EJAIC

Eur J Anaesthesiol Intensive Care Med 2022; **2:**1(e0010)

OPEN

ORIGINAL ARTICLE

Leveraging artificial intelligence for the management of postoperative delirium following cardiac surgery

Janis Fliegenschmidt, Nikolai Hulde, Maria Gedinha Preising, Silvia Ruggeri, Ralph Szymanowsky, Laurent Meesseman, Hong Sun, Michael Dahlweid and Vera von Dossow

BACKGROUND Postoperative delirium is a highly relevant complication of cardiac surgery. It is associated with worse outcomes and considerably increased costs of care. A novel approach of monitoring patients with machine learning enabled prediction software could trigger pre-emptive implementation of mitigation strategies as well as timely intervention.

OBJECTIVE This study evaluates the predictive accuracy of an artificial intelligence (AI) model for anticipating postoperative delirium by comparing it to established standards and measures of risk and vulnerability.

DESIGN Retrospective predictive accuracy study.

SETTING Records were gathered from a database for anaesthesia quality assurance at a specialised heart surgery centre in Germany.

PATIENTS Between January and July 2021, 131 patients had been enrolled into the database and had data available for AI prediction modelling. After exclusion of incomplete follow-ups, a subset of 114 was included in the statistical analysis.

MAIN OUTCOME MEASURES Delirium was diagnosed with the Confusion Assessment Method for the ICU (CAM-ICU) over three days postoperatively with specific follow-up visits. AI predictions were also compared with risk assessment through a frailty screening, a Shulman Clock Drawing Test, and using a checklist of predisposing factors including comorbidity, reduced mobility, and substance abuse.

RESULTS Postoperative delirium was diagnosed in 23.7% of patients. Postoperative AI screening exhibited reasonable performance with an area under the receiver operating curve (AUROC) of 0.79, 95% confidence interval (CI), 0.69−0.87. But pre-operative prediction was weak for all methods (AUROC range from 0.55 to 0.66). There were significant associations with postoperative delirium: open heart surgery versus endovascular valve replacement (33.3% *vs.* 10.4%, $P < 0.01$), postinterventional hospitalisation (12.8 *vs.* 8.6 days, $P < 0.01$), and length of ICU stay (1.7 *vs.* 0.3 days, $P < 0.01$) were all significantly associated with postoperative delirium.

CONCLUSION AI is a promising approach with considerable potential and delivered noninferior results compared with the usual approach of structured evaluation of risk factors and questionnaires. Since these established methods do not provide the desired confidence level, improved AI may soon deliver a better performance.

TRIAL REGISTRATION None.

Published online 8 December 2022

From the Department of Anesthesiology and Pain Therapy, HDZ NRW, University Hospital of Ruhr-University Bochum, Bad Oeynhausen, Germany (JF, NH, MGP, SR, VvD) and Dedalus Healthcare, Roderveldlaan 2, 2600 Antwerp, Belgium (RS, LM, HS, MD)

Correspondence to Janis Fliegenschmidt, Department of Anaesthesiology and Pain Therapy, Heart and Diabetes Centre HDZ NRW, Georgstraße 11, 32545 Bad Oeynhausen, NRW, Germany
Tel: +49 5731 97 1128; fax: +49 5731 97 2196; e-mail: janis.fliegenschmidt@ruhr-uni-bochum.de

DOI:10.1097/EA9.0000000000000010

KEY POINTS

- Postoperative delirium is an under-recognised entity with detrimental impact on affected patients
- Artificial intelligence (AI) can utilise a patient's electronic record to provide a low-effort screening mechanism
- Postoperative screening with an AI-based approach was helpful in identifying at-risk patients
- At the moment, both the traditional approaches to pre-operative risk stratification and AI-based were at best 'marginal' in diagnosing postoperative delirium.
- More comprehensive data availability will continue to improve AI-based solutions in the field of peri-operative care

Introduction

Postoperative delirium (POD) is an important complication of cardiac surgery. Research suggests that despite continuing academic interest, delirium remains an under-recognised entity in clinical practice.[1] This is despite the implications that POD has for the clinical course of surgical patients. POD increases the risk for an extended period of intensive care, leads to longer hospitalisation, increased mortality and decreased cognitive outcome. Cognitive deficits can persist past rehabilitative measures.[2] Consequently, POD presents a clear danger, especially to vulnerable patients undergoing major surgical interventions. In addition to the detrimental effects on an individual patient's short- and long-term outcomes, a high incidence of POD also generates a significant economic burden on the healthcare system, as increased care requires utilisation of scarce resources.[3] Recognising delirium and its prodromes enables healthcare providers to implement safeguarding measures to prevent the development of manifest delirium[4] and to accurately recognise patients who would profit from specialised care. Special emphasis should be placed on the recognition of hypoactive delirium,[5] as it can be considered to be mostly responsible for the large discrepancy in the incidence of delirium reported in studies which utilise active screening *versus* studies relying on coding data.

