

Research and Applications

Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study

Stefanie Jauk ^{1,2} Diether Kramer,¹ Birgit Großauer,³ Susanne Rienmüller,³ Alexander Avian,² Andrea Berghold,² Werner Leodolter,¹ and Stefan Schulz²

¹Department of Information and Process Management, Steiermärkische Krankenanstaltengesellschaft m.b.H. (KAGes), Graz, Austria, ²Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria, and ³Department of Internal Medicine, Steiermärkische Krankenanstaltengesellschaft m.b.H. (KAGes) LKH Graz II, Graz, Austria

Corresponding Author: Stefanie Jauk, MSc, Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2, 8036 Graz, Austria; stefanie.jauk@kages.at

Received 5 November 2019; Revised 11 March 2020; Editorial Decision 14 May 2020; Accepted 20 May 2020

ABSTRACT

Objective: Machine learning models trained on electronic health records have achieved high prognostic accuracy in test datasets, but little is known about their embedding into clinical workflows. We implemented a random forest-based algorithm to identify hospitalized patients at high risk for delirium, and evaluated its performance in a clinical setting.

Materials and Methods: Delirium was predicted at admission and recalculated on the evening of admission. The defined prediction outcome was a delirium coded for the recent hospital stay. During 7 months of prospective evaluation, 5530 predictions were analyzed. In addition, 119 predictions for internal medicine patients were compared with ratings of clinical experts in a blinded and nonblinded setting.

Results: During clinical application, the algorithm achieved a sensitivity of 74.1% and a specificity of 82.2%. Discrimination on prospective data (area under the receiver-operating characteristic curve = 0.86) was as good as in the test dataset, but calibration was poor. The predictions correlated strongly with delirium risk perceived by experts in the blinded ($r=0.81$) and nonblinded ($r=0.62$) settings. A major advantage of our setting was the timely prediction without additional data entry.

Discussion: The implemented machine learning algorithm achieved a stable performance predicting delirium in high agreement with expert ratings, but improvement of calibration is needed. Future research should evaluate the acceptance of implemented machine learning algorithms by health professionals.

Conclusions: Our study provides new insights into the implementation process of a machine learning algorithm into a clinical workflow and demonstrates its predictive power for delirium.

Key words: Machine learning, prospective studies, delirium, electronic health records, clinical decision support

INTRODUCTION

Background and significance

In today's clinical practice, patients are often routinely classified into risk groups, with the purpose to predict future outcomes of treatments and evolution of diseases.¹ Owing to the increasing amount and availability of clinical data stored in electronic health record (EHR) systems, prediction models based on machine learning algorithms have become popular,² as they overcome barriers typical for classical modelling approaches.³

Although the superiority of machine learning models over rule-based models has been shown in test scenarios,^{4–6} there is an urgent need for prospective evaluation studies that demonstrate their actual value in clinical settings.^{7,8} However, the implementation of such complex models in clinical practice faces various obstacles and barriers.^{2,9,10} Little is known about the integration of machine learning algorithms into clinical workflows, or about the performance of predictive models in dynamic situations. There are still many lessons to be learned about the challenges that might occur during the implementation process, and there is little evidence regarding the uptake of machine learning models by health professionals.^{9,11}

The delirium use case

For a successful implementation in clinical practice, the predicted outcome needs to be controllable and actionable,¹² like the prediction of delirium in hospitalized patients. Delirium is a syndrome of acute confusional state and is common among elderly patients. In general medical departments, up to 49% of patients suffer from delirium.¹³ Besides causing a burden for the healthcare personnel, hospitalized delirium patients also have an increased risk of mortality. However, many delirium cases can be prevented using nonpharmacological interventions with multiple components (eg, reinforcement of visual and hearing aids, hydration, reorientation to surroundings, bed time protocols, noise reduction).^{14,15} Targeting patients with highest risk of delirium is therefore crucial, but established methods have their flaws. The use of the Delirium Observation Scale¹⁶ for delirium risk assessment in clinical routine is time-consuming, and it is rather used for assessing first signs and symptoms of delirium than for prediction. The Confusion Assessment Method¹⁷ is widely used as a screening instrument, although it is at the same time the established tool for diagnosing delirium.

Prediction models for delirium

In 2017, we had trained various machine learning models predicting the occurrence of delirium in internal medicine patients.¹⁸ The data were provided by the public hospital provider Steiermärkische Krankenanstaltengesellschaft m.b.H. (KAGes), which hosts longitudinal health records from 2.1 million patients in the province of Styria, Austria. The models had been trained on demographic data, previous International Classification of Diseases–Tenth Revision (ICD-10)–coded diagnoses, laboratory data, nursing assessment and procedures from the EHRs of more than 8500 patients, and the predicted outcome was an ICD-10–coded diagnosis F05 (Delirium due to known physiological condition). A random forest–based model had achieved the best performance with an area under the receiver-operating characteristic curve (AUROC) of 0.910.

Several prediction models for delirium have been reported but were commented to be insufficiently reliable for clinical use,^{19,20} were not generalizable for several populations,²¹ or were never adopted in clinical practice.²² Similar to our approach, other ma-

chine learning models predicting delirium based on EHR have recently been published.^{23–25} The reported performances were similar, in some cases even better than in our model, with AUROCs ranging from 0.86 to 0.94.

Although most of the published models performed very well in retrospective datasets, to our knowledge none of the reported machine learning models and the systematically reviewed rule-based models has been implemented and prospectively evaluated in a clinical workflow.

OBJECTIVE

The aim of this study was (1) to implement a previously developed machine learning algorithm predicting delirium in a clinical workflow involving hospitalized patients and (2) to prospectively evaluate this algorithm in a clinical setting. The goal of the prediction was to identify patients at high risk for the occurrence of delirium during the current hospital stay. The risk was predicted at the beginning of the stay as well as on the evening of the admission day in order to take action as soon as possible. Our evaluation focused on the analysis of the predictive performance of the algorithm, as well as on the validation of its accuracy by comparing its outcome to expert ratings.

