dHealth 2022
G. Schreier et al. (Eds.)

doi:10.3233/SHTI220353

External Validation of a Machine Learning Based Delirium Prediction Software in Clinical Routine

Stefanie JAUK[a,1], Sai Pavan Kumar VEERANKI[a], Diether KRAMER[a], Stefan HÖGLER[b], David MÜHLECKER[b], Erwin EBERHARTL[b], Arthur SCHUELER[b], Christian CHVOSTA[c], Wolfgang STRASSER[c], Reinhold STRASSER[b] and Werner LEODOLTER[a]
[a] *Steiermärkische Krankenanstaltengesellschaft m.b.H, Graz, Austria*
[b] *Krankenhaus der Barmherzigen Schwestern Ried, Ried im Innkreis, Austria*
[c] *Vinzenz Gruppe Krankenhausbeteiligungs- und Management GmbH, Wien, Austria*

Abstract. Background: Various machine learning (ML) models have been developed for the prediction of clinical outcomes, but there is missing evidence on their performance in clinical routine and external validation. Objectives: Our aim was to deploy and prospectively evaluate an already developed delirium prediction software in clinical routine of an external hospital. Methods: We compared updated ML models of the software and models re-trained with the external hospital's data. The best models were deployed in clinical routine for one month, and risk predictions for all admitted patients were compared to the risk ratings of a senior physician. After using the software, clinicians completed a questionnaire assessing technology acceptance. Results: Re-trained models achieved a high discriminative performance (AUROC > 0.92). Compared to clinical risk ratings, the software achieved a sensitivity of 100.0% and a specificity of 90.6%. Usefulness, ease of use and output quality were rated positively by the users. Conclusion: A ML based delirium prediction software achieved a high discriminative performance and high technology acceptance at an external hospital using re-trained ML models.

Keywords. Machine learning, clinical prediction models, electronic health records, external validation, clinical decision support, delirium.

1. Introduction

The occurrence of delirium in hospitalized patients presents a great burden to patients, their families, clinicians and hospital providers [1,2]. Delirium is a syndrome of an acute confusional state together with an acute decline in attention and cognition, and is common in elderly patients [3]. Delirium patients have longer hospital stays [4], an increased risk for long-term care [5] and increased mortality rates [6,7].

For medical inpatients, the incidence of delirium ranges from 3% to 29% [8]. However, there is evidence that delirium can be prevented, especially when using non-pharmacological actions [9,10]. Therefore, patients at high risk of delirium need to be identified as early as possible during their hospital stay.

[1] Corresponding Author: Stefanie Jauk, Steiermärkische Krankenanstaltengesellschaft m.b.H, Graz, Austria, E-Mail: stefanie.jauk@kages.at

Common screening instruments for delirium include the Confusion Assessment Method (CAM) [11] or the Delirium Observation Screening (DOS) scale [12]. However, CAM is considered a diagnostic instrument and DOS is based on diagnostic criteria to assess early symptoms of delirium [13].Thus, such assessments may only detect patients with already existing signs and symptoms. In addition, manual assessments require additional resources, which are very limited in hospitals.

Over the last years, machine learning (ML) methods have become popular for the training of risk prediction models using electronic health record (EHR) data [14]. A major advantage of this approach is that no additional data needs to be collected, which reduces the effort of risk assessment for healthcare professionals substantially. Furthermore, the automatic and fast prediction of patients' risks allows for hospital-wide risk stratification with the need of little resources.

Although various ML models have been developed for the prediction of delirium using EHR data [15,16], there is missing evidence on the performance of the models in clinical routine [17]. Above all, if the distribution of the training data does not match the distribution for deployment, the performance of a model is expected to be lower [18]. Hence, research on the external validation of ML models in clinical routine is strongly needed [19,20].

In 2018, we first implemented an ML based prediction tool for delirium in a hospital of Steiermärkische Krankenanstaltengesellschaft m.b.H. (KAGes) in Graz, Austria [21]. The algorithm of the software uses random forest classification models, which have been developed since 2016 by KAGes, to stratify patients according to their risk of delirium. Within few seconds after admission, the software automatically predicts the risk of delirium based on EHR data such as previous diagnoses, procedures, or laboratory values.

During a seven-months prospective evaluation in clinical routine, the software achieved a sensitivity of 74.1% and a specificity of 82.2% [21]. Furthermore, the evaluation of the user acceptance in the KAGes hospital demonstrated that the software was easy to use and that healthcare professionals perceived it as useful [22].