In cardiac surgery especially, POD is a major concern.[6] Patients are often at high-risk from multimorbidity and are exposed to extensive surgical interventions with long anaesthesia, ventilation, and sedation periods. In patients receiving cardiac surgery, POD has been shown to be an independent risk factor for neurological decline as well as increased 30-day and 1-year-mortality.[7] Potter *et al.* found that POD led to an increased financial burden of 15 592 USD per cardiac surgery patient on average.[8]

A screening approach requiring very little effort by healthcare providers could substantially increase the provision of delirium screening, with a potential to improve outcome.[9] An artificial intelligence (AI) programme relying on data from a hospital's electronic health record (EHR) could provide such a highly accessible screening instrument, with benefit for patients and providers alike. Previous studies report high accuracies for models trained and validated retrospectively within the same dataset and for general populations.[10,11] A prospectively evaluated AI model also exhibited high performance, but utilised additional rules and provided predictions at admission only.[12]

The AI model in this study was evaluated on data from a separate, actively screened cohort, with no additional rules and continuous monitoring. In a previously published case report, it successfully identified a patient who was at risk of developing POD following implantation of a left ventricular assist device (LVAD).[13] This work builds on the findings from the index case by analysing the prognostic accuracy of the AI model on a cohort of cardiac surgery patients from a high-volume cardiac surgery center.

Methods

Ethics

Ethical approval for this study (Az.2021-861) was provided by the Ethics Committee of the Medical Faculty, Ruhr-University Bochum, Division OWL (PO Box 10 03 61, 32503 Bad Oeynhausen), Chairperson Prof. Dr med. Wolfgang Burchert, on 29 October 2021.

Data acquisition

The patients in this study were retrospectively recruited from a quality assurance program database.* The program provides frailty and delirium screening for patients of advanced age undergoing cardiac surgery or transcatheter cardiac interventions at the Heart and Diabetes Centre (HDZ) NRW, Germany; a specialised, high-volume cardiac surgery centre. The programme enrols a representative subset of patients of 65 years of age and over, selected based on membership with the AOK (Allgemeine Ortskrankenkasse, a statutory health insurance provider). After enrolment, patients receive extensive pre-operative screening. It includes measures of frailty[14] (Table 1) such as hand strength and a simple mobility assessment, patient history including psychiatric comorbidities and substance use, cognitive assessments including completion of a preoperative CAM-ICU questionnaire and the Shulman Clock Drawing Test, as well as EEG based delirium scoring from a single channel recording (Fp2-Pz, delta-scan).[15] Additionally, patients undergo risk stratification by checklist according to the hospital's standard operating procedure for delirium

* Quality assurance contract between the hospital and the insurance provider on the prevention of postoperative delirium in elderly patients, monitored by the German Institute for Quality Assurance and Transparency (IQTIG).

Table 1 Preoperative frailty assessment

Age >70 years
>3 Comorbidities: *Arterial hypertension, chronic heart failure, diabetes, atrial fibrillation, stroke, peripheral artery disease, chronic kidney injury*
Neurodegenerative disease
Chronic pain
Impaired vision or hearing
Reduced mobility or immobility
Substance abuse (including alcohol)
Timed-up-and-go test <20 s
Manual strength below age-adjusted threshold on either side
Involuntary weight loss
Reduced general activity
PHQ-2 depression screening

Preoperative frailty assessment. Zero points indicate nonfrail patients, 1 – 2 points indicate prefrail patients, greater than two points indicate frail patients.

Table 2 Risk factors for postoperative delirium

Age >70 years
>3 Comorbidities: *Arterial hypertension, chronic heart failure, diabetes, atrial fibrillation, stroke, peripheral artery disease, chronic kidney injury*
Neurodegenerative disease
Chronic pain
Impaired vision or hearing
Reduced mobility or immobility
Substance abuse (including alcohol)

Checklist for preoperative assessment of patient's risk for postoperative delirium. Patients are considered to be at risk when more than three risk factors are present.

prevention (Table 2). After surgery, patients are followed up for three consecutive days and assessed with the CAM-ICU, the Shulman Clock Drawing Test, and they are asked to report their pain levels and whether they experienced nausea and vomiting. The EEG screening is repeated for all three postoperative assessments as well. The CAM-ICU measurements were used as the reference test for delirium.

Statistical analysis was performed with Python 3.8.8[16], pandas 1.3.5[17], scipy 1.7.3[18] and scikit-learn 1.0.2[19]. Graphs were plotted with matplotlib 3.5.0[20].

Patient collective

From the quality assurance database, all patients whose data were available to compute predictions (see section 2.2) were included from 1 January 2021 to 30 July 2021. Inclusion criteria for enlistment in the quality assurance program were patient age greater than 64 years, planned surgical or endovascular intervention and enrolment with the participating insurance provider. Interventions included surgical valve replacements (aortic valve and mitral valve), off-pump coronary bypass surgery, transcatheter aortic valve replacement and tricuspid valve clipping. Emergency surgery patients were not included. A total of 131 patients were identified in the study period. Patients were excluded if predictions from the day before to three days after surgery were missing. Reasons for this were peri-operative mortality ($n=2$), early discharge ($n=4$), and technical problems pertaining to the data export ($n=10$). One patient was excluded after they rescinded their consent ($n=1$). The remaining 114 cases formed the basis for the statistical analysis in this study. For some cases, data other than AI predictions were missing. These patients were not excluded. The numerical basis for all calculations is denoted in the respective tables.

All patients were subject to the hospital's standard delirium prevention measures. This included benzodiazepine-free premedication, only two hours of drinking restrictions for clear fluids preoperatively, the use of dexmedetomidine intraoperatively as well as intraoperative depth of anesthesia monitoring. Postoperatively, patients were regularly assessed by nursing staff. Adherence to the protocol was not monitored.