Machine learning models are often seen as black boxes.⁹ Therefore, a major goal of the implementation was to explain the prediction results as well as possible and support decision making of the health care personnel.

MATERIALS AND METHODS

We used the STARE-HI (Statement on Reporting of Evaluation Studies in Health Informatics) statement²⁶ as a guideline for reporting (see [Supplementary Table 1](#)).

Delirium prediction algorithm

In a first step, we adapted and expanded the previously trained random forest¹⁸ by:

- Extending the internal medicine cohort by including patients from surgical departments.
- Training the F05 model on a larger cohort with over 19 000 patients.
- Training a second random forest model predicting the ICD-10 code F10.4 (Alcohol withdrawal state with delirium). Although quite distinct from the syndrome coded by F05 in terms of etiology and pathophysiology, clinical experts found it crucial to include both predictions because of their similarity in signs, symptoms and consequences.
- Assigning the highest risk score for all patients with a delirium code F05 from an earlier hospital stay, as a history of delirium is highly associated with the current delirium risk.²⁷

For implementation, both models F05 and F10.4 were combined into one algorithm. Delirium risk was calculated with both models separately and only the higher risk score of both was presented. Examples of features used for prediction are summarized in [Supplementary Table 2](#). [Figure 1](#) shows the receiver-operating characteristic (ROC) curves and calibration plots for the test datasets of both models. Owing to a more heterogeneous cohort, the overall performance of the F05 model decreased slightly compared with our

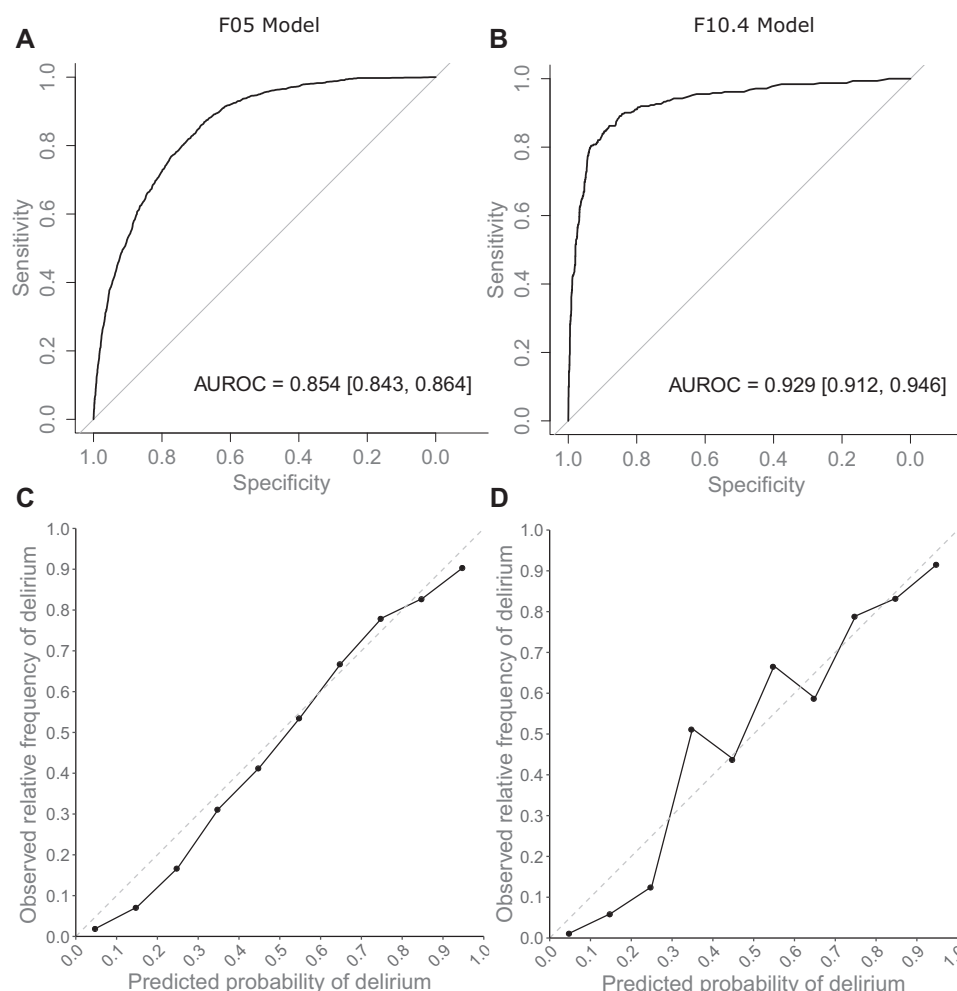


Figure 1. (A, B) Area under the receiver-operating characteristic curve (AUROC) and (C, D) calibration plots for the random forest model predicting International Classification of Diseases–Tenth Revision code F05 (Delirium due to known physiological condition) trained on 19 905 patients and a random forest model predicting International Classification of Diseases–Tenth Revision code F10.4 (Alcohol withdrawal state with delirium) trained on 9872 patients.

previously published model.¹⁸ The F05 model achieved an AUROC of 0.85 (Figure 1A), and the F10.4 model an AUROC of 0.93 (Figure 1B). Predictions of the F05 model were well calibrated (Figure 1C), while those for the F10.4 model showed slight deviations from the optimal calibration line (Figure 1D).

