JAMIA Open, 2025, **8(3)**, ooaf037
https://doi.org/10.1093/jamiaopen/ooaf037
Review

Review

Determining the ground truth for the prediction of delirium in adult patients in acute care: a scoping review

Lili M. Schöler, MSc[1,2], Lisa Graf, MSc*[,3,4], Antti Airola, PhD[5], Alexander Ritzi, MSc[1,6], Michael Simon, PhD[7], Laura-Maria Peltonen, PhD[2,8]

[1]Department of Nursing, Medical Center—University of Freiburg, Freiburg 79106, Germany, [2]Department of Nursing Science, University of Turku, Turku 20520, Finland, [3]Department of Neurology, Medical Center—University of Freiburg, Freiburg 79106, Germany, [4]Neurorobotics Lab, Department of Computer Science, University of Freiburg, Freiburg 79110, Germany, [5]Department of Computing, University of Turku, Turku 20500, Finland, [6]Centre for Geriatric Medicine and Gerontology (ZGGF), Medical Center—University of Freiburg, Freiburg 79106, Germany, [7]Institute of Nursing Science, Department of Public Health, University of Basel, Basel 4056, Switzerland, [8]Research Services, The Wellbeing Services County of Southwest Finland, Turku 20521, Finland

*Corresponding author: Lisa Graf, MSc, Neurorobotics Lab, Department of Computer Science, University of Freiburg, 79110 Freiburg, Georges-Köhler-Allee 201, Germany (lgraf@cs.uni-freiburg.de)

L.M. Schöler and L. Graf contributed equally to this work.

Abstract

Objective: Delirium is a severe condition, often underreported and linked to adverse outcomes such as increased mortality and prolonged hospitalization. Despite its significance, delirium prediction is often hindered by underreporting and inconsistent labeling, highlighting the need for models trained on reliably labeled data (ground truth). This review examines (i) practices for determining labels in delirium prediction models and (ii) how study designs affect label quality, aiming to identify key considerations for improving model reliability.

Materials and Methods: A search of Cochrane, PubMed, and IEEE identified 120 studies that met the inclusion criteria.

Results: To establish the ground truth, 40.8% of studies used routine data, while 42.5% used primary data. The Confusion Assessment Method (CAM) was the most widely used assessment tool (60. 0%). Label and data leakage occurred in 35.0% of studies. High Risk of Bias (RoB) was a recurring issue, with 31.7% of studies lacking sufficient reporting and 36.7% showing inadequate outcome determination. Studies using primary data had lower RoB, whereas those with unclear label sources displayed higher RoB.

Discussion: Our findings underscore the importance of careful planning in determining the ground truth frequently neglected in existing studies. To address these challenges, we provide a decision support flowchart to guide the development of more accurate and reliable prediction models.

Conclusion: This review uncovers significant variability in labeling methods and discusses how this may affect delirium prediction model reliability. Highlighting the importance of addressing underreporting bias and providing guidance for developing more robust models.

Lay Summary

Delirium, a serious condition causing acute confusion in hospitalized patients, is linked to worse outcomes, including longer hospital stays, higher costs, and increased mortality. However, it is often underreported, making it difficult for artificial intelligence (AI) models to predict accurately. When hospitals fail to document delirium cases, AI models may only detect severe cases already flagged by clinicians. Our review of 120 studies examined why prediction models struggle with reliability. Nearly half relied on routine hospital records prone to underreporting, while others used data collected specifically for model development. The Confusion Assessment Method was the most common tool, but over a third of studies had errors, such as including post-outcome data in predictions or using unclear labels, which can falsely inflate accuracy. Models using direct assessments performed better, emphasizing the need for high-quality data. To address these issues, we developed a step-by-step guide to help researchers and clinicians build fairer, more reliable models. This tool promotes careful planning to reduce bias and improve detection of overlooked cases. By improving data quality, healthcare teams can create AI-driven prediction tools that better identify delirium early, ultimately reducing complications and improving patient outcomes.

Key words: delirium; ground truth; electronic health records; prediction models; machine learning.

Introduction

The advent of electronic health record (EHR) data has opened up new possibilities for disease identification and prediction, with numerous studies showcasing the potential of EHR-based prediction models to forecast conditions like delirium. Delirium—characterized by disruptions in attention, cognition, and consciousness due to underlying medical issues[1]—remains a global healthcare challenge. Its far-reaching consequences include increased cognitive decline, morbidity, mortality, dementia progression, healthcare costs,

Received: February 10, 2025; **Revised:** April 10, 2025; **Editorial Decision:** April 22, 2025; **Accepted:** April 25, 2025

and premature transfers to long-term care facilities.[2,3] The delirium incidence varies widely, from 10% to 82%, depending on clinical context.[4]

Despite the high predictive power of delirium prediction models,[5] one fundamental obstacle remains: the lack of reliable ground truth. "Ground truth" refers to the unambiguous labeling of data that defines what a prediction model must learn and enables performance evaluation. For delirium, this means classifying patients as true positives (with delirium) or true negatives (without delirium) according to the delirium labels (hereafter "labels"), sometimes called phenotyping. However, detecting delirium is complex due to its variable and nuanced presentation. Unfortunately, delirium often goes undetected, underreported, and poorly documented in hospitalized patients,[6–9] resulting in many false negatives in EHR data.[10] Barriers to delirium recognition include its variable nature and lack of competence in delirium recognition.[11]

Robust screening and assessment are essential for enhancing delirium detection,[11] crucial for effective delirium prevention and care.[12,13] Yet, only 37% of nursing staff find screening instruments worthwhile,[14] and delirium is often under-documented and under-diagnosed, with International Classification of Diseases (ICD) codes[15] assigned in only 3% to 34% of cases that should have been diagnosed and documented.[9,16]

This review examines methods used to determine the ground truth for delirium prediction models, as model accuracy depends heavily on training data quality. We intentionally avoid comparing the performance of different delirium prediction models, as this would be misleading. Each model's performance is only as good as its underlying ground truth, which we show varies significantly between studies. By neglecting to scrutinize the labeling methods employed, we risk propagating biases and inaccuracies, ultimately undermining the validity of these models in real-world clinical settings. Instead, we provide a detailed analysis of the labeling strategies used in delirium prediction modeling, highlighting their strengths, limitations, and potential pitfalls. We aim to spark a crucial conversation about the importance of finding reliable ground truth and to provide guidance for developing reliable delirium prediction models. To our knowledge, this is the first review in this area, whereas previous reviews have focused on model performance.[5,17–20]

Methods

This scoping review followed Von Elm et al.'s 5-phase framework[21]: (1) identifying the research question; (2) identifying relevant studies; (3) study selection; (4) charting the data; and (5) collating, summarizing, and reporting the results.

Our research questions were as follows:

- What are the common practices for determining the ground truth in delirium prediction models?
- How do different study designs for determining ground truth (eg, routine assessments versus primarily collected delirium labels) influence delirium prediction models, and what are the implications for model development and evaluation?

Here, "study design" refers to 2 distinct approaches for labeling delirium in prediction model datasets: (1) studies that collect delirium labels specifically for model development through a structured research protocol (primary data) and (2) studies using labels from routine clinical assessments (routine data). This distinction is crucial, as protocol-based labeling often results in higher quality but smaller datasets, whereas routine data provide larger datasets with potentially less consistent labeling and data quality.[22]

A search was conducted on December 4, 2023, in PubMed, IEEE, and the Cochrane database. The full search strategy is available in the OSF protocol (https://osf.io/htmsc/). The software tool Covidence[23] facilitated the screening process. Inclusion and exclusion criteria are listed in Table 1. Two independent reviewers, 1 from nursing and 1 from computer science, screened and extracted data.

The PRISMA flow chart (Figure 1) shows the study selection process.

Data were extracted with REDCap (v14.0.30)[24,25] with a form based on the CHARMS checklist[26] and the PROBAST tool,[27] tailored to our research questions. The form was piloted, reviewed, and refined. A condensed version is displayed in Table 2, with the full version in the Supplementary Material.

We assessed sample bias by evaluating 3 exclusion criteria used in studies: (1) incomplete or missing assessments, (2) missing data, and (3) delirium at hospital admission.

Risk of Bias (RoB) was assessed based on the reliability of delirium detection instruments, rater training, usage frequency, dataset biases (eg, patient exclusion due to missing data), and incidence plausibility.

We did not evaluate model performance but checked for label and data leakage.[28] Label leakage occurs when outcome labels are inadvertently included in model inputs, such as using medication for delirium treatment to predict delirium. Data leakage occurs when training and test sets are not strictly separate. This pervasive problem in healthcare prediction models can occur through various mechanisms, such as temporal leakage.[29] We investigated this issue by checking whether post-outcome data were masked during model development, meaning that the input for delirium prediction came exclusively from pre-delirium episode data.