1.1. Aim

The delirium prediction software achieved a high discriminative performance in clinical routine of KAGes hospital wards, but external validation outside of the KAGes network is needed in order to determine the generalizability of the software. Therefore, the aim of this study is to deploy the delirium prediction software of KAGes in a hospital of Vinzenz Gruppe (VG), another healthcare provider in Austria, and to evaluate its performance and technology acceptance in this external setting.

2. Methods

2.1. Update of ML models to external EHR data

The algorithm of the software uses three random forest models for delirium prediction at different prediction times during the hospital stay: at admission, the evening after admission, and the second evening after admission. The original models of the software were trained on EHR data from KAGes exclusively. In order to determine the generalizability of the KAGes models, we first externally validated the original KAGes models on test data hosted by VG.

VG and KAGes host EHR data of around 2.3 million patients. Although both hospital providers use the same software for the hospital information system (HIS), EHR documentation and patient populations differ between them. Therefore, the original random forest models needed to be updated based on a selection of features. All model features which were available in the EHR system of VG were determined and selected for training. Out of 1,033 features of the original models, the selection resulted in 716 features available for VG data. This included demographic data, ICD-10 coded diagnoses, transfer data, procedures, laboratory data, nursing data and medication.

Using the selected features, the random forest models were trained again using KAGes data. The training data was the same as for the original models, including 22,110 admissions to KAGes hospitals between 2011 and 2019. Random forest models were trained in R using the *caret* package [23] with 5-fold cross-validation and down-sampling in order to account for imbalance in the outcome variable. The trained models were evaluated on the unseen test data of 5,347 admissions from VG, including 347 cases of delirium.

2.2. Re-training of ML models with external EHR data

Subsequently, a re-training of the models using VG data only was conducted. New random forest models were trained from scratch on VG data using the same methods as for the updated models with down-sampling and 5-fold cross-validation, and the same 716 features. The models were trained on VG data of 16,042 admissions and tested on the same test data as the updated models (n = 5,347).

These re-trained models were then compared to the updated models. As measures of discrimination, Receiver-Operating Characteristic (ROC) curves and Area Under the Receiver-Operating Characteristic (AUROC) with DeLong confidence intervals [24] were used. In addition, calibration plots were conducted to assess calibration.

2.3. Integration of the delirium prediction software in the HIS

After choosing the best performing models, the software was integrated in the HIS of VG. The process was identical to the implementation process in KAGes hospitals [21].

Delirium prediction was conducted for every patient admitted or transferred to the participating department. At time of admission, an HL7 message was sent from the HIS to a communication server, monitored by a Linux server that retrieves EHR data for the patient from the HIS using http-requests. The prediction was performed in an R environment, and the results were sent back to the communication server.

The delirium risk was visualized in the user interface of the HIS using a red symbol for *very high risk* patients, a yellow symbol for *high risk* patients, and no symbol for *low risk* patients. With a click on this symbol, a web application opened up in the local browser, providing details of the ML based risk prediction. The application presented modelling features ranked by their importance for the prediction and patient specific values for all features groups. This visualization aimed to support healthcare professionals in decision-making and verifying the prediction results of the software [21].

2.4. Study design

The prospective evaluation of the delirium prediction software focused on (a) the discriminative performance and calibration when deployed in clinical routine and (b) the user acceptance reported by healthcare professionals.

After integration in the HIS of VG, the software was made visible to healthcare professionals of the trauma surgery department in the hospital Barmherzige Schwestern Ried. For all patients admitted or transferred to the department between 16th of June, 2021, and 16th of July, 2021, the software predicted the risk of delirium at admission time, the evening of admission, and the evening of the second day of admission.

During the entire pilot phase, a senior physician observed all patients of the department and stratified the patients according to their delirium risk into *low risk* and *high risk* patients. This clinical risk estimation was used as a gold standard to evaluate the risk predictions of the software. In order to calculate metrics of the confusion matrix and to compare the results with the physician's estimation, patients predicted a *high risk* or a *very high risk* were combined in one risk group for analyses. As the senior physician provided his estimation within the first 48 hours of admission, we compared it to the latest prediction result within this time period (prediction time *2nd evening*).