Implementation process and visualization in the hospital information system

The implementation process started in November 2017, when a first expert group meeting took place at the pilot site, in the Austrian public hospital LKH Graz II (see Figure 2). This expert group was set up in order to enhance participation of health professionals, including senior physicians, ward nurses, technicians, and leading employees. In total, the expert group held 4 meetings before and during the evaluation period. Besides deciding how the prediction results should be visualized in the hospital information system (HIS), the expert group defined probability thresholds for 3 delirium risk classes. Prevalence of delirium and current resources for prevention were considered, and finally clinical experts agreed on presenting the top 5% of highest rated patients as “very high risk,” followed by the next 10% as “high risk,” and the remaining 85% as “low risk.” We determined the thresholds of the model in a separate

test dataset from the participating clinical departments using these percentages. Predicted probabilities from the random forest models on this dataset were ranked, and cutoffs for the risk classes were set at the 85th and 95th percentiles.

The calculated delirium risk was displayed to health professionals using 2 presentation methods. First, we added a new column named “Prognose” (German for *prediction*) to the clinical workplace within the HIS. A red icon symbolized patients at “very high risk” (95th to 100th percentile) and a yellow icon those at “high risk” (85th to 94th percentile), shown in Figure 3. In order to avoid an information overflow at the clinical workplace, no symbol was shown for “low-risk” patients.

Second, patient-specific features relevant for prediction were presented in a web application developed in R shiny,²⁸ which opens up with a click on the icon or empty cell. It displays demographic data together with 4 separate boxes with previous ICD-10 codes and diagnosis texts, laboratory results, procedures and remaining information (eg, nursing assessment data, Charlson Comorbidity Index) for each patient (see Figure 4).

In February and March 2018, training sessions for health professionals were offered with the objective to stimulate the understanding and the uptake of the application. After the training

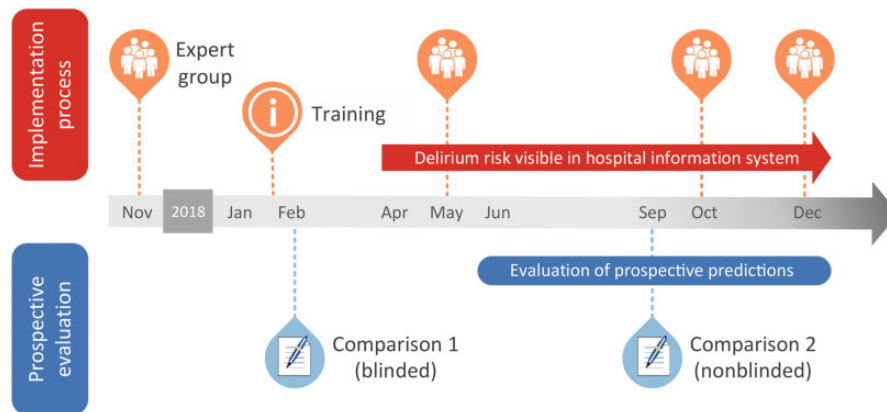


Figure 2. Timeline for the implementation process and evaluation study design.

Belegung MED Gesamt Pflege vom								
Pfl. OE	Zimmer	Bett	Patient/Geschl./Alter	Kw	Prognose	Warn	MIBI St.	
GEMS1	D157	D157-2	Wald, Gertrud (M, 85)	Δ		!		
	D158	D158-1	Mahner, Cornelia (F, 78)	Δ		!	MRE	
		D158-2	Fischer, Barbara (F, 84)	Δ	Δ	!		
	D160	D160-1	Mahner, Cornelia (F, 78)	Δ		!		
		D160-2	Mahner, Cornelia (F, 78)	Δ	Δ	!		
	D161	D161-2	Tangemann, Andrea (F, 84)	Δ	Δ	!	MRE	
	D162	D162-1	Fischer, Barbara (F, 84)	Δ		!		
		D162-2	Fischer, Barbara (F, 84)	Δ		!		
	D163	D163-2	Wald, Gertrud (M, 85)	Δ		!		
	D164	D164-1	Mahner, Cornelia (F, 78)	Δ		!		
GEMS2		D164-2	Mahner, Cornelia (F, 78)	Δ		!		
	D166	D166-2	Fischer, Barbara (F, 84)	Δ				
	D260	D260-1	Mahner, Cornelia (F, 78)	Δ		!	MRE	
		D260-2	Mahner, Cornelia (F, 78)	Δ				

Figure 3. Presentation of delirium risk on the clinical workplace of the KAGes hospital information system *openMEDOCS*, based on IS-H/i.s.h.med information systems and implemented on platforms provided by the software corporation SAP SE.

sessions, the application was started at the pilot site. The use of the application was voluntary, and no recommendation was given by the system on how to proceed with patients at risk.

Study design

For every admission to a surgical or internal medicine department, delirium risk was computed by the algorithm at admission time. The risk was then recalculated on the evening of the admission day if until this time no delirium diagnosis had been coded. This recalculation included the most recent laboratory results and nursing assessment data. All risk predictions and values of the features were stored in a data warehouse.

In our hospital network, diagnoses are often coded in the EHR close to discharge or up to 14 days after discharge. Therefore, occurrence of delirium was defined as being diagnosed and coded with the ICD-10 code F05 (including all subcategories) or F10.4 during the hospital stay and up to 14 days after discharge. In addition, free-text patient summaries from this period were screened for words related to delirium using an approximate string match. Summaries

with a positive screening result were manually checked, and patients with evidence of delirium were added to the delirium patient group.

We performed a prospective evaluation study using 2 methods. The timeline of the evaluation is shown in Figure 2.

First, we evaluated the prospective predictions of the algorithm from June 1, 2018, until December 31, 2018. All predictions for patients admitted during this 7-month period were included in the analysis, excluding patients younger than 18 years of age. To assess the predictive performance of the algorithm, we compared the outcome for every case (delirium or nondelirium) with the calculated risk category (low, high, very high).