We calculated a Cohen's kappa of 0.71 for 10% of the studies (n = 12) on the final RoB judgment. In 2 studies, the reviewers disagreed: the computer science reviewer judged them as "unclear," while the nursing reviewer judged them as "high" RoB. Both assessments were poorly described, with one study misinterpreting assessment results,[30] and the other mentioning missing regular assessments only in the discussion.[31]

Results

We identified 120 studies meeting our inclusion criteria. For a comprehensive overview of those studies, see Table SA2 (study settings and delirium outcomes) and Table SA4 (modeling techniques).

Lengthy reference lists were excluded from the main text to improve clarity and brevity. All extracted numerical data are detailed in the Supplementary Material S3.

Study designs to determine ground truth

Dataset labeling was based on routine data, primary study-specific efforts, or a combination of both. The analysis shows that 40.8% (n = 49) of the studies relied exclusively on routine data, 42.5% (n = 51) used labels collected for study

Table 1. Population, Concept, and Context (PCC) with inclusion and exclusion criteria.

PCC element	Inclusion	Exclusion
Population: Adult acute care patients	Human patients admitted to a hospital.	Pediatric populations as defined by the authors.
Concept: Delirium	Delirium, not induced by alcohol and other psycho-active substances as for example defined by the F05 ICD-10 code.[15]	Individual symptoms of delirium that do not meet the criteria of delirium as a syndrome (like disorientation or agitation).
Context: Predictive models using EHR data as input	Original peer-reviewed studies about the development of a predictive model using input features derived from EHR data.	Diagnosing delirium rather than predicting, identifying individual predictors, and predicting other outcomes such as geriatric syndromes or postoperative complications if this model cannot predict delirium itself.

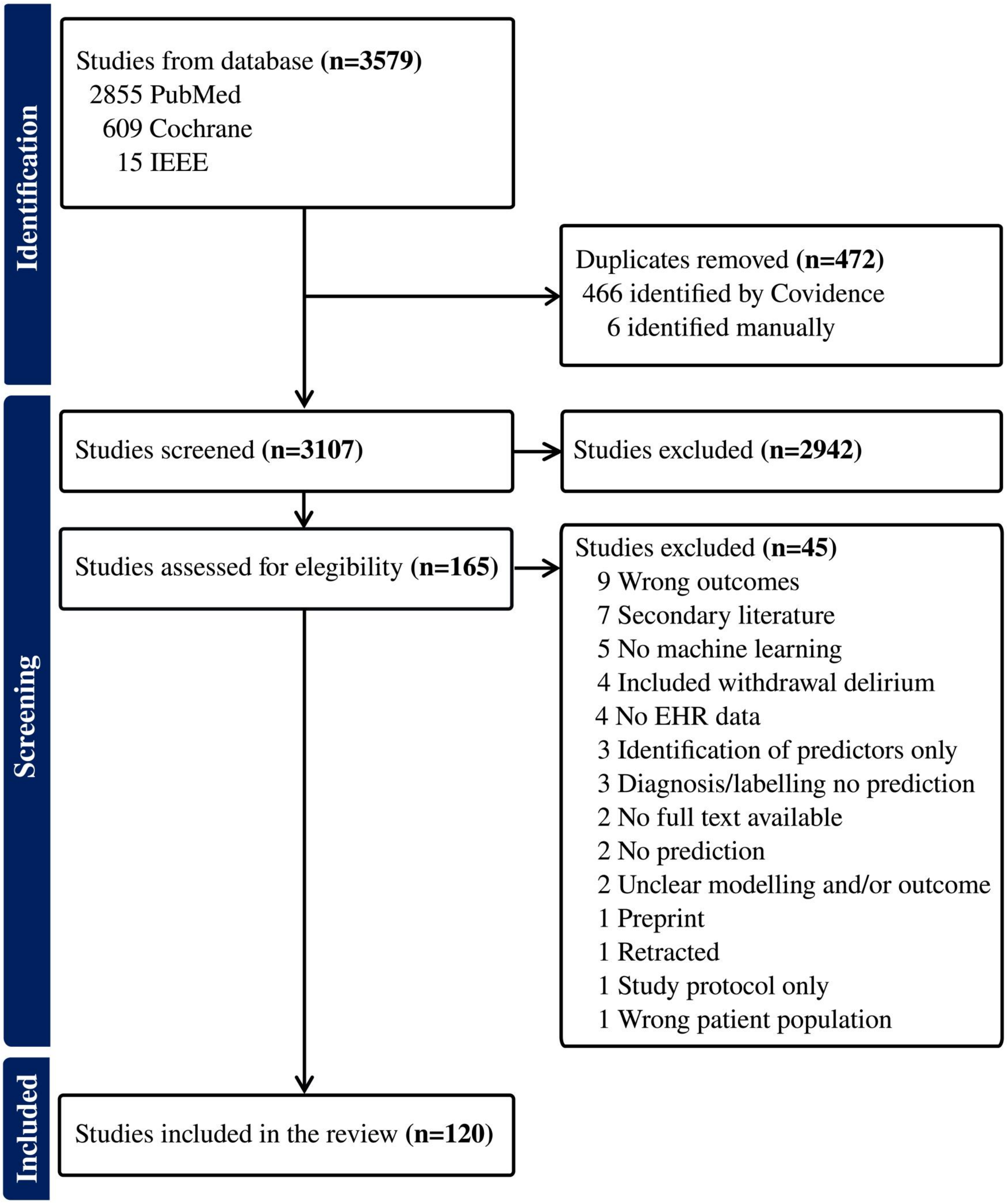
Identification
Studies from database **(n=3579)**
2855 PubMed
609 Cochrane
15 IEEE
Duplicates removed **(n=472)**
466 identified by Covidence
6 identified manually
Screening
Studies screened **(n=3107)**
Studies excluded **(n=2942)**
Studies assessed for elegibility **(n=165)**
Studies excluded **(n=45)**
9 Wrong outcomes
7 Secondary literature
5 No machine learning
4 Included withdrawal delirium
4 No EHR data
3 Identification of predictors only
3 Diagnosis/labelling no prediction
2 No full text available
2 No prediction
2 Unclear modelling and/or outcome
1 Preprint
1 Retracted
1 Study protocol only
1 Wrong patient population
Included
Studies included in the review **(n=120)**

Figure 1. PRISMA flow chart for study selection.

Table 2. Condensed version of the extraction form, full extraction form available in the Supplementary Material.

Dimension	Variable	Response options
Study information	Authors/Year/Title/Journal or Conference	Text fields
Source of data	Dataset	Public Dataset/Registries/Clinical Data/Different study/Unclear/NR
	Study setting/Country/Population/Age/Data collection period/Study sites/Inclusion-Exclusion criteria	Text fields
Participants	Exclusion because of . . .	Yes/No
	. . . missing data	
	. . . missing or incomplete delirium assessment	
	. . . delirium on admission	
	Outcome definition	Text fields
Outcome for prediction	Delirium subgroup distinction	Yes/No/I can't say
	Source of label	Instruments[a]/ICD Code/ Review method/Other/NR
	Description of chart review or "Other"	Text fields
	Routine Data or Primary Data?	Routine care data/Primary data (collected for study purposes)/I can't tell
	Frequency/time point of assessment	Text field (per instrument)
	Who used the instrument?	Text field
	Information on fidelity	
Sample size	Number of participants and Incidences	Text fields
Model	Modeling method and which was the final model?	Text fields
	Number of features (before/after feature selection)	
Development	Alcohol or substance abuse as final feature?	Yes/No/I can't say
	Development of an online tool, score or nomogram?	
	Model implemented in the hospital?	
	External validation?	Yes/No/I can't say
Results	How did they do the internal validation?	Random split/Data from different time/ Data from different department/No validation/Only external validation/NR
	Did they mask post-outcome data?	Yes/No/I can't say
	Were appropriate data sources used?	
Risk of bias assessment	Was the outcome determined appropriately?	
	Outcome defined/determined in similar way for all participants?	For all questions and possible answers of risk of the bias assessment (Table SB1).
	Reasonable number of participants with the outcome?	

[a] Instruments were listed individually.

purposes (hereafter called "primary data"), 8.3% (n = 10) used both, and in 8.3% (n = 10), the method was unclear due to insufficient reporting.

About a quarter of papers (24.2%, n = 29) excluded patients due to missing delirium assessments, 25.8% (n = 31) excluded patients due to missing data, and one third (33.3%, n = 40) excluded patients with delirium within the first 24 hours of inpatient stay or at admission.