Artificial intelligence predictions

Artificial intelligence predictions were computed by a tool named clinalytix.[†] [21] Using clinalytix, a delirium risk prediction model was trained on retrospective EHR entries from HDZ encompassing 6456 cases between 2009 and 2020. The model was trained as a binary classifier predicting an ICD delirium label in the patient's chart at discharge (80:20 train-test-split).[22] The clinalytix tool is integrated with the ORBIS EHR system, the delirium risk prediction model trained by clinalytix is deployed as a prediction service. The model training process uses the Transformer Deep Learning model to train a binary classification model for delirium risk prediction.[23] Features such as lab results, procedures, demographic information, etc. are used as inputs of the model training process. Clinical entities are extracted from free-text clinical notes using Natural Language Processing and serve as additional input features. Diagnosis codes at discharge are used to label the samples and serve as targets for the model training process; more details can be found in our previous work.[21] The generated models are evaluated on their performance on retrospective data using common metrics such as the area under the receiver operating curve (AUROC), sensitivity, specificity, etc. The model that passes the model evaluation is deployed as the delirium risk prediction service which generates predictions on live data.

The prediction service generates a fractional score, with higher values indicating higher confidence that a patient will develop delirium. Evaluation is performed several times a day, following any change to a patient's record in the EHR system, such as the addition of new lab results. When a change happens, a request is sent to the delirium risk prediction service. The prediction service parses each prediction request into an observation which is used to generate a prediction. The predictions are returned and displayed in the EHR system. When deployed into clinical practice, the prediction service generates

[†] clinalytix, DH Healthcare GmbH, CE number (01)4260693990033(8012) 3.0 (11)14012021.

warnings when certain thresholds are crossed. Whenever its internal risk score increases past 0.5, a ‘yellow alert’ is generated, indicating that a patient is at increased risk. The risk score passing 0.75 generates a ‘red alert’, indicating a high risk of delirium. Thresholds were chosen based on the evaluation of the trained model on a separated testing set.[22] Warnings were only visible to the AI working group and were not displayed to the clinicians on the wards or to those conducting the screenings.

For all cases included in the analysis, prediction scores were generated based on a chronological stream of stored observations. In the following statistical analysis, five risk scores were extracted per patient by selecting the highest daily value from the day before surgery, the day of surgery and from all of the first three postoperative days. A chronological stream of the predictions made on the same patient is generated to show the changes of prediction scores following the progress of the patient stay. Observation inputs for each prediction can be analyzed and the top contributing features for each prediction are also provided for further analysis. In this cohort study, patients could only be included from January 2021, since the data were not available for older records. The AI is blinded to the notes of the postoperative visits so as to prevent the reference test from influencing the screening tool.

The AI performance in the postoperative screening use case is evaluated by calculating the correlation of AI scores with POD, generating the contingency table, and calculating related performance measures and a receiver operating characteristic (ROC) curve analysis.

Inspecting the artificial intelligence model: predictive factors

Individual predictions can be analysed using a method published by Ribeiro *et al.* in 2016.[24] This assesses the relative weight of individual factors contributing to a single prediction. All predictions gathered for this study were analysed to find the number of unique contributing and discounting factors. Furthermore, the highest scoring postoperative prediction from all patients with POD that were classified as ‘high risk’ (>0.75) were analysed ($n = 20$). For each patient, the five factors with the most positively contributing impact were extracted and these are summarized into a histogram (Fig. 1). This allows for plausibility checking, but it also serves as evidence of the complexity of the underlying model.

Risk factors and frailty assessment

In addition to CAM-ICU and frailty screening (including a Shulman test), all patients underwent the clinic’s standard pre-operative risk evaluation (Table 1). Predisposing factors for postoperative delirium according to the literature were identified from the patient history and all patients with more than three risk factors were considered to be at risk of POD (Table 2). We analysed the predictive quality of all screening methods by calculating their association with POD as well as their cut-off, sensitivity, specificity and a ROC curve analysis.

Results

Patient collective

The mean age of the patients included in this study was 76.7 years, with women comprising 42.1% of the sample, and 57.9% of patients underwent surgical rather than endovascular interventions. The overall incidence of postoperative delirium was 23.7%. The incidence of POD following endovascular intervention was significantly lower at 10.4% compared to 33.3% for surgical interventions ($P < 0.01$, χ^2 test). Compared to patients who did not develop POD, for those who did develop POD there was a significant increase in the duration of postoperative hospitalisation (12.8 *vs.* 8.6 days, $P < 0.01$, Student’s *t*-test) and in the duration of postoperative intensive care over the first 3 days (1.7 *vs.* 0.3 days, $P < 0.01$, *t*-test). There were no statistically significant associations of POD with gender, age (likely skewed by the preselection of patients of 65 years and older,) postoperative nausea and vomiting (PONV), and postoperative pain (Table 3).