Second, prospective comparisons between the algorithm and clinical experts on a sample of internal medicine patients were performed. We developed a protocol for the clinical assessment to be completed by experienced ward nurses within the first 24 hours of a patient's hospital stay. The protocol included 1 item with a subjective rating of delirium risk on a scale of 1 (very low) to 5 (very high), and all 5 items of the Confusion Assessment Method¹⁷ as a measurement for delirium. For their subjective risk estimation, the nurses considered all information available such as information stored in the HIS as well as the current clinical condition of the patient.

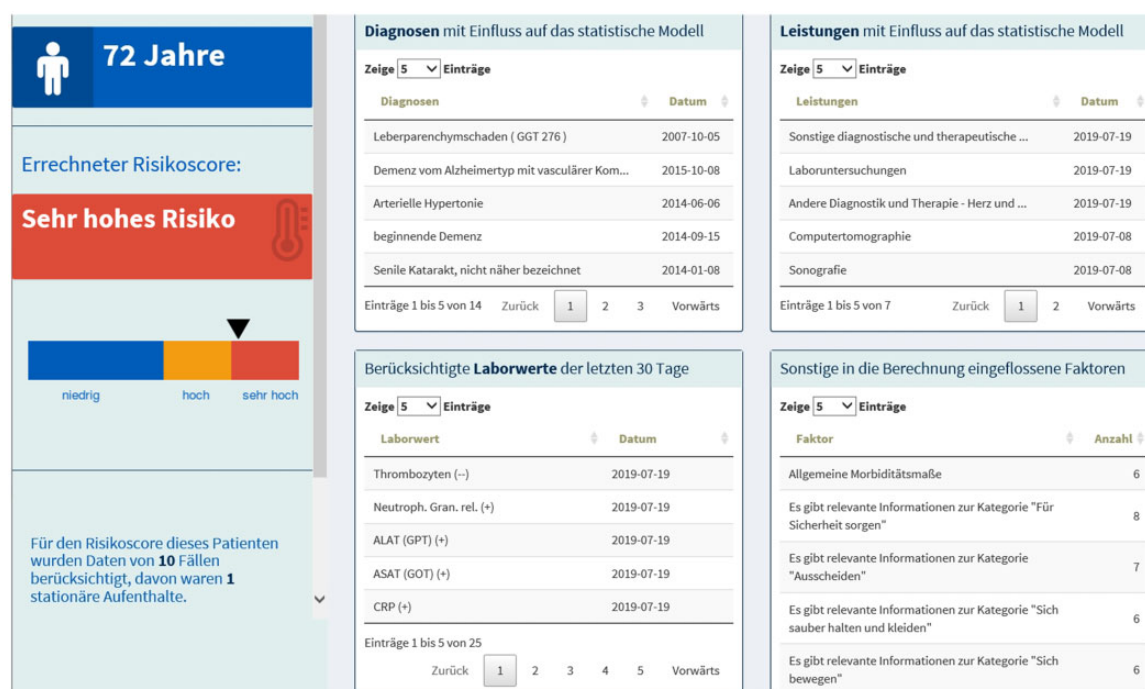


Figure 4. Web application visualizing features used in the random forest model. The delirium risk was presented in the left column of the application, utilizing a bar for the continuous risk probability. Features were categorized into 4 feature groups: International Classification of Diseases–Tenth Revision–coded diagnoses, procedures, laboratory data, and other features.

In addition, nurses reported any comments in a free-text entry field. The risk rating by the nurses (5 categories) was then compared with the risk prediction of the delirium algorithm.

The first comparison (comparison 1—blinded) was conducted in February 2018 before the application was visible in the HIS, and was thus blinded. One ward nurse completed the protocol for all patients admitted to her ward over a period of 14 days ($n=33$). This comparison provided a first quality assessment of the algorithm's accuracy in a clinical workflow.

In September 2018, the second, nonblinded comparison was performed (comparison 2—nonblinded). Two ward nurses from 2 different internal medicine wards recorded the ratings for their patients over 14 days ($n=86$). Again, these expert ratings were compared with the algorithm's results.

Data analysis

All data were analyzed in R version 3.5 (R Foundation for Statistical Computing, Vienna, Austria). First, we analyzed the predictive performance on each admission during the 7-month evaluation period. We evaluated the latest prediction results within the first 24 hours of the hospital stay for each admission. We combined the "high-risk" and "very high-risk" groups, and used the lower threshold for the calculation of sensitivity, specificity, false positive rate, false negative rate, positive predictive value, and negative predictive value. All false negative cases were qualitatively examined in order to investigate potential weaknesses of the algorithm.

The results of the prospective prediction were then compared with the test datasets of the F05 model and the F10.4 model. As a measure of discrimination, we used ROC curves with DeLong confidence intervals.²⁹ To measure calibration, we calculated the frequencies of delirium over the 3 risk classes and computed a calibration plot with a 95% confidence interval.

Second, the protocols completed by the nurses were analyzed using descriptive statistics. Relationships between the nurses' ratings and the algorithm were evaluated with Spearman's rank correlation coefficient (r), with statistical significance defined at an alpha level of 0.05. In addition, patients with differences in risk estimation were analyzed qualitatively focusing on the amount of available information in the HIS and delirium relevant features.

RESULTS

Descriptive statistics of the cohort for prospective prediction as well as of the cohorts of comparisons 1 and 2 are presented in Table 1. For all patients, the entire EHR was screened for coded diagnoses relevant for delirium across all KAGes hospitals since 2004.

Prospective performance of the delirium algorithm

During the 7-month evaluation period, the delirium risk was prospectively calculated for 5647 admissions of 4765 patients. For 113 admissions, a technical error had occurred during the risk estimation and they were excluded from the analysis. Four patients were younger than 18 years of age and therefore were excluded from the analysis. This resulted in a cohort of 5530 admissions of 4663 patients for analysis.

