 2015). Despite these efforts, delirium remains under-diagnosed and under-treated (McCoy et al., 2016).

To optimize existing delirium screening processes, numerous delirium prediction models have been developed over the last two decades (Chua et al., 2021; Lindroth et al., 2018). Early models were designed as screening tools in form of a questionnaire. Therefore, such models had to balance their precision with the resources required to obtain the necessary information (e.g., clinical assessments). Other predictive models focused on risk stratification aiming to predict the severity of delirium based on a few easy-to-obtain items (Inouye et al., 1993). More recently, the wide use of electronic health records (EHR) and the advancements in predictive modeling using machine learning (ML) algorithms paved the way for a new generation of delirium prediction models (Lindroth et al., 2018). These novel models make use of data collected routinely and stored in the EHR to optimize the precision of their prediction. Consequently, such models include several dozens to hundreds of different predictor variables. For example, the authors of one study used EHR data available in the first 24 h after patient admission to an acute care facility to predict the onset of delirium using a series of ML models. Their models achieved excellent performance (Wong et al., 2018). Yet, each of their ML models included almost 800 different clinical variables. Notably, other predictive models with a similar performance also incorporate hundreds of variables (Chua et al., 2021; Jauk et al., 2020).

While many of these ML-based delirium prediction models typically achieve reliable performance, their clinical applicability is limited in two fundamental ways. First, although these models usually rely on routinely collected data, many of the included variables will not be collected for every patient during every hospitalization. For example, in one study, laboratory results of a comprehensive metabolic panel (e.g., including ammonia) were among the variables with the highest predictive value (Wong et al., 2018). However, such a panel is likely only collected from a minority of patients. Consequently, the utility of a model incorporating such variables is limited in clinical practice. Second, the type, structure, and quality of the data obtained and stored in EHRs vary across different hospital networks (and sometimes even within the same facility). Since most of the developed ML-based delirium prediction models try to use the full source of data, these models are affected by the standard operating procedures of the study site. Consequently, many of these models are highly generic and it is doubtful that they can be used in a clinical setting other than the one in which they were trained – at least not without extensive site-specific model retraining and reevaluation (Jauk et al., 2020; Sun et al., 2021).

Hence, the clinical applicability of a delirium prediction model across different sites requires standardized patient data collection at each hospitalization regardless of its location if substantial retraining and validation at every new site should be avoided. Nowadays, many hospitals use the same standardized tools to plan and document the care processes for each patient. Data obtained by these tools are therefore similar across different sites. Consequently, a delirium prediction model based on such data might therefore be more generalizable to multiple sites. The aim of the current study was to provide a proof-of-principle that a delirium prediction model based on such standardized nursing data can achieve good performance. Therefore, we aimed to develop a ML-based delirium prediction model using data collected with standardized nursing instruments at a single study site.

2. Materials and methods

2.1. Procedure

This study included data extracted from the EHR of all patients who were hospitalized at the University Hospital Zurich in Switzerland for more than 24 h between January 1, 2014, and December 31, 2018. Data collected during the calendar year of 2014 (January 1 to December 31) were part of a research program approved by the ethics committee of the Canton of Zurich (see (Schubert et al., 2018)). The investigation was carried out in accordance with the Declaration of Helsinki. Data collected between January 1, 2015, to December 31, 2018, were obtained for quality control reasons. Under Swiss federal law and local regulations, such data is waived from the requirement of written informed consent by individual participants.

2.2. Measures

Demographic information was extracted from the EHR of individual patients, including patients' age and sex. Information on self-reported race and ethnicity is not routinely collected in Switzerland and was therefore not available.

Delirium presence or absence was determined using the Delirium Observation Scale (DOS) (Koster et al., 2009). The DOS is a 13-item scale developed to assess the presence of delirium as defined by the fourth edition of the Diagnostic and Statistical Manual of Mental Disorder (DSM-IV) and administered by nursing staff (Scheffer et al., 2011). Items reflect disturbances in the patient's level of consciousness, attention, thought processes, orientation, memory, psychomotor behavior, and affect (e.g., a sample item "The patient is picking, disorderly, restless"). All items are rated with a dichotomous response format, three items are reversely coded. Individual item ratings are summed to a total score ranging from 0 to 13, with higher scores indicating higher probability of delirium. A total score of three or more is considered to be indicative for delirium (Gemert van and Schuurmans, 2007). The DOS has been previously validated in a variety of populations (Gavinski et al., 2016; Koster et al., 2009). In this study Cronbach's alpha was 0.8 indicating good internal consistency.