Risk stratification

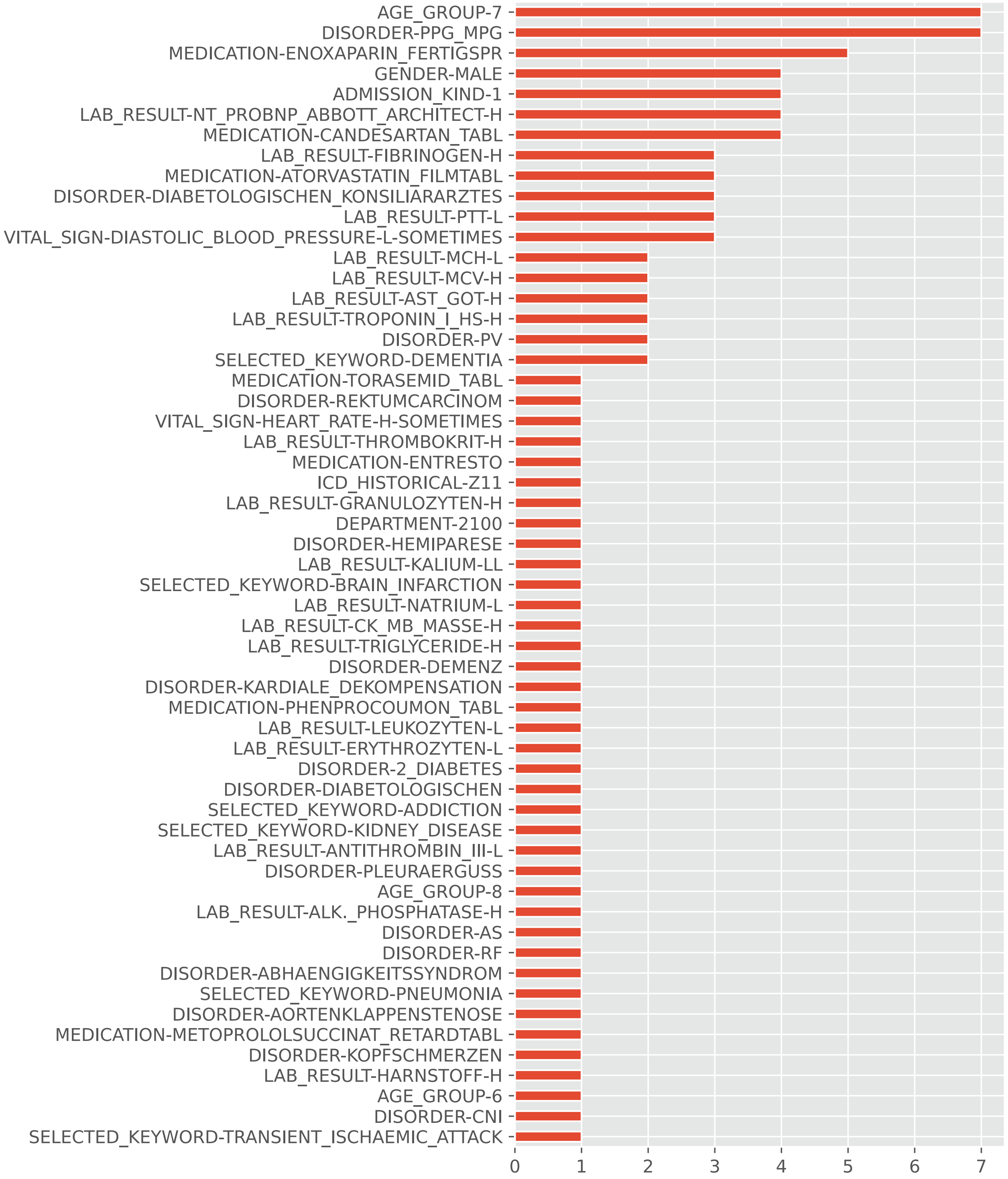
Plotting the AI scores over the peri-operative period monitored for this study reveals obvious differences in the scores of patients who developed POD compared to patients who did not (Fig. 2). There is a visible tendency for patients with POD to have had higher scores preoperatively. However, the graphical presentation already gives a clear indication that no threshold with perfect discriminatory performance can be defined. Adhering to the 0.5 ‘increased risk’ threshold results in an incidence closely matching that of the reference test.

Table 4 shows how the different methods of pre-operative risk stratification and the development of POD. All described methods identified a number of patients closely matching the expected and observed incidence in the study collective. For all methods, sensitivity was below 0.5 and specificity between 0.76 and 0.82. Calculating correlations between screening scores and POD yields significance at a 95% confidence level only for the preoperative Shulman test ($P < 0.01$, Mann–Whitney *U*-test (MWU)) in comparison to AI ($P = 0.06$, *t*-test), the checklist ($P = 0.06$, MWU) and the frailty assessment ($P = 0.48$, MWU). Figure 3 shows the ROC analysis.

Postoperative screening

Postoperatively, patients with and without POD received scores with considerable variance. A good approximation of the predictive performance of AI screening for POD can be calculated by comparing AI scores >0.75 during the first three postoperative days with the CAM-ICU during these 3 days. The highest postoperative AI score is highly correlated with the incidence of postoperative

Fig. 1 Histogram showing the frequency of the top five contributing factors in postoperative, true positive delirium predictions ($n = 20$).



delirium ($P < 0.001$, t-test), and the 'high-risk' threshold of 0.75 already achieves reasonable discriminatory performance. Table 5 shows the contingency table for the AI software as a postoperative screening tool and Fig. 4 shows the ROC analysis (AUROC 0.79; 95% CI, 0.69–0.87)

Exploration of artificial intelligence features

While a few features can be identified as more universally relevant, most features only occur once or twice in the investigated sample. In the case of the analyzed high-risk predictions, 56 individual factors were extracted, with