Results of the prospective prediction are presented in Table 2. Out of the nondelirium cases, 82.2% were identified as "low risk" (specificity). Thus, the false positive rate was 0.178. In total, 81 admissions (1.5%) developed a delirium during the stay ($n_{F05}=67$; $n_{F10.4}=14$), of which 74.1% were predicted as "high-risk" or "very high-risk" patients (sensitivity). The false negative rate was 0.259, with 21 of 81 undetected cases of delirium. Positive predictive value for prospective prediction was 0.058 and negative predictive value was 0.995.

Table 1. Descriptive statistics for all analyzed cohorts including delirium relevant diagnoses in the electronic health records

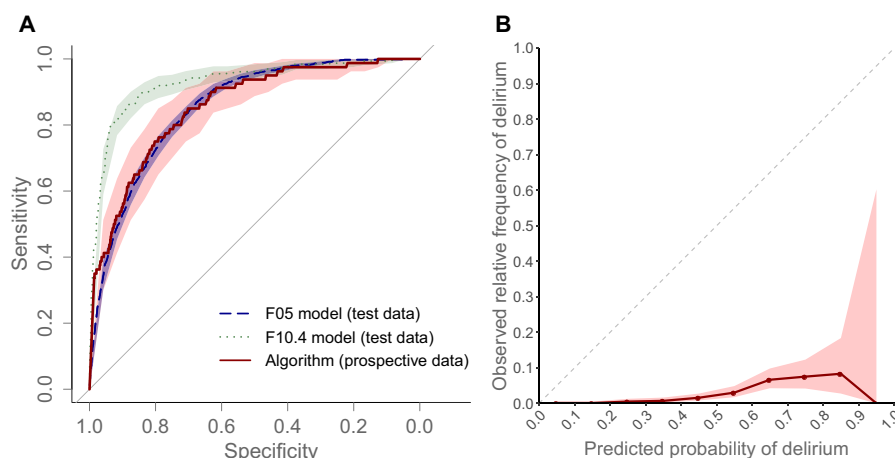
	Comparison 1 (n = 33)	Comparison 2 (n = 86)	Prospective prediction (n = 5530)
Age, y	71 (63-80)	77 (64-84)	71 (57-80)
Sex			
Male	16 (48.5)	44 (51.2)	2924 (52.9)
Female	17 (51.5)	42 (48.8)	2606 (47.1)
Previously coded diagnoses			
Delirium	0 (0.0)	8 (9.3)	104 (1.9)
Dementia	2 (6.1)	10 (12.8)	245 (4.4)
Parkinson's disease	0 (0.0)	3 (3.5)	87 (1.6)
Depression	7 (21.2)	16 (18.6)	574 (10.4)
Alcohol abuse	3 (9.1)	8 (9.3)	236 (4.3)
Substance abuse(excl. alcohol, nicotine)	1 (3.0)	1 (1.2)	98 (1.8)

Values are median (interquartile range) or n (%).

Table 2. Confusion matrix for the prospective prediction of delirium for all evaluated admissions

		Prediction		
		No delirium (low)	Delirium (high, very high)	Total
Outcome	No delirium	4479 (82.2)	970 (17.8)	5449 (100.0)
	Delirium	21 (25.9)	60 (74.1)	81 (100.0)
	Total	4500 (81.4)	1030 (18.6)	5530 (100.0)

Values are n (%).

**Figure 5.** Performance of the machine learning algorithm during prospective prediction. (A) Receiver-operating characteristic curve of the prospective prediction (in red) compared with the 2 random forest models on test datasets. (B) Calibration plot with 95% confidence interval of the algorithm applied to the prospective cohort.

Three of the 21 undetected cases suffered from an alcohol withdrawal state with delirium (F10.4); they were between 30 and 62 years of age, and the prediction only relied on data from the current hospital stay. For another 3 cases with a delirium diagnosis, available information was sparse, too. The remaining 15 cases had a median risk probability of 0.47 (min = 0.33, max = 0.57), and were thus all (except 1 case) above the third quartile of all 4477 cases in the “low-risk” category (median = 0.20 [interquartile range, 0.08-0.36]).

As shown in Figure 5A, discriminative performance of the algorithm during the prospective prediction (red) (AUROC = 0.855) was as good as for the F05 model (blue) (AUROC = 0.854), but lower than for the F10.4 model (green) (AUROC = 0.929). The av-

erage predicted risk was higher than the overall event rate during the prospective prediction over all 3 risk classes (Table 3) and for all percentiles of predicted probabilities (Figure 5B). This indicates a poorer calibration for the prospective data than for the test data.