Standardized documentation of nursing processes was carried out with the Electronic Patient Assessment-Acute Care® (ePA-AC®). The ePA-AC® is a 56-item nursing instrument developed to measure abilities and impairments of patients in 11 domains (e.g., state of consciousness or motor skills). Individual items are rated dichotomously or on four-point Likert scales and can be used to calculate a variety of scores (for more information see (Hunstein, 2012)). In the current analysis, only the raw scores of the 56 items were included.

2.3. Study population

During the study period all patients of 65 years or older, as well as patients younger than 65 years old with delirious symptoms or a known neurological disorder or cognitive impairments, were included in the delirium screening program. Delirium screening with the DOS was conducted 3 times a day (observations) until screening was negative for three consecutive days, the patient died or was discharged. Standardized nursing documentation was completed for every patient once every 24 h at the end of the morning shift (at 4 p.m.). The results of each patient's delirium screening events were included in the study sample (i.e., observation) when the standardized nursing documentation was completed for the same assessment day. Consequently, up to three screening events per patient per day could be included, with all three referring to the same nursing documentation. We conducted a complete-case analysis (i.e., if one of the 13 DOS or 56 ePA-AC® items was

missing, the whole scale was missing for this observation).

Overall, 48,840 patients with cumulatively 536,233 screening events between January 1, 2015, and December 31, 2018, were considered for this study. Of these, 18,846 (38.7%) patients and 212,183 (39.6%) screening events were excluded due to missing data. The remaining 324,050 outcome observations were split into: (1) **Training Set** comprising of 233,024 (71.9%) observations from 21,147 patients screened between January 1, 2015, and December 31, 2017; (2) **Test Set** comprising of 91,026 (28.1%) screening observations from 8,820 patients obtained between January 1, 2018, and December 31, 2018.

2.4. Model development & statistical analysis

The outcome of each model was coded as a binary variable with two levels, delirium presence defined as a DOS total score ≥3) and delirium absence (DOS total score <3). Model predictors included sex, age (as a continuous predictor) and all 56-items of the ePA-AC® in their original coding.

Model training and validation followed procedures were adapted from similar publications (Wong et al., 2018). The performance of the following models was tested (implemented in the following package): non-regularized logistic regression (LR; *glm*), penalized logistic regression (PLR; *glmnet*), random forest (RF; *randomForest*), linear support vector machine (SVM; *e1071*), gradient boosting machine (GBM; *gbm*), and an artificial neural network with a single hidden layer (Nnet; *nnet*). Because the outcome was highly imbalanced with less than 7% of outcome observations indicating delirium presence, we used down sampling to balance the training set. We then first optimized the hyperparameters of each model using three repeats of five-fold cross--validation. In a second step, the optimized models were trained using the full training data set. Third, model performance was assessed using the area under the receiver operating characteristic curve (AUC). In addition, sensitivity, specificity, positive predictive values (PPV), and negative predictive value (NPV) are reported. Exact binomial confidence intervals were calculated for sensitivity, specificity, PPV, and NPV. Confidence intervals for AUC were obtained using a bootstrapping procedure. Fourth, AUCs were compared using DeLong's test for correlated receiver operating characteristic curves (DeLong et al., 1988). Fifth, the relative importance of the included predictors was calculated for all models but the linear support vector machine.

In a sensitivity analysis, we assessed the performance of a series of models limited to ten predictors. We selected ten predictors based on their variable importance in the full models aiming to choose predictors with top variable importance in more than one full model (see Table S3). Training, testing, and evaluation of these models followed the same procedure as outlined above.

The α-level was set at .05 and all tests were conducted two-sided. Because the analysis was not aimed at formal hypothesis testing, the resulting *P* values were not adjusted for multiple testing. We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guideline (TRIPOD) (Collins et al., 2015). All analyses were carried out in the R statistical environment with R version 4.0.2 (R Core Team, 2020).