To assess user acceptance, a previously developed questionnaire was used [22]. It is based on the Technology Acceptance Model (TAM) and its extension TAM2 [25,26], and assesses perceived ease of use, perceived usefulness, actual system use and output quality. Reliability of the TAM model is generally high with Cronbach's alpha above 0.8 [27]. For the developed questionnaire alpha was ranging from 0.66 to 0.85 [22].

Responses on 15 items were measured using a five point Likert-type response scale (strongly agree to strongly disagree; very frequently to very rarely). One item assessed the absolute frequency of use per month in numbers. In addition, user comments were collected in a free-text field.

The user acceptance questionnaire was distributed to all healthcare professionals of the department, including senior physicians, junior physicians, nurses and care assistants. Results were analyzed using heat maps. Two items measuring perceived ease of use had been formulated negatively and were recoded for analysis.

The study received approval from the Ethics Committee of the Medical University of Graz (30–146 ex 17/18) and from the Ethics Committee of the Faculty of Medicine of the Johannes Kepler University in Linz (1297/2020).

3. Results

3.1. Model validation on test data

Table 1. Discriminative performance of the random forest models trained on KAGes data (updated models) and on VG data (re-trained models), when tested on the same VG test data.

Version	Training	Test	Prediction time	AUROC	95%-CI	Sensitivity[a]	Specificity[a]
Updated model	KAGes	VG	Admission	0.863	0.8475-0.8784	75.9%	85.3%
			1st evening	0.851	0.8348-0.8664	74.1%	85.2%
			2nd evening	0.857	0.8412-0.8721	74.1%	85.3%
Re-trained model	VG	VG	Admission	0.924	0.9123-0.9364	82.8%	85.3%
			1st evening	0.926	0.9140-0.9386	81.0%	85.3%
			2nd evening	0.931	0.9202-0.9427	86.2%	85.5%

Note: [a]Values were calculated on the test data, setting the top 15% of highest risk patients as *very high* and *high risk* patients.

Table 1 presents the AUROC values and Figure 1 the corresponding ROC curves including 95%-CI for the three updated models and re-trained models. The AUROC values for the updated models were above 0.85 for all prediction times. For the re-trained models, the AUROC was even higher with values above 0.92. Calibration plots of all six models indicated an overestimation of delirium risk. Because of the superior discriminative performance, the re-trained models were selected for deployment.

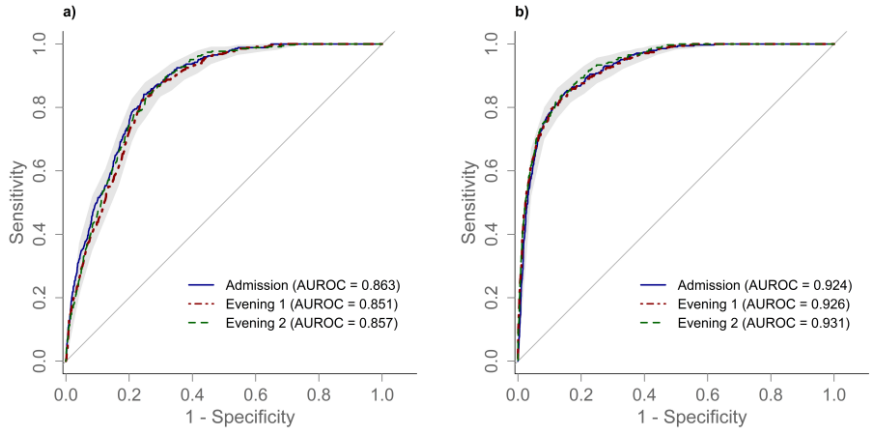

Figure 1. ROC curve including 95%-CI for the updated models trained on KAGes data (a) and for the re-trained models trained on VG data (b), both tested on the VG test data (n = 5,347).