Table 3 Patient characteristics, and the occurrence of postoperative delirium

Patient characteristic	Postoperative delirium[†] (+)	Postoperative delirium[†] (–)	Combined
All patients	27/114, 23.7%	87/114, 76.3%	114, 100%
Male	15/66, 22.7%	51/66, 77.3%	66, 100%
Female	12/48, 25.0%	36/48, 75.0%	48, 100%
Age	77.7 (SD 6.7, IQR 72.5–82.5, $n=27$)	76.4 (SD 7.7, IQR 68.5–83, $n=87$)	76.7 (SD 7.5, IQR 69–83, $n=114$)
Surgical	22/66, 33.3%	44/66, 66.7%	66, 100%
Endovascular	5/48, 10.4%	43/48, 89.6%	48, 100%
Length of postoperative stay[°]	12.8 d (SD 5.5, IQR 9–14, $n=25$)	8.6 d (SD 3.8, IQR 6–10.75, $n=86$)	9.5 d (SD 4.6, IQR 7–11, $n=111$)
Length of postop. ICU-stay[&]	1.7 (SD 1.2, IQR 1–3, $n=27$)	0.3 (SD 0.7, IQR 0–0, $n=87$)	0.66 (SD 1.1, IQR 0–1, $n=114$)
NRS[$]	2.4 (SD 2.7, IQR 0–5, $n=24$)	2.5 (SD 3.0, IQR 0–5, $n=75$)	2.5 (SD 2.9, IQR 0–5, $n=100$)
PONV[*]	5/27, 18.5%	15/87, 17.2%	20/114, 17.5%

[†]Diagnosed by CAM-ICU screenings on the 3 days following intervention. [°]Days from intervention to discharge. [&]Only reported for the three days of postoperative screening. [$]"Numerical Rating Scale", self-reported pain on a scale from 0 (no pain) to 10 (worst imaginable), averaged over three days post intervention. *Postoperative nausea and vomiting.

just 18 factors having been observed in more than one prediction. Analysing all predictions evaluated in this study yields 426 unique contributing and 318 unique discounting factors. Furthermore, features are context-dependent, meaning that a contributing feature in one prediction can appear as a discounting feature in another prediction. One instance of context sensitivity can be read from Fig. 1: even within this rather small group ($n=20$), the AI identified male as well as female gender as contributing factors in different circumstances. Across all predictions in this study cohort, there are 214 different factors appearing as contributing as well as discounting factors.

Discussion

The detrimental effects of POD on the recovery of cardiac surgery patients are well established. A multitude of factors contribute to the development, many of them

Fig. 2 Boxplot showing dispersion of AI predictions per day in a 5-day perioperative timeframe. AI, artificial intelligence.