3. Results

A demographic description of the dataset including delirium prevalence is provided in Table 1. Overall, 40.8% of the patients were female and mean age was 71.1 years (SD = 12.2). In n = 21,612 (6.1%) of included observations, DOS total scores were ≥3. The relative share of observations with a DOS total score ≥3 differed between the training (6.9%) and test set (6.0%).

The performance of all included models is presented in Table 2 and visualized in Fig. 1. The GBM model had the highest AUC of 0.933 (95% CI, 0.929 - 0.936) significantly outperforming all other models. However, all models had comparable accuracy and performance metrics, apart from the neural net model (Nnet) which performed the worst. Positive predictive values ranged from 0.986 (95% CI, 0.865 - 0.870; Nnet) to.990 (95% CI, 0.989 - 0.991; GBM), and negative predictive values ranged from 0.278 (95% CI, 0.271 - 0.285; Nnet) to 0.326 (95% CI, 0.318 - 0.334; PLR).

The variables with the highest predictive value mainly included variables assessing quantitative and qualitative aspects of consciousness as well as basic bodily and motor functions. For example, orientation (person, place, time, and situation), attention. and a history of falls were among the ten most predictive variables in several of the assessed predictive models. The performance of the series of models restricted to a selection of ten predictors is also shown in Table 2. Overall, these models performed similarly to the full models except for the random forest model which had lower performance than the corresponding model with all predictors. Additional information including confusion matrices and the relative importance of the included variables are outlined in the supplementary materials (Table S4-Table S19).

4. Discussion

Our findings demonstrate that a machine learning classifier solely based on daily collected structured nursing data can be used to reliably identify patients at risk for delirium. Although the GBM model performed significantly better than the models based on other algorithms, the differences in performance between the models were minimal (except for the artificial neural net model). The models developed and validated in this study rely exclusively on data that were routinely collected in the same standardized manner in many hospitals. If replicated in an independent dataset, these models could be readily implemented across different hospitals to optimize delirium screening protocols.

All developed models (with an exception of the artificial neural net model) performed similarly well and achieved a better or similar performance as the best delirium prediction models currently available

Table 1
Demographics and delirium presence by data set.

Variable	Data set, No (%) of Patients		
	Overall (N = 29,967)	Training data (n = 21,147)	Testing data (n = 8820)
Age, mean (SD), y	71.1 (12.2)	71.1 (12.1)	71.1 (12.4)
Gender			
Woman	12,231 (40.8)	8630 (40.8)	3601 (40.8)
Man	17,736 (59.2)	12,517 (59.2)	5219 (59.2)
No (%) of Observations			
DOS total score ≥3	21,612 (6.1)	16,167 (6.9)	5445 (6.0)
DOS total score <3	324,050 (93.9)	216,857 (93.1)	85,581 (94.0)

Abbreviations: DOS, Delirium observation scale.

Table 2
Model performance metrics.

	Full models[a]					
	LR	PLR	RF	SVM	GBM	Nnet
Variable						
Sensitivity	.890 (.887 -. 891)	.890 (.887–.892)	.889 (.887–.891)	.888 (.887–.891)	.874 (.872–.876)	.868 (.865–.870)
Specificity	.838 (.828–.847)	.838 (.828–.848)	.843 (.833–.852)	.841 (.832–.852)	.863 (.854–.872)	.800 (.789–.810)
NPV	.989 (.988–.989)	.989 (.988–.989)	.989 (.988–.990)	.989 (.988–.990)	.990 (.989–.991)	.986 (.985–.986)
PPV	.324 (.316–.332)	.326 (.318–.334)	.326 (.318 -. 333)	.325 (.317–.333)	.304 (.297–.311)	.278 (.271–.285)
AUC	.931 (.928–.935)	.931 (.928–.935)	.928 (.924–.932)	.930 (.927–.934)	.933 (.929–.936)	.904 (.900–.909)
Restricted models[b]						
Sensitivity	.889 (.886–.891)	.889 (.886–.891)	.867 (.864–.868)	.882 (.880–.884)	.874 (.871–.876)	.874 (.872–.876)
Specificity	.844 (.834–.853)	.844 (.834–.853)	.814 (.804–.824)	.835 (.825–.845)	.863 (.853–.872)	.863 (.854–.872)
NPV	.989 (.988–.990)	.989 (.988–.990)	.987 (.986–.987)	.988 (.987–.989)	.990 (.989–.991)	.990 (.989 -. 991)
PPV	.325 (.317–.333)	.325 (.317–.333)	.279 (.272–.286)	.311 (.303–.318)	.303 (.295–.310)	.303 (.296–.311)
AUC	.931 (.927–.935)	.931 (.927–.935)	.909 (.905–.913)	.929 (.925–.933)	.932 (.928–.935)	.928 (.925–.932)