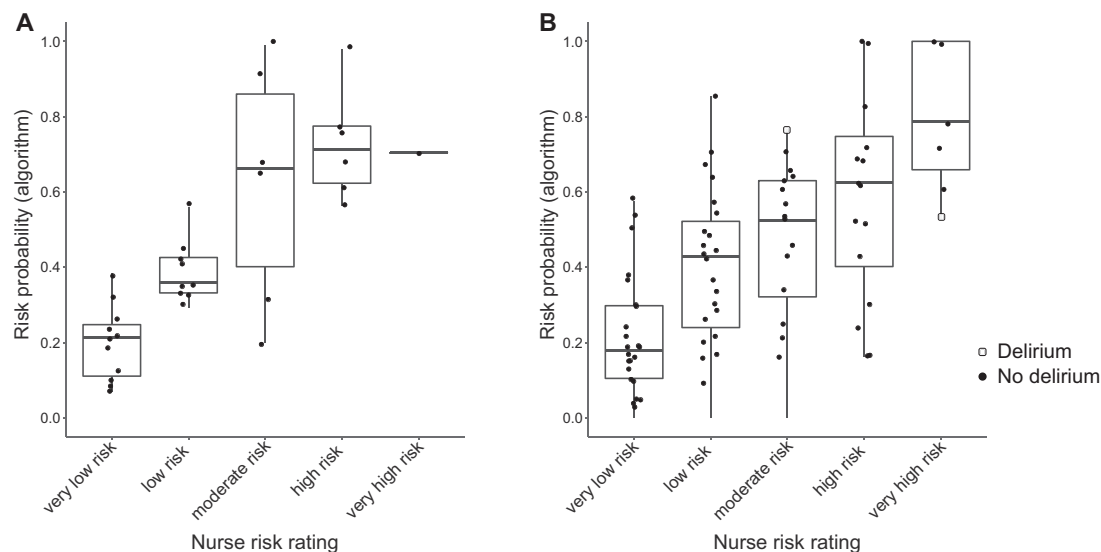
Comparison with expert ratings

In the blinded comparison 1, the ward nurse observed 33 patients within 24 hours after admission (see Figure 6A). There was a high correlation between the nurse's rating and the algorithm's estimation ($r=0.81$, $P<.001$). No patient of the cohort was coded as delirium. For all patients that were classified as “very low risk” by the nurse, the estimation of the algorithm was low as well.

Table 3. Frequencies of the outcome delirium over the 3 risk classes for test data F05, test data F10.4, and the prospective data

	Predicted risk			Total
	Low	High	Very high	
Probability thresholds	0-0.576	0.576-0.714	0.714-1	
Test data F05	3854	593	528	4975
With outcome	806 (20.9)	382 (64.4)	436 (82.4)	1624 (32.6)
Test data F10.4	2276	69	123	2468
With outcome	164 (7.2)	44 (63.8)	104 (84.6)	312 (12.6)
Prospective data	4500	585	445	5530
With outcome	21 (0.5)	16 (2.7)	44 (9.9)	81 (1.5)

Values are probabilities (0 - 1), n or n (%).

**Figure 6.** Comparisons of nurses' risk ratings and algorithm's estimation for internal medicine patients in (A) the blinded setting (comparison 1; n = 33) and (B) the non-blinded setting (comparison 2; n = 86). Two cases of delirium in comparison 2 are highlighted.

For the nonblinded comparison 2 (n = 86), shown in Figure 6B, there was also a correlation between the nurses' assessments and the algorithm's calculation ($r=0.62$, $P<.001$). Again, all 24 patients identified as "very low-risk" patients by the nurses were accordantly estimated a low-risk probability by the algorithm.

Six patients of comparison 2 had an occurrence of delirium in their history but no delirium diagnosis during the hospital stay. All of them were classified as high-risk patients by both the nurses and the algorithm. Two patients of the cohort developed a delirium during their stay. One patient was correctly identified by the nurse only, with the other one by the algorithm only.

The qualitative analysis showed 2 main reasons for contrary risk estimations. First, in some cases with little information stored in the EHR, the algorithm's predictions were low, whereas the nurses' estimations were high. Second, for some other cases, the nurses commented that they were uncertain about the delirium risk. One reported reason for this uncertainty was a communication barrier due to sedation or language. The algorithm predicted high-risk probabilities for these cases.

DISCUSSION

In this study, we implemented a machine learning algorithm in a clinical workflow to identify patients at high risk for delirium, and evaluated it prospectively during 7 months. The implementation

process included several meetings with clinical experts as well as training for nurses and physicians in the departments. We implemented a visualization of the delirium risk prediction inside the HIS in order to highlight patient-specific EHRs used by the machine learning algorithm.

In the prospective evaluation, we first analyzed the performance of the algorithm for 5530 hospitalizations. During the evaluation period of 7 months, its discriminative performance was very high, with a specificity of 82.2% and a sensitivity of 74.1%. Compared with a similar delirium prediction model,²⁴ the predictive performance of our algorithm was superior, even though it was evaluated in a clinical workflow and not on test data only. Compared with our own test data, the algorithm's performance was as good as the performance of the model prediction the ICD-10 diagnosis code F05.

The overall goal of the implementation was to identify patients at high risk of delirium at the beginning of the hospital stay. As preventive actions for delirium are not harmful for patients,¹⁴ false positive cases may only lead to inefficient use of clinical resources because of many interventions. In contrast, an occurrence of delirium is associated with higher mortality rates and complications,¹³ and for this reason the false negative rate must be kept as low as possible.

Our clinical setting did not allow controlling for staged interventions to prevent delirium. Such interventions might influence the delirium occurrence rate and thus the results of our analysis. We

observe a self-destroying prophecy, a phenomena known from sociology.³⁰ The algorithm predicts a high risk of delirium, and the delirium is successfully prevented due to interventions in the ward. In our analysis however, such a patient is categorized as false positive.

Another scenario might lead to an overestimation of the false positive rate. All cases with a delirium diagnosis in their past were classified as “very high risk” ($n = 104$). Of these, 28 (26.9%) cases developed another delirium during their recent admission. Even though the patients of the remaining 76 cases did not develop a delirium, we still perceive them as high-risk patients. Considering this, we assume the real specificity and positive predictive value to be slightly higher than reported.

Because we could not control for all effects in the analysis of the prospective predictions, we used a second method for evaluation that compared the estimations of the algorithm with those of experienced ward nurses. For the majority of the cases, the ratings of the ward nurses and the algorithm were in agreement. The qualitative analysis revealed not only a known weakness of the algorithm (eg, risk prediction based on little information), but also its strengths. Especially in cases of uncertainty, the application seems to provide a good support for delirium management on the ward. Patients unable to communicate with the health professionals (eg, patients under sedation or with language barriers) benefit from the support by the delirium application.

There are 2 more major advantages of the algorithmic approach compared with well-established screening methods in the clinical workflow. First, no additional information is needed to compute the delirium risk because the algorithm uses already documented information of the EHR only. Second, the algorithmic prediction is thus faster than standardized screening methods and not dependent on any clinical resources.