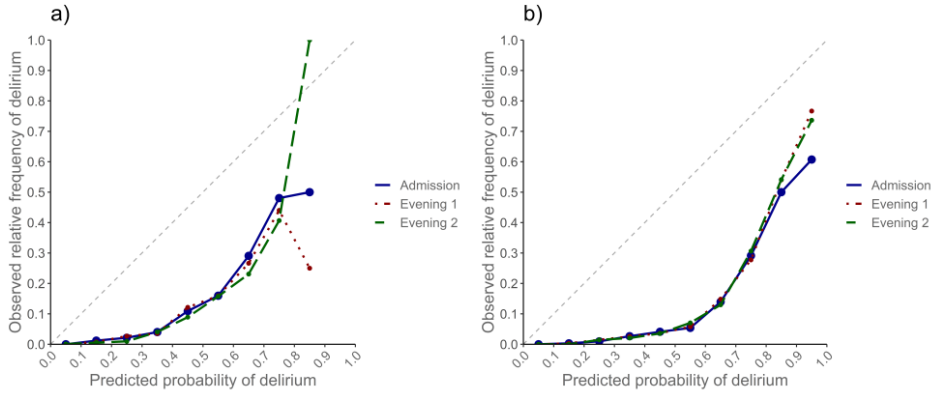

Figure 2. Calibration plots of the updated models trained on KAGes data (a) and of the re-trained models trained on VG data (b), evaluated on the VG test data (n = 5,347) including 347 cases of delirium.

3.2. Model performance in clinical routine

During the one-month evaluation between 16th of June 2021 and 16th of July 2021, 93 patients were admitted to the trauma surgery department. Their median age was 60 years (min = 18; max = 95), and median length of stay 3 days (min = 0; max = 24).

Table 2 presents the confusion matrix comparing the clinical estimation of delirium risk and the prediction of the software. The senior physician estimated a high risk of delirium for 29 of 93 patients (31.2%). All 29 patients were predicted a high risk or very

high risk by the delirium prediction software, leading to a sensitivity of 100.0%. With six false positive cases, the specificity was 90.6%.

Table 2. Confusion matrix comparing the prediction of the delirium prediction software with the clinical estimation by a senior physician. Values are presented as absolute frequencies and row percentages.

		Delirium Prediction Software					
		Low risk		High/Very high risk		Total	
		n	%	n	%	n	%
Clinical estimation	**Low risk**	58	90.6	6	9.4	64	100.0
	High risk	0	0.0	29	100.0	29	100.0
	Total	58	62.4	35	37.6	93	100.0

3.3. User acceptance

One senior physician, one junior physician, nine nurses and two care assistants of the department completed the questionnaire (n = 13). Thus, the response rate was 100%.

Figure 3 presents their answers on the questionnaire items. For the factor *perceived usefulness*, all 13 users (100.0%) agreed or strongly agreed that the software is useful for their work and a useful support for delirium and to detect delirium at an early stage.

For *perceived ease of use*, the purpose was clear and understandable for all users (100.0%). The majority (strongly) agreed that information was understandable (10 users, 76.9%), that they successfully integrated the software into their clinical workflow (11 users, 84.6%) and that it did not increase their workload (9 users, 69.2%).

Regarding *actual system use*, all 13 users (100.0%) (strongly) agreed that they used the application regularly. 7 users (53.8%) reported that they considered the output of the application in their clinical decisions.

Regarding *output quality,* 12 users (92.3%) reported that the calculated delirium risk matched their own estimation; 6 users (46.2%) sometimes estimated the risk higher.

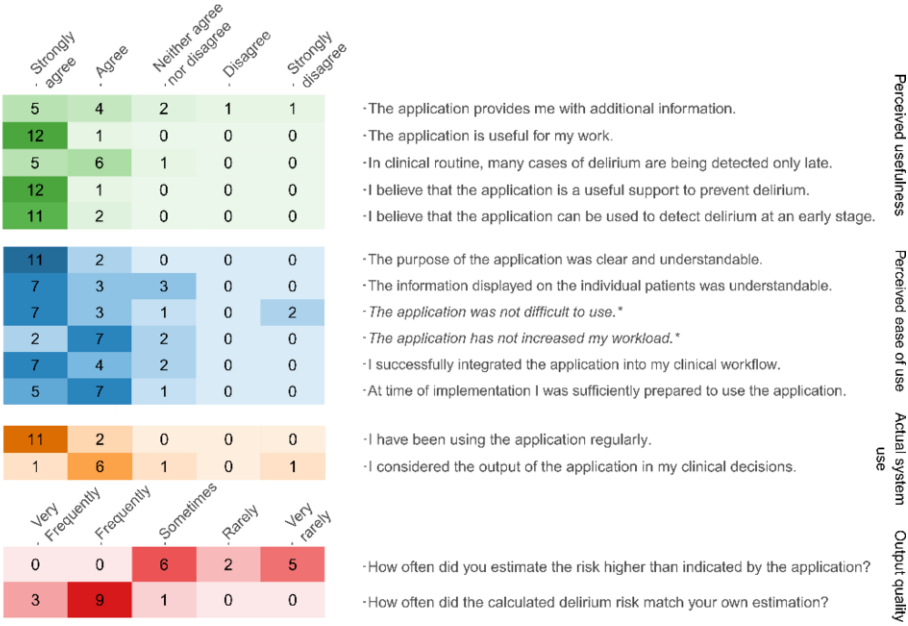

Figure 3. Heat map of healthcare professionals' answers (n = 13) on the items of the technology assessment questionnaire developed in [22]. The two items marked with '*' were recoded for analysis.