Abbreviations: AUC, Area under the receiver operating characteristic curve; GBM, Gradient boosting machine; LR, Logistic regression; Nnet, Artificial neural network with a single hidden layer; NPV, Negative predictive value; PLR, Penalized logistic regression; PPV, Positive predictive value; SVM, Support vector machine.

[a] Models including all available predictors (56 ePA-AC® items, age, gender).

[b] Models including ten predictors with the highest variable importance.

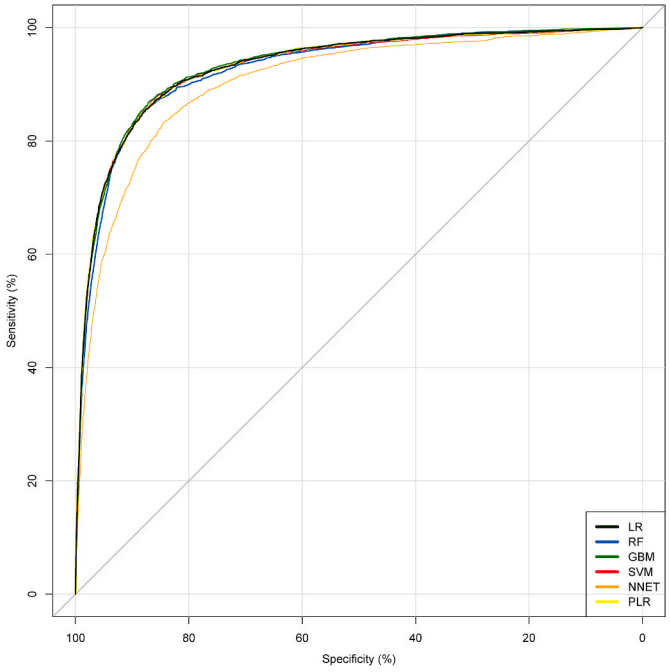

Fig. 1. Legend: GBM, Gradient boosting machine; LR, Logistic regression; Nnet, Artificial neural network with a single hidden layer; PLR, Penalized logistic regression; SVM, Support vector machine.

(Chua et al., 2021; Lindroth et al., 2018). For example, Wong et al. (2018) reported an AUC of 0.855 for their gradient boosting model, and Jauk and colleagues' (2020) random forest model achieved an AUC of 0.86. Notably, the models presented in this study relied on fewer predictors suggesting that nursing data can be used effectively in a parsimonious model for delirium prediction. There are multiple reasons why this could be the case. First, all nursing staff was trained in collecting structured and standardized data, which reduces the measurement error, increases inter-rater reliability, and results in higher data consistency (Weiskopf and Weng, 2013). Second, the staff collecting these data do rely on them for numerous clinical and documentation tasks. This minimizes the loss of data and increases the representativeness of the collected data for the study population. Third, some items included in the nursing assessment ask for a clinical judgment. Compared to results from a blood analysis or a patient's administrative arrival (e.g., from home or a nursing home (Zipser et al., 2022)), the clinical assessment of a patient's alertness is much more complex and might contain more relevant information. Fourth, the fact that different algorithms achieved similar performance with similar variables among those with top importance further emphasizes the high information value of these variables for delirium prediction. Taken together, this underscores that structured nursing data is a rich resource for delirium prediction models.

The variables with the highest predictive value included symptoms and known risk factors for delirium, emphasizing the clinical plausibility of the prediction models. For example, several items included in the nursing assessment cover the quantitative and qualitative state of consciousness, including changes in attention, a key diagnostic criterion for delirium (American Psychiatric Association, 2013). Similarly, other items with high predictive value include assessments of basic neuropsychiatric functions, such as interpersonal interaction or the need for support in daily activities, which are often disturbed in patients with delirium (Inouye et al., 2014). Individuals' age, a well-documented risk factor for delirium (Oh et al., 2017), was also among the ten variables with the highest predictive value in several of the developed models. Moreover, the relative importance of the five most important variables was greatly higher than the importance of the remaining variables. In accordance, the sensitivity analyses of a second series of delirium prediction models which were solely based on the ten predictors with highest predictive value across all assessed models revealed comparable performance like the full models.