Strengths and weaknesses of the study

For more than 5530 hospitalizations evaluated during 7 months, the occurrence rate of delirium was 1.5%, which is lower than reported in studies and guidelines ranging from 10% to 40%.^{31,32} One reason for this is that delirium is not always coded in the participating hospital, and sometimes it is not even mentioned in the discharge summary. Our application might improve the administrative documentation of delirium, as it sensitizes physicians to the topic of delirium, and demonstrates the importance of a precise documentation in EHR for secondary use.

This highlights a limitation of our study. Although standardized tools for diagnosing delirium are available, delirium diagnoses are often based on subjective perceptions of a patient’s condition. Different types (eg, hypoactive or hyperactive delirium) and different degrees of severity complicate the clinical evaluation even more. This lack of clear diagnostic criteria might be one reason why the incidence of delirium according to ICD codes in an administrative database is lower than the one reported in prospective studies.³²

The low incidence of delirium is also reflected in the calibration plot of the prospective data. While calibration in test data was good, predictions for the prospective cohort were systematically too high. However, with an incidence of only 1.5% and the bias of self-destroying prophecy, strong calibration for our clinical setting is hard to achieve. The comparison with clinical expert ratings showed that some patients were perceived as high-risk patients in the ward, but did not develop a delirium or showed no record of delirium. Neither calibration plots nor confusion matrices are able to account for such cases. Nevertheless, the poor calibration highlights the need

for calibrating machine learning classifiers in prospective scenarios. Further analyses need to determine whether refitting the models or calibrating the predictions using Platt scaling or isotonic regression can improve calibration results.

Another shortcoming of our study concerns the applicability of the algorithm to other hospital environments. The algorithm was trained on a cohort of patients across all institutions of the same hospital network, and it is unclear whether the prediction will be accurate enough for other hospitals with different EHR systems, workflows, and coding policies. We plan to address this question in future studies.

The last limitation to be mentioned is the availability of information stored in the EHR. We observed that for some patients with few previous stays in the hospital network, the algorithm underestimated the delirium risk compared with clinical experts. We tried to overcome this problem by updating the delirium risk in the evening of admission considering new laboratory data, which are available within 1 hour after admission, and nursing data, which are available within the first hours after admission. However, the information provided between admission and recalculation did not always seem sufficient for reliable risk estimation. More research is needed on this topic to determine the amount of information needed for reliable prediction.

Future research opportunities

In the implemented version of the algorithm, the most recent nursing assessment and laboratory results from the admission day were used for training. However, if a prediction is made right at time of admission, for some patients important data might still be missing.³⁴ Previous research showed that clinical decision support based on EHR is influenced by late data entry in some cases.³⁵ Hence, in a next step we will test different models for 3 calculation times (admission, first evening, and second evening) in order to account for time-delayed information availability. In the future, this problem might be overcome with the use of nationwide EHR systems that better represent a whole patient history.

Finally, and most importantly, future studies should focus on the perception and uptake of the application by health professionals. Without their support and their trust in algorithmic decision support, the success of such approaches will not last long, especially when using complex machine learning models that are not easily explainable. Hence, an in-depth assessment of the acceptance by health professionals interacting with machine learning applications will further contribute to the research of implementation in the field of predictive analytics.

CONCLUSION

Many published models predicting delirium have achieved high accuracy in retrospective datasets, but to our knowledge, none of these machine learning models has ever been implemented in clinical practice. This study showed that a machine learning based algorithm predicting the risk of delirium achieved the same discriminative performance during prospective prediction as that achieved in test scenarios. We revealed new insights in the implementation process of a machine learning application into a clinical workflow and were able to demonstrate a high agreement between the algorithm’s risk estimation and independent ratings by clinical experts. The final results of the evaluation and its clinical validation indicate that our algorithm is a reliable and accurate support for the delirium management in hospitals.

DATA AVAILABILITY

The data that support the findings of this study are available from KAGes (Steiermärkische Krankenanstaltengesellschaft m.b.H., Stiftingtalstraße 4, 8010 Graz, Austria) but restrictions apply to their availability, as they are not publicly available. Data are however available from the authors upon reasonable research proposals. In any case, permission of KAGes is required.

FUNDING

SJ was partly financed by a PhD project of the K1 COMET Competence Centre CBmed. CBmed is funded by the Federal Ministry of Transport, Innovation and Technology; the Federal Ministry of Science, Research and Economy; Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency; and the Vienna Business Agency. The COMET program is executed by the Austrian Research Promotion Agency (FFG). KAGes and SAP SE provided significant resources, manpower, and data as basis for research and innovation for this project.

AUTHOR CONTRIBUTIONS

SJ and DK developed the evaluation concept and guided the whole implementation process, including expert group meetings and trainings. SJ performed all analyses for evaluation and was the main contributor in writing the manuscript. DK developed the machine learning algorithm, the web application shown in Figure 4, and the technical solution for the implementation of the algorithm in the hospital information system. SJ and AA developed the protocol for clinical assessment, which was then completed in the clinical department by BG. BG and SR were involved in the expert group meetings and contributed on clinical discussions of the model results as clinical experts. AB, AA, WL, and SS critically reviewed the study design and results of the evaluation. All authors critically revised the manuscript and approved its final version.

ETHICS APPROVAL

The study received approval from the Ethics Committee of the Medical University of Graz (30-146 ex 17/18).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We gratefully acknowledge all employees at the participating departments at the hospital LKH Graz II who were involved in the implementation and evaluation process, especially those participating in the expert group meetings. We acknowledge Herbert Wurzer, Hubert Hauser, and Ewald Tax, who supported the implementation in the clinical departments. Special thanks go to Christian Jagsch, who was at our disposal for clinical questions during the development; to Elisabeth Lampl for her support in data assessment for evaluation; and to the team of Medical Informatics and Process Management of KAGes for their technical support. DK and SJ further acknowledge Günter Schreier, Dieter Hayn, Sai Veeranki, and Franz Quehenberger, who contributed to the development of the machine learning models with

their technical expertise. SJ acknowledges Michel Oleynik for his feedback on the manuscript, Harry Freitas Da Cruz for sharing his methodological knowledge on calibration, and the PhD school AMBRA at the Medical University of Graz.