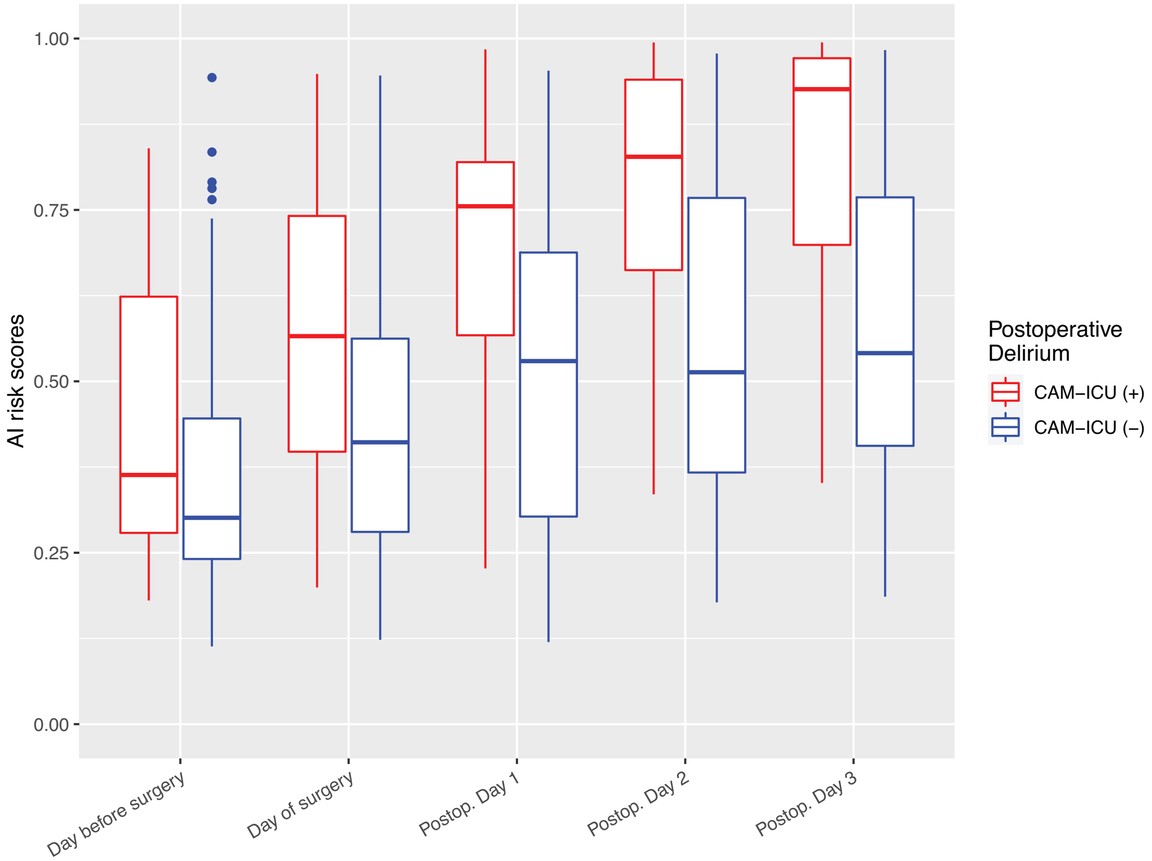


Table 4 Risk score and discrimination by cut-off

Risk scores by POD	Postoperative delirium[†] (+)	Postoperative delirium[†] (−)
AI pre-op[*]	0.45 (IQR 0.28 to 0.62, $n=27$)	0.35 (IQR 0.24 to 0.45, $n=87$)
AI pre-op[*] >0.5	9/25, 36.0%	16/25, 64.0%
AI postop[‡]	0.83 (IQR 0.73 to 0.97, $n=27$)	0.61 (IQR 0.45 to 0.80, $n=87$)
AI postop[‡] >0.5	24/83, 28.9%	59/83, 71.1%
AI postop[‡] >0.75	20/52, 38.5%	32/52, 61.5%
Risk checklist[#]	3.1 (IQR 2 to 4, $n=26$)	2.5 (IQR 2 to 3, $n=86$)
Risk checklist[#] >3	12/32, 37.5%	20/32, 62.5%
Frailty score[§]	2.0 (IQR 0 to 3, $n=26$)	1.5 (IQR 0.5 to 2, $n=87$)
Frailty score[§] >2	10/31, 32.3%	21/31, 67.7%
Preop Shulman test	2.5 (IQR 1 to 3, $n=27$)	1.7 (IQR 1 to 2, $n=84$)
Preop Shulman test >2	13/33, 39.4%	20/33, 60.6%

IQR, interquartile range; POD, postoperative delirium. [†]Diagnosed by CAM-ICU screenings on one of the 3 days following intervention; [*]Highest score on the day before surgery. [‡]Highest AI score at any point during the 3 days following surgery. [#]See Table 2. [§]See Table 1.

Table 5 Contingency table: postoperative AI screening by CAM-ICU

	POD (+)	POD (−)	Sum	
AI screening (+)	20	32	52	PPV = 0.38
AI screening (−)	7	55	62	NPV = 0.89
Sum	27	87		
	Sens = 0.74	Spec = 0.63		

AI, artificial intelligence; NPV, negative predictive value; POD, postoperative delirium; PPV, positive predictive value; Sens, sensitivity; Spec, specificity. Absolute frequencies of manifest postoperative delirium by artificial intelligence screening result (AI screening (+), AI risk score ≥ 0 0.75).

subject to ongoing research, like neuroinflammation.[25] Since there is no effective therapy for POD, risk stratification and prevention are paramount.[26] But while pre-operative assessment for other organ systems is commonplace, the pre-operative neurological status, let alone psychiatric comorbidities, are not as rigorously explored.[27]

Our data show an expected incidence of POD and well established associations like increased length of stay on ICU and wards. These findings stress the relevance of prophylaxis and care for POD and the motivation for, and potential of, a potent low-barrier screening mechanism.

In previous studies, AI's predictions demonstrated high performance when evaluated at hospital discharge, both retrospectively (AUROC 0.931; 95% CI, 0.929 to 0.932) and prospectively (AUROC 0.934; 95% CI, 0.927–0.942).[21,22] However, the performance dropped markedly in the current study with an AUROC of 0.79, (95% CI, 0.69–0.87) in the 3-day postoperative period.

This performance degradation is caused by multiple factors. First, the delirium prediction model is trained with a different labelling strategy in contrast to the CAM-ICU labels used in this study. The postoperative assessments that were used as reference in this study were not available throughout the training dataset. The AI was trained to assign high scores to patients who had a delirium diagnosis coded at discharge. If a patient was categorised as 'high-risk' but developed delirium only after the three-day follow-up, they were classified as false positives in this study while being true positives with regard to the training labelling strategy. Additionally, due to under-coding, the labelling strategy using discharge codes is generally more restrictive.

Fig. 3 Receiver operating characteristic (ROC) curves of preoperative risk stratification tools.

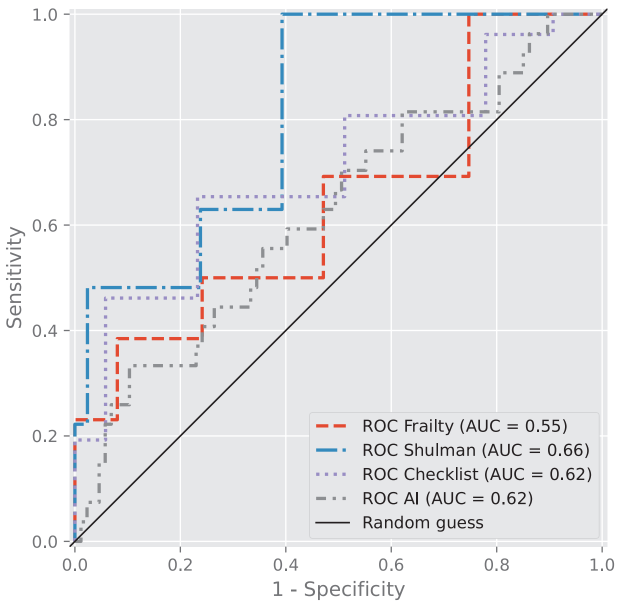
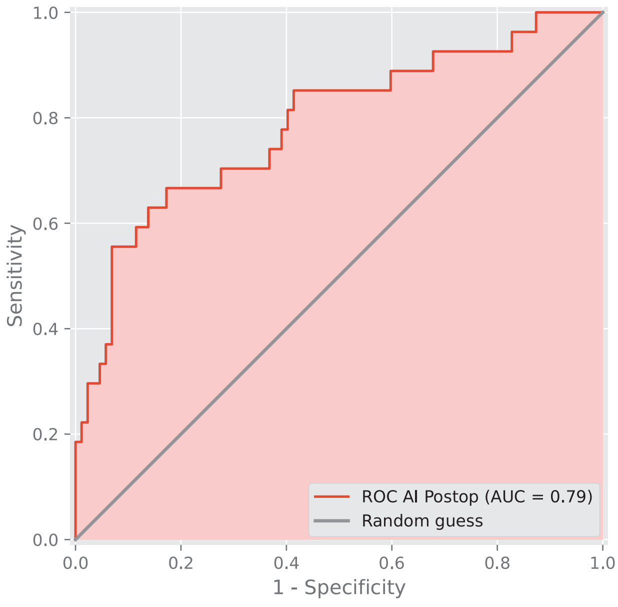