Although the current findings highlight the clinical potential of a nursing data-based delirium prediction model, additional research is needed before such a model can be implemented in clinical practice. First, several studies demonstrated lower performance of predictive models at other clinical sites than the ones in which the models were developed (Zech et al., 2018). We would also expect a decrease in predictive precision of our models at another site, for example due to differences in the underlying hospital population. However, because our models are based on structured nursing data collected in the same standardized and structured manner at many different sites, our models should generalize well. Second, any implementation of a delirium prediction model in a clinical setting should be tested for its efficacy (i.e., whether it outperforms current screening protocols in a controlled study), as well as for its cost-effectiveness (Labarère et al., 2014; Salazar de Pablo et al., 2021) Still, the current findings provide a proof-of-principle for the use of clinical nursing data to reliably predict delirium and are therefore the first of many steps towards this goal.

4.1. Limitations

This study is subject to several limitations. First, the use of highly structured clinical data limits the generalizability and applicability of our models to sites using the same standardized nursing assessment (ePA-AC©). However, this specific nursing assessment is being used in hundreds of hospitals representing a population of millions of patients in which our prediction model could be employed. Nonetheless, the use of additional or alternative data sources could help to increase the number of sites in which such a delirium prediction model could be implemented. Furthermore, the use of additional data sources could also increase the performance and reliability of the prediction models. Second, the rate of outcome observations indicating delirium in our sample (6.1%) was lower than the prevalence of delirium documented in other large cohort studies (7–35%) (Boettger et al., 2020; Inouye et al., 2014; Siddiqi et al., 2006). Therefore, it is likely that some patients in our cohort were screened false negatively for delirium. Our predictive models are thus likely affected by this bias as well. Third, the generalizability of our findings is limited by using data from a single clinical site. As studies have shown that predictive models designed at a single site can have substantially worse performance at another clinical site (Beede et al., 2020), future research should aim to develop a predictive model that includes data from multiple sites. Fourth, the exclusion of patients with missing data (see methods) further limits the generalizability. However, we did not aim to develop a predictive model that is readily transferable into clinical use but designed this study to provide proof-of-principle for the utility of using structured nursing data. Given this proof, future work should focus on optimizing the model for maximal generalizability (e.g., also including data from different sites) and clinical utility. Fifth, our main outcome measure (i.e., presence of delirium based on total DOS scores $\geq$3) indicated that a patient is at-risk for delirium and does not correspond to the clinical diagnosis of delirium, which can only be made by structed interview with a certified clinician. Nevertheless, this DOS cut-off score has been validated as a delirium screener in numerous studies (Scheffer et al., 2011; Zipser et al., 2022) to determine presence or absence of delirium and therefore represents a widely used outcome in delirium research. Apart from these limitations, the current study has several important strengths. First, more than 200,000 complete observations of the outcomes and the predictors were used to train the models, providing a sufficiently large pool of training data. Second, in addition to the machine-learning models, a simpler logistic regression model was also trained and evaluated, which performed comparably to the more complex models. Although much attention has been paid to machine-learning models, simpler, more interpretable models are preferred in clinical practice and therefore they should be included in the prediction model development process. In conclusion, ML-based prediction models relying only on structured nursing data collected daily can reliably identify patients at risk for delirium. The models presented in this study achieved similar performance as more complex delirium prediction models. Yet, the presented models are likely more easily transferable to other hospitals collecting the same structured nursing data. However, validation across different sites is necessary before the models can be implemented in clinical practice.

Author statement

Conceptualization: TRS, ET, SB, JE.
Methodology: TRS, OD, ZBZ, NK, SF.
Investigation: TRS, SB.
Visualization: TRS, OD, ZBZ, NK.
Funding acquisition: HP, RvK.
Project administration: HP.
Supervision: HP, IHR, RvK.
Writing – original draft: TRS, ET, JE.
Writing – review & editing: All authors.

Declaration of competing interest

There are no conflict of interests.