CONFLICT OF INTEREST STATEMENT

SJ is currently employed by KAGes.

REFERENCES

1. Steyerberg EW, Moons KGM, van der Windt DA, *et al.* Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; 10 (2): e1001381.
2. Jiang F, Jiang Y, Zhi H, *et al.* Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2 (4): 230–43.
3. Goldstein BA, Navar AM, Pencina MJ, *et al.* Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; 24 (1): 198–208.
4. Weng SF, Reps J, Kai J, *et al.* Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017; 12 (4): e0174944.
5. Kourou K, Exarchos TP, Exarchos KP, *et al.* Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015; 13: 8–17.
6. Meyer A, Zverinski D, Pfahringer B, *et al.* Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med* 2018; 6 (12): 905–14.
7. Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; 1: 18.
8. Lee CH, Yoon H-J. Medical big data: promise and challenges. *Kidney Res Clin Pract* 2017; 36 (1): 3–11.
9. Amarasingham R, Patzer RE, Huesch M, *et al.* Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff (Millwood)* 2014; 33 (7): 1148–54.
10. Liberati EG, Ruggiero F, Galuppo L, *et al.* What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. *Implement Sci* 2017; 12: 113. doi : 10.1186/s13012-017-0644-2.
11. Islam M, Hasan M, Wang X, *et al.* A systematic review on healthcare analytics: application and theoretical perspective of data mining. *Healthcare (Basel)* 2018; 6 (2): 54.
12. Bates DW, Saria S, Ohno-Machado L, *et al.* Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)* 2014; 33 (7): 1123–31.
13. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet* 2014; 383 (9920): 911–22.
14. Hsieh TT, Yue J, Oh E, *et al.* Effectiveness of multicomponent nonpharmacological delirium interventions: a meta-analysis. *JAMA Intern Med* 2015; 175 (4): 512.
15. Inouye SK, Bogardus ST, JrCharpentier PA, *et al.* A multicomponent intervention to prevent delirium in hospitalized older patients. *N Engl J Med* 1999; 340 (9): 669–76.
16. Schuurmans M, Shortridge-Baggett L, Duursma S. The Delirium Observation Screening Scale: a screening instrument for delirium. *Res Theory Nurs Pract* 2003; 17 (1): 31–50.
17. Inouye SK, Christopher M, Dyck H, *et al.* Clarifying confusion: the confusion assessment method. A new method for detection of delirium. *Ann Intern Med* 1990; 113 (12): 941–8.
18. Kramer D, Veeranki S, Hayn D, *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.
19. van Meenen LCC, van Meenen DMP, de Rooij SE, *et al.* Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *J Am Geriatr Soc* 2014; 62 (12): 2383–90.

20. Lee A, Mu JL, Joynt GM, *et al.* Risk prediction models for delirium in the intensive care unit after cardiac surgery: a systematic review and independent external validation. *Br J Anaesth* 2017; 118 (3): 391–9.
21. Lindroth H, Bratzke L, Purvis S, *et al.* Systematic review of prediction models for delirium in the older adult inpatient. *BMJ Open* 2018; 8 (4): e019223.
22. Newman MW, O'Dwyer LC, Rosenthal L. Predicting delirium: a review of risk-stratification models. *Gen Hosp Psychiatry* 2015; 37 (5): 408–13.
23. 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. doi : 10.1007/s10916-018-1109-0.
24. 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.
25. Kim MY, Park UJ, Kim HT, Cho WH. DELirium prediction based on hospital information (Delphi) in general surgery patients. *Medicine (Baltimore)* 2016; 95 (12): e3072.
26. Talmon J, Ammenwerth E, Brender J, Dekeizer N, Nykanen P, Rigby M. STARE-HI -statement on reporting of evaluation studies in health informatics. *Int J Med Inform* 2009; 78 (1): 1–9.
27. Watt J, Tricco AC, Talbot-Hamon C, *et al.* Identifying older adults at risk of delirium following elective surgery: a systematic review and meta-analysis. *J Gen Intern Med* 2018; 33 (4): 500–9.
28. Veeranki S, Hayn D, Eggerth A, *et al.* On the representation of machine learning results for delirium prediction in a hospital information system in routine care. *Stud Health Technol Inform* 2018; 251: 97–100.
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics* 1988; 44 (3): 837–45.
30. Sabetta L. Self-defeating prophecies: when sociology really matter. In: Poli R, Valerio M, eds. *Anticipation, Agency and Complexity*. Cham, Switzerland: Springer; 2019: 51–9.
31. DELIRIUM: Diagnosis, Prevention and Management. London, United Kingdom: National Clinical Guideline Centre; 2010.
32. Siddiqi N, House AO, Holmes JD. Occurrence and outcome of delirium in medical in-patients: a systematic literature review. *Age Ageing* 2006; 35 (4): 350–64.
33. Katznelson R, Djaiani G, Tait G, *et al.* Hospital administrative database underestimates delirium rate after cardiac surgery. *Can J Anaesth* 2010; 57 (10): 898–902.
34. Jauk S, Kramer D, Quehenberger F, *et al.* Information adapted machine learning models for prediction in clinical workflow. *Stud Health Technol Inform* 2019; 260: 65–72.
35. Perry WM, Hossain R, Taylor RA. Assessment of the Feasibility of automated, real-time clinical decision support in the emergency department using electronic health record data. *BMC Emerg Med* 2018; 18 (1): 19. doi : 10.1186/s12873-018-0170-9.