When we excluded studies using routine data which also excluded patients due to missing assessments (38.8%, n = 19), the delirium incidence, and variability decreased (see Figure 2).

Datasets, settings, and patient populations

Most datasets (78.3%, n = 94) were derived from EHR data directly from clinical records. Additionally, 9.2% (n = 11) used registry data based on EHR data, and 7.5% (n = 9) relied on datasets from other studies. Public datasets, like MIMIC-III,[32] MIMIC-IV,[33] and eICU,[34] were used in 5.0% (n = 6) of studies. Most models (75.8%, n = 91) were developed and internally validated using data from a single hospital. In 36.7% (n = 44) of the studies, models were developed for intensive care unit (ICU) populations, 29.2% (n = 35) for mixed populations, 8.3% (n = 10) for non-ICU populations, or 3.3% (n = 4) for the emergency department. In 22.5% (n = 27) of studies, it was unclear if they included ICU patients.

Most models (48.3%, n = 58) were developed for all ages (≥ 18), 14.2% (n = 17) for patients aged 50 or older, and 20.8% (n = 25) for patients 65 years or older. In 5.8% (n = 7) of the studies, an age limit was used (eg, < 89 years), and in 10.8% (n = 13) the age of the population was not reported.

Delirium outcome definition and determination

In 51 studies (43.5%), the models predicted postoperative delirium (POD). Delirium subgroups such as hypoactive, hyperactive or mixed delirium were distinguished in 5.0% of the studies. In 47 studies (39.1%), a specific time period (eg, within 5 postoperative days) was defined for outcome prediction. In 12 studies (10.0%), the distinction between substance-induced delirium and delirium according to for example ICD-10 code F05 was unclear. In 10 studies (8.3%), the features used for delirium prediction were not listed.

Assessment methods and frequencies varied widely. The Confusion Assessment Method (CAM)[35] including its variations was the most commonly used assessment tool (n = 72, 60.0%). In 29 studies (24.1%) validated delirium screening instruments such as 4 A's Test (4-AT),[36] Intensive Care Delirium Screening Checklist (ICDSC),[37] Delirium Observation Screening Scale (DOS),[38] NEECHAM,[39] Nursing Delirium Scale (NuDeSc),[40] and DRS-R-98[41] were used. Furthermore, 29 (24.1%) studies used a variety of chart review methods, eg, the validated Chart-DEL,[9] and other non-validated chart

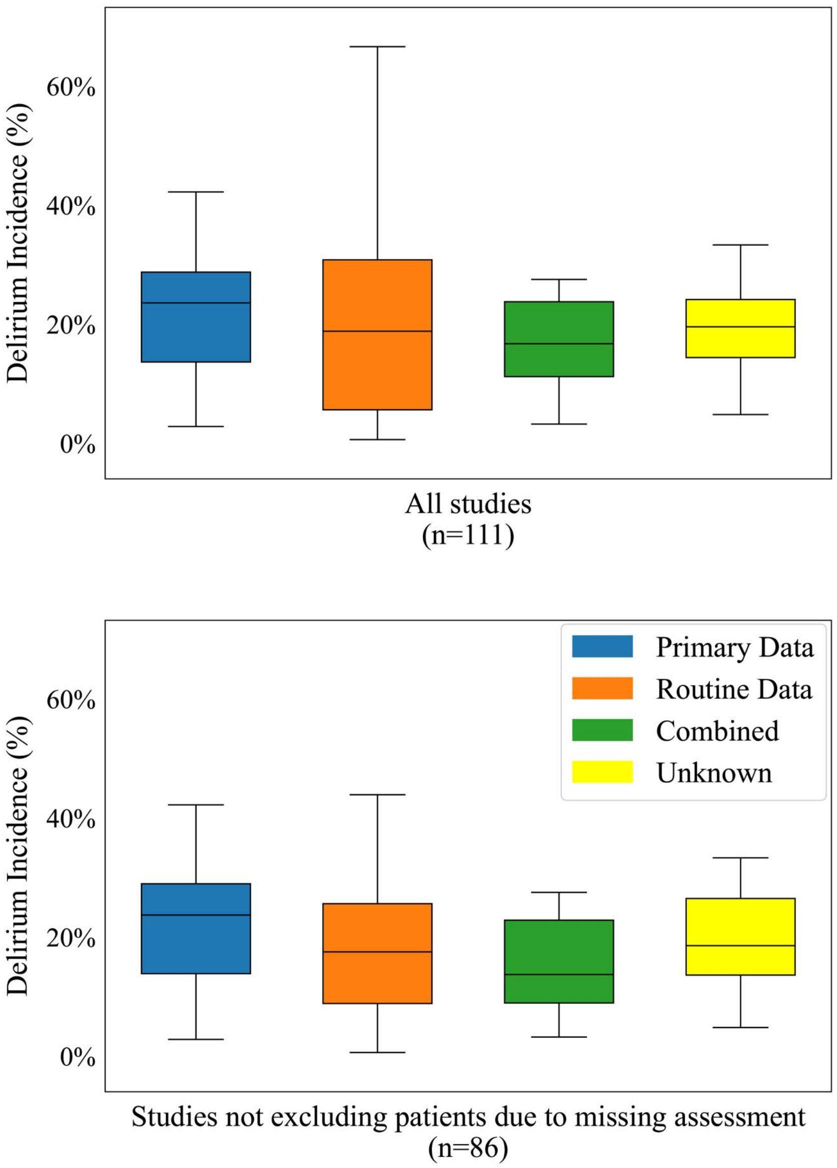

Figure 2. Boxplot showing delirium incidence for all studies and for those including patients regardless of missing assessments. Excludes 9 studies where incidence was not reported or reported in a different format (eg, percentage of positive assessments).

reviews. The Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria[42] were used in 19 (15.8%) studies, which by definition are not a tool but diagnostic criteria. Another 7 (5.8%) studies used delirium-related ICD codes. Additionally 19 studies (15.8%) used other labeling methods (Supplementary Material). In 6 studies, the Richmond Agitation Sedation Scale (RASS)[43] was used to determine whether a screening or assessment instrument could be applied, of which one[44] used it to distinguish delirium subtypes. Of the studies using routine data for labeling, 15 (28%) reported on the fidelity of the screening or assessment tool.

Table 3 provides a complete overview of the different assessment methods and their usage across different label sources.

Delirium incidence and sample size

Table 4 presents the population sizes used to train and test the models, which varied by data source. Routine datasets typically had larger sample sizes with slightly lower incidences as primary datasets.

Modeling methods and features

The use of EHR data in delirium prediction models began in 1996, marking the inception of EHR-based approaches to this task. However, the field has grown exponentially since 2012. A majority (65.0%) of studies used a single modeling approach, with Logistic Regression (LR) being chosen in 92.0% of these cases. Historically dominant in delirium prediction, LR's popularity has declined in recent years in favor of Neural Networks (NN), XGBoost, and Gradient Boosting Machines (GBM). For a comprehensive overview, refer to Table SA3. Detailed trends over time can be found in Table 3.

Label and data leakage

Our review of delirium prediction models found that 10 studies (8.3%) did not mask post-outcome data during model development, while 32 studies (26.7%) did not provide enough information to assess whether they did.

Table 3. Assessment tools and their corresponding labeling strategies.

Delirium assessment method	∑ n (% out of 120)	Routine data	Primary data	Combined	Unknown
CAM[a]	72 (60.0%)	31 (43.1%)	27 (37.5%)	8 (11.1%)	6 (8.3%)
Record review	29 (24.2%)	–	20 (69.0%)	8 (27.6%)	1 (3.4%)
DSM criteria	19 (15.8%)	7 (36.8%)	5 (26.3%)	4 (21.1%)	3 (15.8%)
Other	19 (15.8%)	4 (21.0%)	9 (47.4%)	5 (26.3%)	1 (5.3%)
ICDSC	8 (6.7%)	5 (62.5%)	3 (37.5%)	–	–
Nu-DeSc	8 (6.7%)	3 (37.5%)	3 (37.5%)	2 (25.0%)	–
ICD-Code	7 (5.8%)	6 (85.7%)	–	1 (14.3%)	–
DOS	7 (5.8%)	3 (42.9%)	1 (14.3%)	1 (14.3%)	2 (28.6%)
RASS	6 (5.0%)	3 (50.0%)	3 (50.0%)	–	–
4-AT	4 (3.3%)	1 (25.0%)	1 (25.0%)	1 (25.0%)	1 (25.0%)
NEECHAM	1 (0.8%)	–	1 (100.0%)	–	–
DRS-R-98	1 (0.8%)	–	1 (100.0%)	–	–

In some cases, authors did not report the label origin, leading to the "Unknown" label.
CAM, Confusion Assessment Method; DSM, Diagnostic and Statistical Manual of Mental Disorders; ICDSC, Intensive Care Delirium Screening Checklist; Nu-DeSc, Nursing Delirium Scale; ICD, International Classification of Diseases; DOS, Delirium Observation Screening Scale; RASS, Richmond Agitation Sedation Scale; 4-AT: 4 A's Test; DRS-R-98, Delirium Rating Scale-Revised-98.
[a] CAM, CAM with all variations.