Four users left comments in the free-text field. Two appreciated the work of the senior physician in charge of the delirium risk assessment. One user commented that the software is a great support in daily routine, and the other user commented that the use of such solutions is a duty for every modern hospital.

4. Discussion

In this study, we externally validated the performance of an ML based delirium prediction software in clinical routine. For this purpose, we deployed an already developed and evaluated software in a trauma surgery department of another hospital provider in Austria.

During the one-month evaluation period at the external hospital, the delirium prediction software achieved a sensitivity of 100% and a specificity of 90.6% compared to the risk estimation of a senior physician. Healthcare professionals using the software in clinical routine rated the overall ease of use, usefulness and output quality positively.

The random forest models trained with KAGes data achieved high AUROC values (above 0.85) when externally validated on the VG test data. This result demonstrates that models trained on KAGes data generalize well and can be used for prediction at other hospital providers with high discriminative performance. However, for future deployments, model re-training using data of external hospital providers might be preferred if the amount of EHR data hosted by the provider allows for it and if resources are provided. As expected, the re-trained models reached even a higher performance for VG data (AUROC values above 0.92) and were thus preferred for integration in the software and for prediction in clinical routine.

A main limitation of this study is that the model performance in clinical routine was assessed using the clinical estimation of delirium risk by one senior physician only. Although the physician was a very critical rater, subjective estimations by single physicians are prone to be biased. For healthcare professionals in the department it was essential to perform preventive actions for patients at risk, and delirium could have been prevented in some cases. Therefore, it was not feasible to use delirium diagnoses to determine the performance of the software. In future, the performance of the software at VG needs to be surveilled using more data sources and longer observation times.

Another limitation regards the comparison of clinical estimations with the software. While the physician documented his clinical estimation using two risk groups (*low risk* and *high risk*), the software stratifies patients into three risk groups (*low risk*, *high risk*, *very high risk)*. The three groups support a further differentiation for the users between *high risk* and *very high risk* patients, but for analyses these two groups were combined.

Future work should validate the software in an international setting. An external validation in other countries with different cultural backgrounds can help to detect potential biases of the algorithm. In addition, a validation in other HIS will further increase the insights on the generalizability of the software.

Finally, more research is needed to determine how model performances in external settings can be increased. One opportunity for future developments is federate learning (FL). FL provides the opportunity to train models with data from various partners without the need of centralized data sharing [28]. Methods of FL are based on the idea to train models at each partner and share the pre-trained models instead of sharing data. In future, such methods could help to develop more generalizable risk prediction models and thus make it easier to distribute successful decision support tools among hospital providers.

References

[1] D.L. Leslie, and S.K. Inouye, The Importance of Delirium: Economic and Societal Costs, *J Am Geriatr Soc*. **59** (2011) S241–S243. doi:10.1111/j.1532-5415.2011.03671.x.

[2] E.M. Schmitt, J. Gallagher, A. Albuquerque, P. Tabloski, et al., Perspectives on the Delirium Experience and Its Burden: Common Themes Among Older Patients, Their Family Caregivers, and Nurses, *The Gerontologist*. **59** (2019) 327–337. doi:10.1093/geront/gnx153.

[3] S.K. Inouye, Delirium in Older Persons, *NEJM*. **354** (2006) 1157–1165. doi:10.1056/NEJMra052321.

[4] J. McCusker, M.G. Cole, N. Dendukuri, and E. Belzile, Does Delirium Increase Hospital Stay?, *J Am Geriatr Soc*. **51** (2003) 1539–1546. doi:10.1046/j.1532-5415.2003.51509.x.

[5] H. Bickel, R. Gradinger, E. Kochs, and H. Förstl, High Risk of Cognitive and Functional Decline after Postoperative Delirium, *Dement Geriatr Cogn Disord*. **26** (2008) 26–31. doi:10.1159/000140804.