Acknowledgment

Funding/Support: TRS was supported by an Early.Postdoc Mobility Fellowship of the Swiss National Science Foundation, grant no. [P2ZHP3_195191].

Role of the Funder/Sponsor: The funding sources had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data sharing statement: Data are subject to legal restrictions on confidential data of the University Hospital Zurich (USZ) and cannot be shared by the authors. Free access to the data can be requested from the USZ, however, access might be denied. The corresponding author can be contacted for further information.

Disclaimer: The views expressed in this article are solely those of the authors and do not reflect an endorsement by or the official policy or position of the U.S. Department of Veterans Affairs or the United States Government.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpsychires.2022.10.018.

References

American Psychiatric Association, 2013. In: Diagnostic and Statistical Manual of Mental Disorder, fifth ed. Washington, DC.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., Vardoulakis, L.M., 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '20: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, pp. 1–12. https://doi.org/10.1145/3313831.3376718.

Boettger, S., Zipser, C.M., Bode, L., Spiller, T., Deuel, J., Osterhoff, G., Ernst, J., Petry, H., Volbracht, J., von Känel, R., 2020. The prevalence rates and adversities of delirium: too common and disadvantageous. Palliat. Support Care 1–9. https://doi.org/10.1017/s1478951520000632.

Chua, S.J., Wrigley, S., Hair, C., Sahathevan, R., 2021. Prediction of delirium using data mining: a systematic review. J. Clin. Neurosci. 91, 288–298. https://doi.org/10.1016/j.jocn.2021.07.029.

Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.M., 2015. Transparent reporting of a multivariable prediction model for individual Prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann. Intern. Med. 162, 55–63. https://doi.org/10.7326/M14-0697.

De, J., Wand, A.P.F., 2015. Delirium screening: a systematic review of delirium screening tools in hospitalized patients. Gerontol. 55, 1079–1099. https://doi.org/10.1093/geront/gnv100.

de Wit, H.A.J.M., Winkens, B., Mestres Gonzalvo, C., Hurkens, K.P.G.M., Mulder, W.J., Janknegt, R., Verhey, F.R., van der Kuy, P.-H.M., Schols, J.M.G.A., 2016. The development of an automated ward independent delirium risk prediction model. Int. J. Clin. Pharm. 38, 915–923. https://doi.org/10.1007/s11096-016-0312-7.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837. https://doi.org/10.2307/2531595.

Gavinski, K., Carnahan, R., Weckmann, M., 2016. Validation of the delirium observation screening scale in a hospitalized older population: delirium Screening in Older Patients. J. Hosp. Med. 11, 494–497. https://doi.org/10.1002/jhm.2580.

Gemert van, L.A., Schuurmans, M.J., 2007. The neecham confusion scale and the delirium observation screening scale: capacity to discriminate and ease of use in clinical practice. BMC Nurs. 6 https://doi.org/10.1186/1472-6955-6-3.

Hunstein, D., 2012. ePA-AC©: Ergebnisorientiertes PflegeAssessment AcuteCare, Version 2.0.

Inouye, S.K., Viscoli, C., Horowitz, R., Hurst, L., Tinetti, M., 1993. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. Ann. Intern. Med. 119, 474. https://doi.org/10.7326/0003-4819-119-6-199309150-00005.

Inouye, S.K., Westendorp, R.G., Saczynski, J.S., 2014. Delirium in elderly people. Lancet 383, 911–922. https://doi.org/10.1016/S0140-6736(13)60688-1.

Jauk, S., Kramer, D., Großauer, B., Rienmüller, S., Avian, A., Berghold, A., Leodolter, W., Schulz, S., 2020. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. J. Am. Med. Inf. Assoc. 27, 1383–1392. https://doi.org/10.1093/jamia/ocaa113.

Kim, M.S., Rhim, H.C., Park, A., Kim, H., Han, K.-M., Patkar, A.A., Pae, C.-U., Han, C., 2020. Comparative efficacy and acceptability of pharmacological interventions for the treatment and prevention of delirium: a systematic review and network meta-analysis. J. Psychiatr. Res. 125, 164–176. https://doi.org/10.1016/j.jpsychires.2020.03.012.

Koster, S., Hensens, A.G., Oosterveld, F.G.J., Wijma, A., van der Palen, J., 2009. The delirium observation screening scale recognizes delirium early after cardiac surgery. Eur. J. Cardiovasc. Nurs. 8, 309–314. https://doi.org/10.1016/j.ejcnurse.2009.02.006.