Table 4. Comparison of population sizes and incidences in development datasets between studies using routine data labels and those using study-specific labeling (primary data) or a combination of both (9 studies excluded due to missing information).

	Sample size (incidence)			
Labels from	Min	Max	Mean	Median
Routine data	66 (0.6%)	203 374 (66.6%)	17 477 (21.7%)	6672 (18.8%)
Primary data	87 (2.8%)	57 180 (74.8%)	4284 (23.0%)	627 (23.5%)
Combination	394 (3.2%)	29 756 (27.5%)	4644 (16.6%)	1802 (16.7%)
Unknown	159 (4.8%)	3284 (48.3%)	799 (21.6%)	470 (19.6%)

Model implementation in hospitals

Delirium prediction models were rarely integrated into clinical practice. Only 1.7% of the studies implemented their models in hospitals, and an unclear implementation status in 3.3% of studies. However, half of the studies (51.2%) developed online tools, scores, or nomograms.

Risk of bias

Adequate information to assess the RoB was lacking in 31.7% (n = 38) of the studies, with insufficiency in almost all areas investigated, from missing incidences to unknown delirium assessment. About half of the studies (n = 12) that failed to use appropriate data sources (eg, large sample bias) acknowledged this limitation. Studies using routine data had a higher sample bias excluding patients with missing assessments (38.8%, n = 19), inflating incidence numbers. Among the 44 (36.7%) studies with inadequate outcome determination, 21 did not acknowledge this limitation. Our assessment of the RoB in determining the outcome indicated a trend towards lower RoB when labels were derived from primary data (Figure 4).

Additionally, we found that the RoB for determining the outcome is lower when the incidence is higher (Figure 4). In contrast, studies with unclear label source had a higher incidence and RoB.

A total of 38. 3% (n = 46) of the included studies had a low number of participants with delirium. Of these, 43.5% (n = 20) did not address the issue, while 43.5% (n = 20) recognized it as a limitation and 13.0% (n = 6) mentioned it in their discussion without considering it a limitation.

For a full overview of the RoB for all included papers, refer to Table SB2.

Decision support to find a strategy for determining ground truth

We devised a flowchart outlining various decision points to assist researchers in establishing suitable study designs for delirium prediction and to streamline the process of selecting an optimal study design tailored to specific research objectives and constraints (Figure 5).

Discussion

ICU versus non-ICU patients

In developing delirium prediction models, it is essential to acknowledge the distinct differences between ICU and non-ICU populations with regard to delirium presentations, risk factors, and assessment challenges. The inclusion of mixed populations in 29.2% of reviewed studies (n = 35) suggests developers may have overlooked these variations. Moreover, in 22.5% (n = 27) of studies, it was unclear whether ICU patients were included, reflecting a lack of transparency and consistency in population definitions—factors that can undermine model applicability. To address this, we recommend developing separate models for ICU and non-ICU populations. Given the substantial delirium risk during ICU admission,[4] tailored models for this high-risk group are especially warranted.

Regarding input features, significant differences exist between ICU and non-ICU patients, primarily due to the enhanced monitoring and the presence of unique measurements and events in the ICU. Patients in non-ICU settings often receive less observation and have a lower nurse-to-patient ratio.[45] When assessing delirium, ICU patients are frequently non-verbal or sedated, requiring adapted tools (eg, CAM-ICU vs CAM).

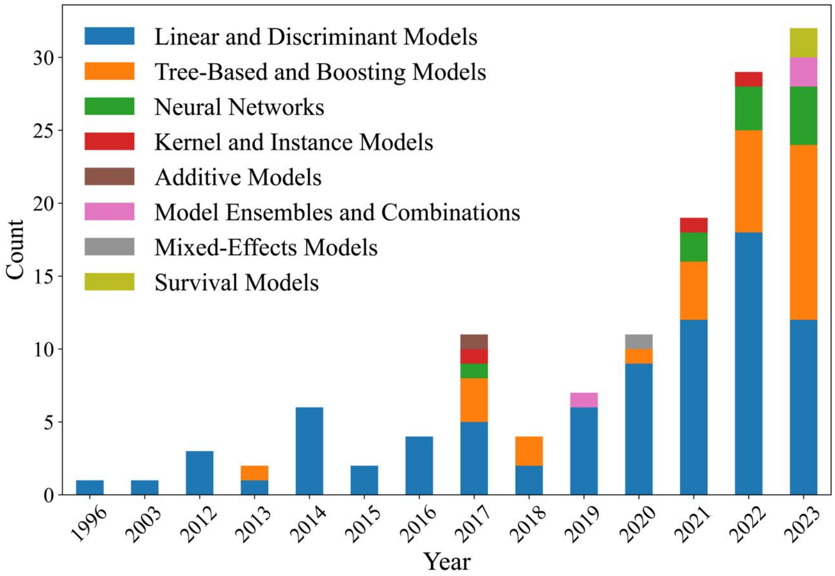

Figure 3. Only the model or final model of each included study, grouped by year. For the breakdown of model grouping, see Table SA3.

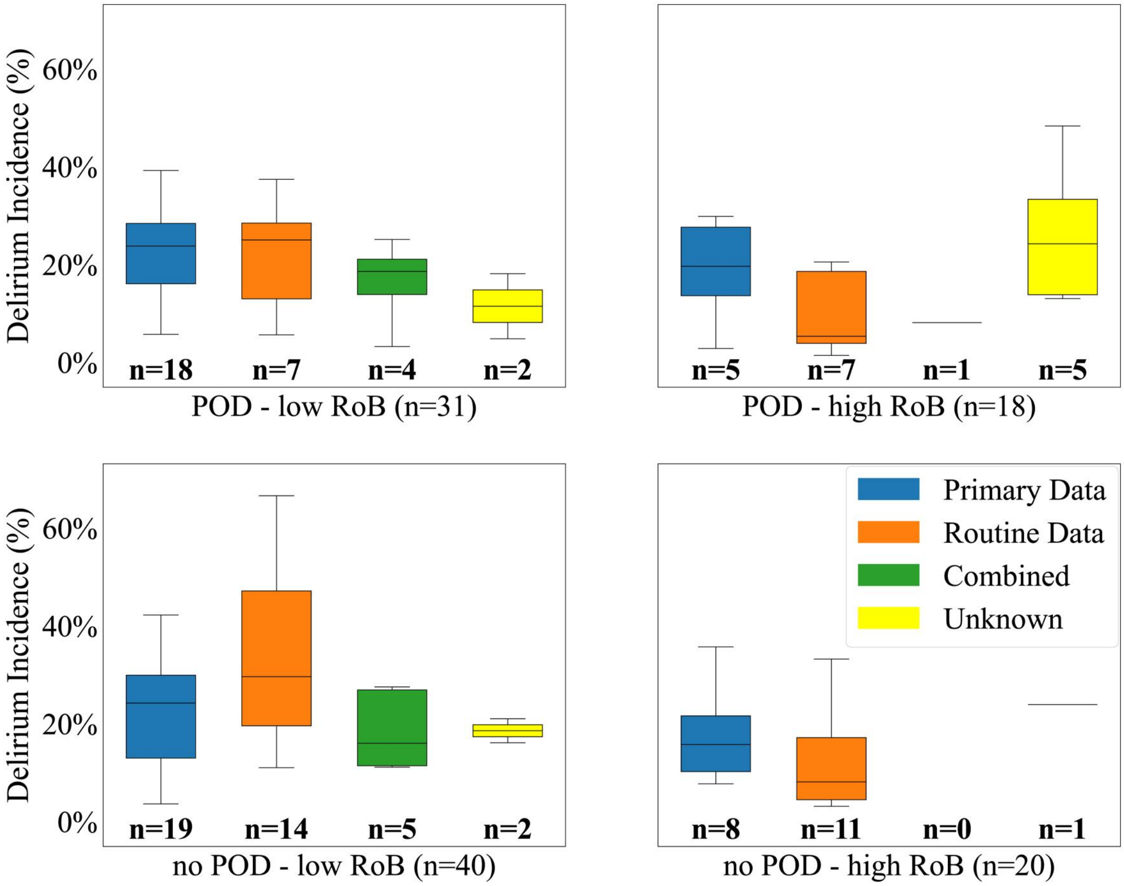

Figure 4. Boxplot displaying the incidence distribution of studies with low or high risk of bias (RoB) in the "determination of outcome" domain, categorized by study design (POD: postoperative delirium). Excludes 9 studies without reported incidences or different formats and 2 with indeterminate bias risk due to missing data.