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

[7] S.-M. Lin, C.-Y. Liu, C.-H. Wang, H.-C. Lin, et al., The impact of delirium on the survival of mechanically ventilated patients*:, *Critical Care Medicine*. **32** (2004) 2254–2259. doi:10.1097/01.CCM.0000145587.16421.BB.

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

[9] N. Siddiqi, J.K. Harrison, A. Clegg, E.A. Teale, et al., Interventions for preventing delirium in hospitalised non-ICU patients, *Cochrane Database of Systematic Reviews*. (2016). doi:10.1002/14651858.CD005563.pub3.

[10] T.T. Hshieh, J. Yue, E. Oh, M. Puelle, et al., Effectiveness of Multicomponent Nonpharmacological Delirium Interventions: A Meta-analysis, *JAMA Internal Medicine*. **175** (2015) 512. doi:10.1001/jamainternmed.2014.7779.

[11] S.K. Inouye, M. Christopher, H. Dyck, M. Cathy, et al., Clarifying confusion: the confusion assessment method. A new method for detection of delirium., *Annals of Internal Medicine*. **113** (1990) 941–948.

[12] M. Schuurmans, L. Shortridge-Baggett, and S. Duursma, The Delirium Observation Screening Scale: a screening instrument for delirium, *Res Theory Nurs Pract*. **17** (2003) 31–50.

[13] S. Grover, and N. Kate, Assessment scales for delirium: A review, *World J Psychiatry*. **2** (2012) 58–70. doi:10.5498/wjp.v2.i4.58.

[14] B.A. Goldstein, A.M. Navar, M.J. Pencina, and J.P.A. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, *J Am Med Inform Assoc*. **24** (2017) 198–208. doi:10.1093/jamia/ocw042.

[15] J.P. Corradi, S. Thompson, J.F. Mather, C.M. Waszynski, et al., Prediction of Incident Delirium Using a Random Forest classifier, *Journal of Medical Systems*. **42** (2018). doi:10.1007/s10916-018-1109-0.

[16] A. Wong, A.T. Young, A.S. Liang, R. Gonzales, et al., Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment, *JAMA Network Open*. **1** (2018) e181018.

[17] M.W. Newman, L.C. O'Dwyer, and L. Rosenthal, Predicting delirium: a review of risk-stratification models, *General Hospital Psychiatry*. **37** (2015) 408–413. doi:10.1016/j.genhosppsych.2015.05.003.

[18] J.P. Cohen, T. Cao, J.D. Viviano, C.-W. Huang, et al., Problems in the deployment of machine-learned models in health care, *CMAJ*. **193** (2021) E1391–E1394. doi:10.1503/cmaj.202066.

[19] T. Hernandez-Boussard, S. Bozkurt, J.P.A. Ioannidis, and N.H. Shah, MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care., *J Am Med Inform Assoc*. **27** (2020) 2011–2015. doi:10.1093/jamia/ocaa088.

[20] T. Antoniou, and M. Mamdani, Evaluation of machine learning solutions in medicine, *CMAJ*. **193** (2021) E1425–E1429. doi:10.1503/cmaj.210036.

[21] S. Jauk, D. Kramer, B. Großauer, S. Rienmüller, et al., Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study, *J Am Med Inform Assoc*. **27** (2020) 1383–1392. doi:10.1093/jamia/ocaa113.

[22] S. Jauk, D. Kramer, A. Avian, A. Berghold, et al., Technology Acceptance of a Machine Learning Algorithm Predicting Delirium in a Clinical Setting: a Mixed-Methods Study, *J Med Syst*. **45** (2021). doi:10.1007/s10916-021-01727-6.

[23] M. Kuhn, caret: Classification and Regression Training. R package version 6.0-78., 2017.

[24] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, *Biometrics*. **44** (1988) 837. doi:10.2307/2531595.

[25] F.D. Davis, Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology, *MIS Quarterly*. **13** (1989) 319. doi:10.2307/249008.

[26] V. Venkatesh, and F.D. Davis, A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies, *Management Science*. **46** (2000) 186–204.

[27] W.R. King, and J. He, A meta-analysis of the technology acceptance model, *Information & Management*. **43** (2006) 740–755. doi:10.1016/j.im.2006.05.003.

[28] N. Rieke, J. Hancox, W. Li, F. Milletarì, et al., The future of digital health with federated learning, *Npj Digit. Med.* **3** (2020) 119. doi:10.1038/s41746-020-00323-1.