Labarère, J., Bertrand, R., Fine, M.J., 2014. How to derive and validate clinical prediction models for use in intensive care medicine. Intensive Care Med. 40, 513–527. https://doi.org/10.1007/s00134-014-3227-6.

Lindroth, H., Bratzke, L., Purvis, S., Brown, R., Coburn, M., Mrkobrada, M., Chan, M.T.V., Davis, D.H.J., Pandharipande, P., Carlsson, C.M., Sanders, R.D., 2018. Systematic review of prediction models for delirium in the older adult inpatient. BMJ Open 8, e019223. https://doi.org/10.1136/bmjopen-2017-019223.

McCoy, T.H., Snapper, L., Stern, T.A., Perlis, R.H., 2016. Underreporting of delirium in statewide claims data: implications for clinical care and predictive modeling. Psychosomatics 57, 480–488. https://doi.org/10.1016/j.psym.2016.06.001.

Oh, E.S., Fong, T.G., Hshieh, T.T., Inouye, S.K., 2017. Delirium in older persons: advances in diagnosis and treatment. JAMA 318, 1161. https://doi.org/10.1001/jama.2017.12067.

R Core Team, 2020. R: A Language and Environment for Statistical Computing.

Rudolph, J.L., Doherty, K., Kelly, B., Driver, J.A., Archambault, E., 2016. Validation of a delirium risk assessment using electronic medical record information. J. Am. Med. Dir. Assoc. 17, 244–248. https://doi.org/10.1016/j.jamda.2015.10.020.

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E.W., Stahl, D., Fusar-Poli, P., 2021. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. Schizophr. Bull. 47, 284–297. https://doi.org/10.1093/schbul/sbaa120.

Scheffer, A.C., van Munster, B.C., Schuurmans, M.J., de Rooij, S.E., 2011. Assessing severity of delirium by the delirium observation screening scale. Int. J. Geriatr. Psychiatr. 26, 284–291. https://doi.org/10.1002/gps.2526.

Schubert, M., Schürch, R., Boettger, S., Garcia Nuñez, D., Schwarz, U., Bettex, D., Jenewein, J., Bogdanovic, J., Staehli, M.L., Spirig, R., Rudiger, A., 2018. A hospital-wide evaluation of delirium prevalence and outcomes in acute care patients - a cohort study. BMC Health Serv. Res. 18 https://doi.org/10.1186/s12913-018-3345-x.

Siddiqi, N., House, A.O., Holmes, J.D., 2006. Occurrence and outcome of delirium in medical in-patients: a systematic literature review. Age Ageing 35, 350–364. https://doi.org/10.1093/ageing/afl005.

Sun, H., Depraetere, K., Meesseman, L., De Roo, J., Vanbiervliet, M., De Baerdemaeker, J., Muys, H., von Dossow, V., Hulde, N., Szymanowsky, R., 2021. A scalable approach for developing clinical risk prediction applications in different hospitals. J. Biomed. Inf. 118, 103783 https://doi.org/10.1016/j.jbi.2021.103783.

Weiskopf, N.G., Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inf. Assoc. 20, 144–151. https://doi.org/10.1136/amiajnl-2011-000681.

Wong, A., Young, A.T., Liang, A.S., Gonzales, R., Douglas, V.C., Hadley, D., 2018. 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. https://doi.org/10.1001/jamanetworkopen.2018.1018.

Young, J., Murthy, L., Westby, M., Akunne, A., O'Mahony, R., on behalf of the Guideline Development Group, 2010. Diagnosis, prevention, and management of delirium: summary of NICE guidance. BMJ 341. https://doi.org/10.1136/bmj.c3704 c3704–c3704.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 15, e1002683 https://doi.org/10.1371/journal.pmed.1002683.

Zipser, C.M., Spiller, T.R., Hildenbrand, F.F., Seiler, A., Ernst, J., von Känel, R., Inouye, S.K., Boettger, S., 2022. Discharge destinations of delirious patients: findings from a prospective cohort study of 27,026 patients from a large health care system. J. Am. Med. Dir. Assoc. https://doi.org/10.1016/j.jamda.2022.01.051.