Generalizability and utility of the models

Implementing a delirium prediction model in hospitals could significantly enhance the awareness, enable early prevention, and improve overall care for these patient populations. The heavy reliance on single-hospital data (78.3%) raises concerns regarding generalizability. This reliance on EHR data underscores the critical dependence on the quality of these records. In contrast, the limited use of registry-based EHR data (9.2%), public datasets (5.0%), and data from other studies (7.5%) suggests an underutilization of more

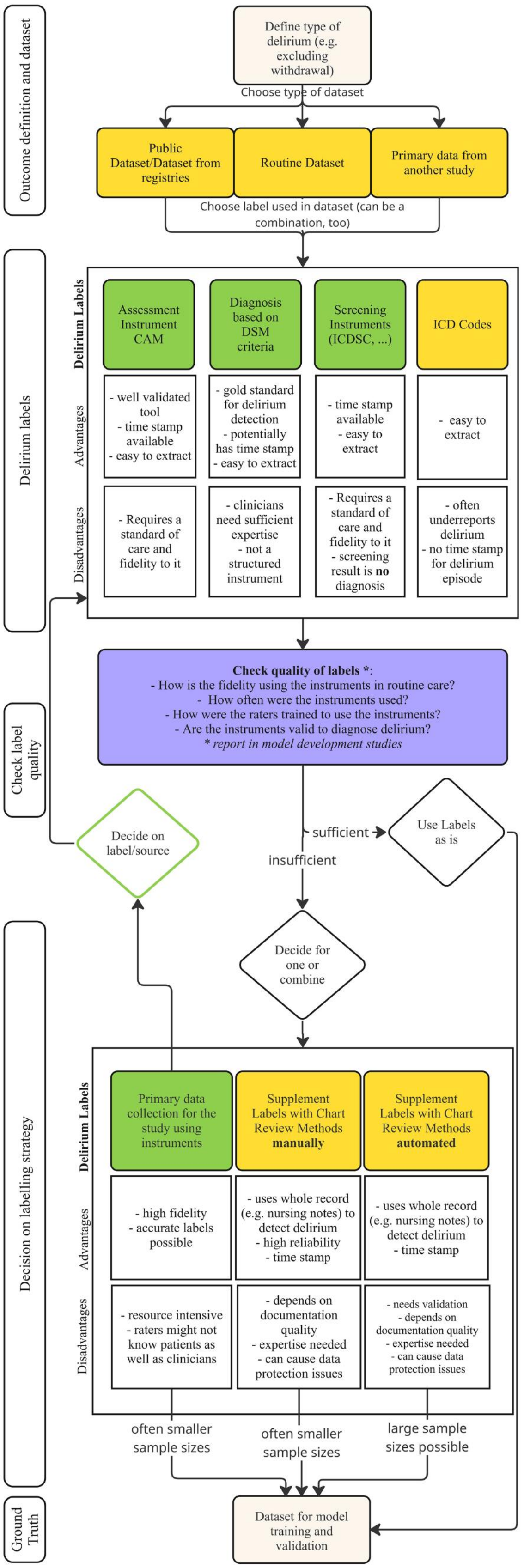
Outcome definition and dataset
Define type of delirium (e.g. excluding withdrawal)
Choose type of dataset
Public Dataset/Dataset from registries
Routine Dataset
Primary data from another study
Choose label used in dataset (can be a combination, too)
Delirium labels
Delirium Labels
Assessment Instrument CAM
Diagnosis based on DSM criteria
Screening Instruments (ICDSC, ...)
ICD Codes
Advantages
- well validated tool
- time stamp available
- easy to extract
- gold standard for delirium detection
- potentially has time stamp
- easy to extract
- time stamp available
- easy to extract
- easy to extract
Disadvantages
- Requires a standard of care and fidelity to it
- clinicians need sufficient expertise
- not a structured instrument
- Requires a standard of care and fidelity to it
- screening result is no diagnosis
- often underreports delirium
- no time stamp for delirium episode
Check label quality
Check quality of labels *:
- How is the fidelity using the instruments in routine care?
- How often were the instruments used?
- How were the raters trained to use the instruments?
- Are the instruments valid to diagnose delirium?
* report in model development studies
Decide on label/source
sufficient
Use Labels as is
insufficient
Decide for one or combine
Decision on labelling strategy
Delirium Labels
Primary data collection for the study using instruments
Supplement Labels with Chart Review Methods manually
Supplement Labels with Chart Review Methods automated
Advantages
- high fidelity
- accurate labels possible
- uses whole record (e.g. nursing notes) to detect delirium
- high reliability
- time stamp
- uses whole record (e.g. nursing notes) to detect delirium
- time stamp
Disadvantages
- resource intensive
- raters might not know patients as well as clinicians
- depends on documentation quality
- expertise needed
- can cause data protection issues
- needs validation
- depends on documentation quality
- expertise needed
- can cause data protection issues
often smaller sample sizes
often smaller sample sizes
large sample sizes possible
Ground Truth
Dataset for model training and validation

Figure 5. Flowchart for selecting a ground truth determination strategy in delirium prediction model development studies.

diverse data sources, which often undergo rigorous quality checks. While single-hospital models may suffice for internal use, our findings stress the necessity for diverse and collaborative approaches in developing delirium prediction models. Such models need to effectively translate across various healthcare settings. However, specialized models tailored to specific hospitals or departments might outperform more general models.[46]

Routine datasets tend to be larger (median 6672) than primary data collections (median 627), but larger size does not guarantee improved quality, as shown in our RoB assessment (Figure 4). Larger datasets often face label quality issues like incorrect labels,[47] which aligns with our findings. This poses a challenge for machine learning (ML) models, typically requiring large datasets for training, as lower quality can compromise their effectiveness. However, as previously reported by Ding et al.,[47] deep learning-based models can tolerate larger amounts of incorrect labels and experience less performance degradation compared to traditional ML methods. Deep learning approaches are scarce due to the limited availability of large, high-quality labeled datasets. Small datasets, often with higher quality labels, are more suitable for simpler models like LR, but are prone to overfitting and random effects, inflating model performance.[48] Balancing model complexity and dataset quality is key. Developing large datasets with high-quality labels is essential for advanced ML models to ensure reliable predictions. The trend towards complex models is evident (Figure 3). Most studies use a single modeling method without clear reasoning or comparison to support its effectiveness. Without evaluating different approaches, it is difficult to determine which method provides the best outcomes, potentially compromising predictive accuracy.

Strategies to determine ground truth

Routinely versus primary collected data

Two primary approaches to labeling delirium are evident: 40.8% of studies relied on routine data, while 42.5% used primary data collected specifically for the study. An advantage of using routine care data is the extended observation period and the familiarity of raters with their patients, allowing them to capture the fluctuating course and brief episodes of delirium (including hypoactive delirium cases[49]) more effectively than the snapshot provided by primary collected data. However, primary collected data, in turn, facilitates easier training for raters to assess patients accurately. Routine data provided larger datasets with slightly lower incidences to those using primary data (Table 4) and potentially higher RoB. Primary data offered higher quality but smaller sample sizes. The surprising relatively small difference in delirium incidence despite differences in RoB prompted further investigation into the data sources. In studies using routine data, we found more than a third excluded patients without an assessment. Excluding those patients can lead to sample bias, particularly in routine data, and may overestimate delirium incidence by selectively capturing higher-risk patients who are more likely to get assessed (Figure 2). This limitation can hinder the generalization of models to lower-risk patients who may be underrepresented in the dataset.

A distinctive approach combines routine data with a primary data collection method, involving chart reviews conducted with validated tools such as Chart-DEL.[9] It ensures robust delirium identification by integrating nursing notes with comprehensive EHR data. However, employing such methods requires trained and experiences raters, whose expertise should be clearly documented. While only 8.3% of studies combined routine and primary data, this approach shows promise after thorough label quality assessments.

Screening and assessment instruments used

Instruments for delirium assessment were utilized in all types of study designs. The key advantage of using instruments is the presence of a time stamp for delirium incidence, enabling masking post-outcome data and predicting its timing. Additionally, direct observations through these instruments consistently outperform other methods, such as chart reviews.[9]

The adoption of the CAM and its variations (60.0%) underscores its clinical validation and widespread acceptance as a diagnostic tool. Nevertheless, a variety of alternative methods were used, including the 4-AT and ICDSC, none of which are validated tools for diagnosing delirium. Moreover, the potential for confusion with other conditions like dementia should be acknowledged as a limitation in their use for outcome detection. Using the CAM requires training, which should be acknowledged and reported in studies. The frequency of instrument use is also a crucial factor to consider, as it directly impacts the likelihood of missing positive cases due to the fluctuating nature of delirium symptoms.