Fig. 4 Receiver operating characteristic (ROC) curve with indicated area under the curve (AUC) for artificial intelligence (AI) predictions of postoperative delirium in the postoperative screening use case.



Second, the training dataset was not preselected for age and need for surgery. The training dataset comprised of several years of data, with delirium having been significantly undercoded, especially in the earlier years. The cohort in this study, in contrast, is preselected by age and severity of illness and closely monitored. It follows that the model was trained on a more general population with an overall lower risk and without the intervention of surgery. Advanced age and major surgery are relevant risk factors all patients in this cohort share, making it harder to discern what are essentially very-high-risk patients from a generally high-risk population. High-risk assessments, even if classified as false positives in this study, may still be warranted.

Third, the medical centre hosting this cohort study is a highly specialized clinic, with many patients only undergoing their procedure there but with the postinterventional care in domo. This results in very little data being available to the AI screening tool for pre-operative evaluation. A larger amount of data might help resolve this problem in other settings and potentially in the studied centre as well, if more healthcare data were available, e. g. from primary care physicians.

Overall, none of the examined methods for pre-operative evaluation stand out as particularly powerful. The risk checklist and Shulman test perform comparably but miss more than half of the recorded POD-instances while producing about twice as much false positives than true positives. A pre-operative frailty assessment certainly provides more value than simply identifying patients at risk for POD, but it does not perform better in this regard than either the Shulman test or the checklist. The AI system performs at a comparable level, identifying a similar number of true positives with a significant number of false positives as well. However, it should be remembered that while the CAM-ICU as a reference test is highly specific, it has imperfect sensitivity.[28] Also, a once-daily screening might miss instances of POD.[29]

Postoperatively, in the absence of regular screening, the use of the AI as a trigger for more specific diagnostic measures would have caught about two thirds of the patients affected. Revisiting the data from previous studies, where active screening brought about a tenfold increase in the incidence of POD, this reveals a significant potential for AI as a low-barrier screening method. There is still room for improvement as the AI does not currently evaluate anesthesia records. Information on medication, vital signs, duration of the procedure and other details would improve postoperative screening performance.[30]

Strengths and weaknesses

This study evaluates a model trained on eleven years of EHR data and compares its predictions against data from an active screening protocol. Consequently, the data quality of the reference screening is substantially better than that of the training data. The stream data representation explained above provides strong protection against information leakage. Predictions are generated continuously over the hospitalisation period with the availability of new data, enabling more applications for this approach than fixed timepoints for predictions, especially in contrast to solutions which score patient risk only on the day of admission, the day after admission or the day of surgery. Another point of note is the mode of model generation and generalisability described in detail in our previous work.[21]

The most notable weakness of this study is the rather small dataset, which despite high data quality and a realistic POD incidence, is not quite large enough to claim stable correlations.[31] Also of note, the training data set is comprised of cases from patients who, for the most part, were not actively screened for POD. Remarkably, although trained from a relatively undercoded set with a different labeling strategy, the AI produced a predicted incidence closely matching the reference test.

An aspect to be addressed in future work is the prediction and distinction of delirium sub-phenotypes. Due to coding challenges, a comparably large database to enable subphenotype learning and prediction is not yet available at our center.

Conclusion

Low-effort, highly available screening for postoperative delirium is an important field of research, and strong tools would probably contribute to improvements in peri-operative care. While artificial intelligence' as evaluated in this cohort' does not yet outperform other methods of pre-operative risk stratification, the potential for development continues to be significant and should be explored further.

Using artificial intelligence predictions as a screening aid during postoperative care provides an opportunity to improve awareness of POD in general, and to bring a reasonably effortless screening method to peripheral, non-intensive care wards. The data from this cohort inspires optimism in the ability of AI, especially when incorporating intra-operative records, to approach reasonable diagnostic accuracy, to the benefit of patients and healthcare providers.

Elderly patients, a demographic of growing relevance for decades to come, stand to benefit the most from improvements in this area. A significant impact on short- and long-term functional outcomes is evident from current literature, and recovery metrics such as length of stay and need of ICU care in this cohort support such claims. Furthermore, it becomes clear that POD is also a critical factor in a healthcare system which needs to optimise resource allocation to be able to provide adequate care for the growing proportion of the elderly.

Acknowledgements relating to this article

Assistance with this article: we would like to thank Patricia Cabanillas Silva and Kristof Depraetere for their valuable contributions with processing the predictions results.