The CAM is a well-validated tool for delirium detection and performs well when used by trained individuals, but the DSM criteria remain the diagnostic reference standard. Applied by senior, specialized professionals (eg, geriatricians or psychiatrists), the DSM criteria is considered the gold standard, especially in studies collecting primary data.

However, in both clinical practice and research, the CAM is a widely used and validated bedside tool for delirium detection, especially where formal DSM-based diagnosis is not feasible. Commonly used by non-specialists, it has high sensitivity and specificity. While sometimes called a gold standard for screening,[50] it is best described as a widely accepted detection tool rather than a definitive diagnostic method.

In routine care, specialists are not always available and are usually consulted only in cases with a very high risk or when symptoms may be confused with those of other conditions. Thus, diagnoses based on the DSM criteria are less frequent than those obtained through tools, such as the CAM, which nurses can use more frequently.

To our knowledge, there is no validated tool to distinguish between delirium subtypes, but assessing a patient's level of consciousness allows to classify the symptoms of the subtypes. Consciousness can be measured using validated tools such as the Glasgow Coma Scale or the RASS. Although the RASS was used in 6 studies, only 1 study utilized it to differentiate between delirium subtypes, showing it is underused. Overall, only 5.0% of models distinguished between delirium subtypes. Training models to identify hypoactive delirium could significantly improve clinical practice by addressing its frequent underdetection.

Assessing fidelity of delirium screening and assessment instruments within routine data was reported in only 28% of studies, complicating the reliability assessment of their labels.

ICD codes

ICD codes are readily available, but potentially underreport delirium[51,52] and are often assigned at discharge without temporal context. This leads to a lack of crucial timestamps for accurate model development excluding post-outcome data.

Delirium label and data leakage

Post-outcome data were masked in only 78 (65.0%) of the included studies. The complexity of healthcare data amplifies the challenges of data leakage in delirium prediction models. Assessments like CAM or applying the DSM criteria provide precise labels at specific time points, unlike labels from ICD codes, allowing masking of post-outcome data.

To prevent inadvertent use of labels in model input, collaboration between technical and clinical teams is crucial. Clear documentation of feature selection and choice of observation and prediction windows is essential. Some studies predict delirium using admission data for the entire stay (static), while others use dynamic prediction throughout hospitalization. Researchers should explicitly mask post-outcome data and rigorously assess label and data leakage potential, as highlighted by Kapoor and Narayanan.[53]

Despite existing reporting guidelines for model development studies,[54] we frequently encountered difficulties in evaluating studies due to insufficient reporting. For instance, 10.0% of the studies reviewed included alcohol withdrawal symptoms as a feature without clearly specifying the type of delirium they predicted, causing doubts in whether they included withdrawal delirium as an outcome. Additionally, 31.7% of studies did not adequately describe how delirium was detected.

Risk of bias

The lack of transparency in reporting was a recurring theme, with more than half (55.8%) of the included studies failing to provide sufficient detail to assess outcome determination. This lack of transparency hindered our ability to accurately evaluate the RoB and undermines confidence in the findings. More than 43% of the studies with a low incidence failed to acknowledge this in their publications. This underscores the importance of our work, emphasizing that a reliable ground truth is crucial in predicting delirium and that data should not be used without careful scrutiny. Similarly, half of the studies relying on inappropriate data sources did not recognize this as a limitation, raising concerns about the reliability of their models.

Overall, our RoB assessment indicates that many studies developing delirium prediction models are susceptible to biases and methodological flaws. These findings highlight the critical need for rigorous methodologies, transparent reporting practices, and careful consideration of potential biases in future research to ensure the development of reliable and generalizable prediction models. To address this, we have developed a flowchart to guide researchers in implementing more robust study designs (Figure 5).

Limitations

A Cohen's kappa of 0.71 indicates a substantial agreement between the 2 raters[55] for the RoB, but the results should be interpreted with caution due to the subjective nature of the RoB assessment. To mitigate this bias, we conducted the assessment jointly with 2 reviewers possessing complementary backgrounds in nursing and computer science. Nevertheless, certain questions, such as "were there a reasonable number of participants with the outcome?," proved challenging to answer, particularly when dealing with special case study populations lacking references for comparison.

Conclusions

This study highlights the challenges of determining an effective delirium labeling strategy in the development of prediction models. While various approaches exist for establishing a ground truth, the optimal method depends heavily on the data available from clinical practice. The underreporting bias, particularly in the field of delirium prediction, warrants special attention. Furthermore, this review emphasizes the inconsistency in reporting practices and aims to provide guidance for establishing a robust ground truth, which is essential for developing reliable prediction models.

Author contributions

Lili M. Schöler (Conceptualization, Data curation, Formal analysis, Visualization, Writing—original draft), Lisa Graf (Conceptualization, Data curation, Formal analysis, Visualization, Writing—original draft), Alexander Ritzi (Writing—review & editing), Michael Simon (Conceptualization, Supervision, Writing—review & editing), Antti Airola (Writing—review & editing), and Laura-Maria Peltonen (Conceptualization, Supervision, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This review paper is part of the KIDELIR[56] project, in collaboration with healthcare professionals, aimed at developing an AI system for delirium prediction. Funding for this research is generously provided by the Bundesministerium für Bildung und Forschung (BMBF) and the Federal Ministry of Education and Research, Germany (funding number 16SV8864). Furthermore, we acknowledge support by the Open Access Publication Fund of the University of Freiburg.

Conflicts of interest

No competing interest is declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

Declaration of generative AI in scientific writing

During the preparation of this work, the author(s) used Open AI's Chat GPT and DeepL in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1. Davis D, Agar M. The DSM-5 criteria, level of arousal and delirium diagnosis: inclusiveness is safer. *BMC Med*. 2014;12:141.
2. Gleason LJ, Schmitt EM, Kosar CM, et al. Effect of delirium and other major complications on outcomes after elective surgery in older adults. *JAMA Surg*. 2015;150:1134-1140.
3. Witlox J, Eurelings LSM, de Jonghe JFM, Kalisvaart KJ, Eikelenboom P, Van Gool WA. Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *JAMA*. 2010;304:443-451.
4. Inouye SK, Westendorp RGJ, Saczynski JS. Delirium in elderly people. *Lancet*. 2014;383:911-922.
5. Xie Q, Wang X, Pei J, et al. Machine learning–based prediction models for delirium: a systematic review and meta-analysis. *J Am Med Dir Assoc*. 2022;23:1655-1668.e6.
6. Hope C, Estrada N, Weir C, Teng C-C, Damal K, Sauer BC. Documentation of delirium in the VA electronic health record. *BMC Res Notes*. 2014;7:208-206.
7. Chuen VL, Chan ACH, Ma J, Alibhai SMH, Chau V. The frequency and quality of delirium documentation in discharge summaries. *BMC Geriatr*. 2021;21:307-310.
8. Bellelli G, Nobili A, Annoni G, REPOSI (REgistro POliterapie SIMI) Investigators, et al. Under-detection of delirium and impact of neurocognitive deficits on in-hospital mortality among acute geriatric and medical wards. *Eur J Intern Med*. 2015;26:696-704.
9. Inouye SK, Leo-Summers L, Zhang Y, Bogardus ST, Jr, Leslie DL, Agostini JV. A chart-based method for identification of delirium: validation compared with interviewer ratings using the Confusion Assessment Method. *J Am Geriatr Soc*. 2005;53:312-318.
10. Wang M, Sushil M, Miao BY, Butte AJ. Bottom-up and top-down paradigms of artificial intelligence research approaches to healthcare data science using growing real-world big data. *J Am Med Inform Assoc*. 2023;30:1323-1332.
11. El Hussein M, Hirst S, Salyers V. Factors that contribute to underrecognition of delirium by registered nurses in acute care settings: a scoping review of the literature to explain this phenomenon. *J Clin Nurs*. 2015;24:906-915.
12. Siddiqi N, Holt R, Britton AM, Holmes J. Interventions for preventing delirium in hospitalised patients. *Cochrane Database Syst Rev*. 2007;Cd005563.
13. Inouye SK. The importance of delirium and delirium prevention in older adults during lockdowns. *JAMA*. 2021;325:1779-1780.
14. Gavinski K, Carnahan R, Weckmann M. Validation of the Delirium Observation Screening Scale in a hospitalized older population. *J Hosp Med*. 2016;11:494-497.
15. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Vol. 1. World Health Organization; 1992.
16. Casey P, Cross W, Webb-St Mart M, Baldwin C, Riddell K, Dārziņš P. Hospital discharge data under-reports delirium occurrence: results from a point prevalence survey of delirium in a major Australian Health Service. *Intern Med J*. 2019;49:338-344.
17. Ruppert MM, Lipori J, Patel S, et al. ICU delirium-prediction models: a systematic review. *Crit Care Explor*. 2020;2:e0296.
18. Lindroth H, Bratzke L, Purvis S, et al. Systematic review of prediction models for delirium in the older adult inpatient. *BMJ Open*. 2018;8:e019223.
19. van Meenen LCC, van Meenen DMP, de Rooij SE, ter Riet G. Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *J Am Geriatr Soc*. 2014;62:2383-2390.
20. Chua SJ, Wrigley S, Hair C, Sahathevan R. Prediction of delirium using data mining: a systematic review. *J Clin Neurosci*. 2021;91:288-298.
21. Von Elm E, Schreiber G, Haupt CC. Methodische anleitung für scoping reviews (JBI-Methodologie). *Zeitschrift Für Evidenz, Fortbildung Und Qualität im Gesundheitswesen*. 2019;143:1-7.

22. Kostopoulou O, Tracey C, Delaney BC. Can decision support combat incompleteness and bias in routine primary care data? *J Am Med Inform Assoc*. 2021;28:1461-1467.
23. Covidence. *Covidence Systematic Review Software, Veritas Health Innovation*. Melbourne, Covidence; 2014. Accessed October 16, 2023. https://www.covidence.org/
24. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377-381.
25. Harris PA, Taylor R, Minor BL, REDCap Consortium, et al. The REDCap consortium: building an international community of software platform partners. *J f Biomed Inform*. 2019;95:103208.
26. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11: e1001744.
27. Wolff RF, Moons KGM, Riley RD, PROBAST Group, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51-58.
28. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data*. 2012;6:1-21.
29. Davis SE, Matheny ME, Balu S, Sendak MP. A framework for understanding label leakage in machine learning for health care. *J Am Med Inform Assoc*. 2024;31:274-280.
30. Yang Y, Wang T, Guo H, et al. Development and validation of a nomogram for predicting postoperative delirium in patients with elderly hip fracture based on data collected on admission. *Front Aging Neurosci*. 2022;14:914002.
31. Liu Y, Shen W, Tian Z. Using machine learning algorithms to predict high-risk factors for postoperative delirium in elderly patients. *Clin Interv Aging*. 2023;18:157-168.
32. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035-160039.
33. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*. 2023;10:219.
34. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data*. 2018;5:180178-180113.
35. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegal AP, Horwitz RI. Clarifying confusion: the confusion assessment method: a new method for detection of delirium. *Ann Intern Med*. 1990;113:941-948.
36. Bellelli G, Morandi A, Davis DHJ, et al. Validation of the 4AT, a new instrument for rapid delirium screening: a study in 234 hospitalised older people. *Age Ageing*. 2014;43:496-502.
37. Bergeron N, Dubois M-J, Dumont M, Dial S, Skrobik Y. Intensive Care Delirium Screening Checklist: evaluation of a new screening tool. *Intensive Care Med*. 2001;27:859-864.
38. Schuurmans MJ, Shortridge-Baggett LM, Duursma SA. The Delirium Observation Screening Scale: a screening instrument for delirium. *Res Theory Nurs Pract*. 2003;17:31-50.
39. Champagne MT, Neelon VJ, McConnell ES, Funk S. The NEECHAM confusion scale: assessing acute confusion in the hospitalized and nursing home elderly. *The Gerontologist*. 1987;27:473-480.
40. Hargrave A, Bourgeois J, Bastiaens J, et al. Validation of a nurse-based delirium screening tool for hospitalized patients (P6.220). *Neurology* 2016;86:P6.220.
41. Trzepacz PT, Mittal D, Torres R, Kanary K, Norton J, Jimerson N. Validation of the Delirium Rating Scale-Revised-98: comparison with the delirium rating scale and the cognitive test for delirium. *J Neuropsychiatry Clin Neurosci*. 2001;13:229-242.
42. Cole MG, Dendukuri N, McCusker J, Han L. An empirical study of different diagnostic criteria for delirium among elderly medical inpatients. *J Neuropsychiatry Clin Neurosci*. 2003;15:200-207.
43. Sessler CN, Gosnell MS, Grap MJ, et al. The Richmond Agitation–Sedation Scale: validity and reliability in adult intensive care unit patients. *Am J Respir Crit Care Med*. 2002;166:1338-1344.
44. de la Varga-Martínez O, Gómez-Pesquera E, Muñoz-Moreno MF, et al. Development and validation of a delirium risk prediction preoperative model for cardiac surgery patients (DELIPRECAS): an observational multicentre study. *J Clin Anesth*. 2021;69:110158.
45. Driscoll A, Grant MJ, Carroll D, et al. The effect of nurse-to-patient ratios on nurse-sensitive patient outcomes in acute specialist units: a systematic review and meta-analysis. *Eur J Cardiovasc Nurs*. 2018;17:6-22.
46. Miller L. Healthcare algorithms don't always need to be generalizable, June 2022. Accessed July 1, 2024. https://hai.stanford.edu/news/healthcare-algorithms-dont-always-need-be-generalizable
47. Ding C, Pereira T, Xiao R, Lee RJ, Hu X. Impact of label noise on the learning based models for a binary classification of physiological signal. *Sensors*. 2022;22:7166.
48. Ying, X. An overview of overfitting and its solutions. *J Phys Conf Ser* 2019;1168:022022.
49. Inouye SK, Foreman MD, Mion LC, Katz KH, Cooney LM, Jr. Nurses' recognition of delirium and its symptoms: comparison of nurse and researcher ratings. *Arch Intern Med*. 2001;161:2467-2473.
50. American Delirium Society. AGS COCARE®: CAM and help tools. Accessed April 1, 2025. https://americandeliriumsociety.org/healthcare-professionals/ags-cocare-cam-and-help-tools/
51. McCoy TH, Jr, Snapper L, Stern TA, Perlis RH. Underreporting of delirium in statewide claims data: implications for clinical care and predictive modeling. *Psychosomatics*. 2016;57:480-488.
52. Bui LN, Pham VP, Shirkey BA, Swan JT. Effect of delirium motoric subtypes on administrative documentation of delirium in the surgical intensive care unit. *J Clin Monit Comput*. 2017;31:631-640.
53. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4:100804.
54. Flanagin A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting use of ai in research and scholarly publication—JAMA Network Guidance. *JAMA*. 2024;331:1096-1098.
55. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. October 2012;22:276-282.
56. KIDELIR. Project funded by the Bundesministerium für Bildung und Forschung, 2022. Accessed October 16, 2023. https://www.interaktive-technologien.de/projekte/kidelir
57. Ahmed N, Kuo YH. Delirium risk in geriatric hip hemi-arthroplasty (DRIGHA): development and validation of a novel score using a national data. *Injury*. 2022;53:1469-1476.
58. Barreto Chang OL, Whitlock EL, Arias AD, et al. A novel approach for the detection of cognitive impairment and delirium risk in older patients undergoing spine surgery. *J Am Geriatr Soc*. 2023;71:227-234.
59. Bhattacharyya A, Sheikhalishahi S, Torbic H, et al. Delirium prediction in the ICU: designing a screening tool for preventive interventions. *JAMIA Open*. 2022;5:ooac048.
60. Bishara A, Chiu C, Whitlock EL, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiol*. 2022;22:8.
61. van den Boogaard M, Pickkers P, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. *BMJ* 2012;344:e420.
62. Böhner H, Hummel TC, Habel U, et al. Predicting delirium after vascular surgery: a model based on pre- and intraoperative data. *Ann Surg*. 2003;238:149-156.