Financial support and sponsoring: authors from Dedalus received funding from the EU Horizon 2020 projects Genomed4All (101017549) and PERSIST (875406).

Conflicts of interest: authors from HDZ have no competing interests to declare. RS, LM, HS, MD work for Dedalus Healthcare.

Presentation: presented at the annual meetup of the cardiac anesthesia working group of The German Society of Anaesthesiology and Intensive Care Medicine (DGAI) on 26th November 2021 in Fulda, Germany.

This manuscript was handled by Pierre Gregoire Guinot.

References

1 Lütz A, Heymann A, Radtke FM, Spies CD. Was wir nicht messen, detektieren wir meist auch nicht. *AINS – Anästhesiologie · Intensivmedizin · Notfallmedizin · Schmerztherapie* 2010; **45**:106–111.

2 Ely EW, Shintani A, Truman B, *et al.* Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *JAMA* 2004; **291**:1753–1762.

3 Leslie DL, Inouye SK. The importance of delirium: economic and societal costs. *J Am Geriatr Soc* 2011; **59 (Suppl. 2)**:S241–S243.

4 Marcantonio ER, Flacker JM, Wright RJ, Resnick NM. Reducing delirium after hip fracture: a. randomized trial. *J Am Geriatr Soc* 2001; **49**:516–522.

5 Safavynia SA, Arora S, Pryor KO, Garcia PS. An update on postoperative delirium: clinical features, neuropathogenesis, and perioperative management. *Curr Anesthesiol Rep* 2018; **8**:252–262.

6 Rudolph JL, Inouye SK, Jones RN, *et al.* Delirium: an independent predictor of functional decline after cardiac surgery. *J Am Geriatr Soc* 2010; **58**:643–649.

7 Maniar HS, Lindman BR, Escallier K, *et al.* Delirium after surgical and transcatheter aortic valve replacement is associated with increased mortality. *J Thorac Cardiovasc Surg* 2016; **151**:815–823.

8 Potter BJ, Thompson C, Green P, Clancy S. Incremental cost and length of stay associated with postprocedure delirium in transcatheter and surgical aortic valve replacement patients in the United States. *Catheteriz Cardiovasc Interv* 2019; **93**:1132–1136.

9 Codling D, Hood C, Bassett P, *et al.* Delirium screening and mortality in patients with dementia admitted to acute hospitals. *Aging Ment Health* 2021; **25**:889–895.

10 Corradi JP, Thompson S, Mather JF, *et al.* Prediction of incident delirium using a random forest classifier. *J Med Syst* 2018; **42**:1–10.

11 Wong A, Young AT, Liang AS, *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 Netw Open* 2018; **1**:e181018.

12 Jauk S, Kramer D, Großauer B, *et al.* Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study. *J Am Med Inform Assoc* 2020; **27**:1383–1392.

13 Fliegenschmidt J, Hulde N, Preising MG, *et al.* Artificial intelligence predicts delirium following cardiac surgery: a case study. *J Clin Anesth* 2021; **75**:110473.

14 Fried LP, Tangen CM, Walston J, *et al.* Frailty in older adults: evidence for a phenotype. *J Gerontol: Med Sci* 2001; **56**:M146–M157.

15 Numan T, van den Boogaard M, Kamper AM, *et al.* Delirium detection using relative delta power based on 1-min single-channel EEG: a multicentre study. *Br J Anaesth* 2019; **122**:60–68.

16 Van Rossum G, Drake FL. *Python 3 reference manual.* Scotts Valley, CA: CreateSpace; 2009.

17 McKinney W, others. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 51–6.

18 Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 2020; **17**:261–272.

19 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 2011; **12**:2825–2830.

20 Hunter JD. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 2007; **9**:90–95.

21 Sun H, Depraetere K, Meesseman L, *et al.* A scalable approach for developing clinical risk prediction applications in different hospitals. *J Biomed Inform* 2021; **118**:103783.

22 Sun H, Depraetere K, Meesseman L, *et al.* Machine learning–based prediction models for different clinical risks in different hospitals: evaluation of live performance. *Journal of Medical Internet Research* 2022; **24**:e34295.

23 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017; **30**:.

24 Ribeiro MT, Singh S, Guestrin C. *Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Association for Computing Machinery; 2016. pp. 1135–1144.

25 Zhu Y, Zhou M, Jia X, *et al.* Inflammation disrupts the brain network of executive function after cardiac surgery. *Annals of surgery* 2021.

26 Jin Z, Hu J, Ma D. Postoperative delirium: perioperative assessment, risk reduction, and management. Vol.125. *Br J Anaesth* 2020; **125**:492–504.

27 Alam A, Ma D. Is it time to assess neurological status before surgery to improve postoperative outcomes? *Ann Surg* 2022; **275**:644–645.

28 Gusmao-Flores D, Salluh JI, Chalhub RÁ, Quarantini LC. The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: a systematic review and meta-analysis of clinical studies. *Crit Care* 2012; **16**:1–10.

29 Hamadnalla H, Sessler DI, Troianos CA, *et al.* Optimal interval and duration of CAM-ICU assessments for delirium detection after cardiac surgery. *J Clin Anesth* 2021; **71**:110233.

30 Andrási TB, Talipov I, Dinges G, *et al.* Risk factors for postoperative delirium after cardiac surgical procedures with cardioplegic arrest. *Eur J Cardio-Thorac Surg* 2022; **62**:17.

31 Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Academic Emergency Medicine* 1996; **3**:895–900.