63. Carrasco GM, Villarroel DL, Calderón PJ, Martínez FG, Andrade AM, González TM. [Development and validation of a clinical predictive model for delirium in hospitalized older people]. *Rev Med Chil*. 2014;142:826-832.
64. Carrasco MP, Villarroel L, Andrade M, Calderón J, González M. Development and validation of a delirium predictive score in older people. *Age Ageing*. 2014;43:346-351.
65. Castro VM, Sacks CA, Perlis RH, McCoy TH. Development and external validation of a delirium prediction model for hospitalized patients with coronavirus disease 2019. *J Acad Consult Liaison Psychiatry*. 2021;62:298-308.
66. Chaiwat O, Chanidnuan M, Pancharoen W, et al. Postoperative delirium in critically ill surgical patients: incidence, risk factors, and predictive scores. *BMC Anesthesiol*. 2019;19:39.
67. Chen D, Li Y, Li Q, et al. Risk factors and a nomogram model establishment for postoperative delirium in elderly patients undergoing arthroplasty surgery: a single-center retrospective study. *Biomed Res Int*. 2021;2021:6607386.
68. Chen J, Ji X, Xing H. Risk factors and a nomogram model for postoperative delirium in elderly gastric cancer patients after laparoscopic gastrectomy. *World J Surg Oncol*. 2022;20:319.
69. Chen Y, Du H, Wei BH, Chang XN, Dong CM. Development and validation of risk-stratification delirium prediction model for critically ill patients: a prospective, observational, single-center study. *Medicine (Baltimore)*. 2017;96:e7543.
70. Chen D, Wang W, Wang S, et al. Predicting postoperative delirium after hip arthroplasty for elderly patients using machine learning. *Aging Clin Exp Res*. 2023;35:1241-1251.
71. Cherak SJ, Soo A, Brown KN, Ely EW, Stelfox HT, Fiest KM. Development and validation of delirium prediction model for critically ill adults parameterized to ICU admission acuity. *PLoS One*. 2020;15:e0237639.
72. Choi JY, Yoo S, Song W, Kim S, Baek H, Lee JS, Yoon YS, Yoon S, Lee HY, and Kim KI. Development and validation of a prognostic classification model predicting postoperative adverse outcomes in older surgical patients using a machine learning algorithm: retrospective observational network study. *J Med Internet Res*. 2023;25:e42259.
73. Choi NY, Kim EH, Baek CH, Sohn I, Yeon S, Chung MK. Development of a nomogram for predicting the probability of postoperative delirium in patients undergoing free flap reconstruction for head and neck cancer. *Eur J Surg Oncol*. 2017;43:683-688.
74. Contreras M, Silva B, Shickel B, et al. Dynamic delirium prediction in the intensive care unit using machine learning on electronic health records. *IEEE EMBS Int Conf Biomed Health Inform* 2023;2023:1-5.
75. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of incident delirium using a random Forest classifier. *J Med Syst*. 2018;42:261.
76. Davoudi A, Ebadi A, Rashidi P, Ozrazgat-Baslanti T, Bihorac A, Bursian AC. Delirium prediction using machine learning models on preoperative electronic health records data. *Proc IEEE Int Symp Bioinform Bioeng* 2017;2017:568–573.
77. Dodsworth BT, Reeve K, Falco L, et al. Development and validation of an international preoperative risk assessment model for postoperative delirium. *Age Ageing*. 2023;52: afad086.
78. Douglas VC, Hessler CS, Dhaliwal G, et al. The AWOL tool: derivation and validation of a delirium prediction rule. *J Hosp Med*. 2013;8:493-499.
79. Eschweiler GW, Czornik M, Herrmann ML, et al. Presurgical screening improves risk prediction for delirium in elective surgery of older patients: the PAWEL RISK study. *Front Aging Neurosci*. 2021;13:679933.
80. Fan H, Ji M, Huang J, et al. Development and validation of a dynamic delirium prediction rule in patients admitted to the intensive care units (DYNAMIC-ICU): a prospective cohort study. *Int J Nurs Stud*. 2019;93:64-73.
81. Gong KD, Lu R, Bergamaschi TS, et al. Predicting intensive care delirium with machine learning: model development and external validation. *Anesthesiology*. 2023;138:299-311.
82. Gu Q, Yang S, Fei D, Lu Y, Yu H. A nomogram for predicting sepsis-associated delirium: a retrospective study in MIMIC III. *BMC Med Inform Decis Mak*. 2023;23:184.
83. Guo Y, Ji H, Liu J, et al. Development and validation of a delirium risk prediction model for elderly patients undergoing elective orthopedic surgery. *Neuropsychiatr Dis Treat*. 2023;19:1641-1654.
84. Guo R, Zhang S, Yu S, et al. Inclusion of frailty improved performance of delirium prediction for elderly patients in the cardiac intensive care unit (d-FRAIL): a prospective derivation and external validation study. *Int J Nurs Stud*. 2023;147:104582.
85. Haight TN, Marsh EB. Identifying delirium early after stroke: a new prediction tool for the intensive care unit. *J Stroke Cerebrovasc Dis*. 2020;29:105219.
86. Harasawa N, Mizuno T. A novel scale predicting postoperative delirium (POD) in patients undergoing cerebrovascular surgery. *Arch Gerontol Geriatr*. 2014;59:264-271.
87. Hata M, Miyazaki Y, Nagata C, et al. Predicting postoperative delirium after cardiovascular surgeries from preoperative portable electroencephalography oscillations. *Front Psychiatry*. 2023;14:1287607.
88. He J, Ling Q, Chen Y. Construction and application of a model for predicting the risk of delirium in postoperative patients with type A aortic dissection. *Front Surg*. 2021;8:772675.
89. Heinrich M, Woike JK, Spies CD, Wegwarth O. Forecasting postoperative delirium in older adult patients with fast-and-frugal decision trees. *J Clin Med*. 2022;11:5629.
90. Hornor MA, Ma M, Zhou L, et al. Enhancing the American College of Surgeons NSQIP surgical risk calculator to predict geriatric outcomes. *J Am Coll Surg*. 2020;230:88-100.e1.
91. Hu Y, Yang M. A predictive scoring system for postoperative delirium in the elderly patients with intertrochanteric fracture. *BMC Surg*. 2023;23:154.
92. Hu XY, Liu H, Zhao X, et al. Automated machine learning-based model predicts postoperative delirium using readily extractable perioperative collected electronic data. *CNS Neurosci Ther*. 2022;28:608-618.
93. Huang HW, Zhang GB, Li HY, et al. Development of an early prediction model for postoperative delirium in neurosurgical patients admitted to the ICU after elective craniotomy (E-PRE-POD-NS): a secondary analysis of a prospective cohort study. *J Clin Neurosci*. 2021;90:217-224.
94. Huang W, Wu Q, Zhang Y, et al. Development and validation of a nomogram to predict postoperative delirium in type b aortic dissection patients underwent thoracic endovascular aortic repair. *Front Surg*. 2022;9:986185.
95. Hur S, Ko RE, Yoo J, Ha J, Cha WC, Chung CR. A machine learning-based algorithm for the prediction of intensive care unit delirium (PRIDE): retrospective study. *JMIR Med Inform*. 2021;9:e23401.
96. Ida M, Takeshita Y, Kawaguchi M. Preoperative serum biomarkers in the prediction of postoperative delirium following abdominal surgery. *Geriatr Gerontol Int*. 2020;20:1208-1212.
97. Isfandiaty R, Harimurti K, Setiati S, Roosheroe AG. Incidence and predictors for delirium in hospitalized elderly patients: a retrospective cohort study. *Acta Med Indones*. 2012;44:290-297.
98. Jeong YM, Lee E, Kim KI, et al. Association of pre-operative medication use with post-operative delirium in surgical oncology patients receiving comprehensive geriatric assessment. *BMC Geriatr*. 2016;16:134.
99. Jung JW, Hwang S, Ko S, et al. A machine-learning model to predict postoperative delirium following knee arthroplasty using electronic health records. *BMC Psychiatry*. 2022;22:436.
100. O'Keeffe ST, Lavan JN. Predicting delirium in elderly patients: development and validation of a risk-stratification model. *Age Ageing*. 1996;25:317-321.

101. Kennedy M, Enander RA, Tadiri SP, Wolfe RE, Shapiro NI, Marcantonio ER. Delirium risk prediction, healthcare use and mortality of elderly adults in the emergency department. *J Am Geriatr Soc*. 2014;62:462-469.
102. Kim EM, Li G, Kim M. Development of a risk score to predict postoperative delirium in patients with hip fracture. *Anesth Analg*. 2020;130:79-86.
103. Kim MY, Park UJ, Kim HT, Cho WH. DELirium Prediction based on Hospital Information (DELPHI) in general surgery patients. *Medicine (Baltimore)*. 2016;95:e3072.
104. Kim MK, Oh J, Kim JJ, Park JY. Development and validation of simplified delirium prediction model in intensive care unit. *Front Psychiatry*. 2022;13:886186.
105. Kobayashi D, Takahashi O, Arioka H, Koga S, Fukui T. A prediction rule for the development of delirium among patients in medical wards: Chi-Square Automatic Interaction Detector (CHAID) decision tree analysis model. *Am J Geriatr Psychiatry*. 2013;21:957-962.
106. 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;236:32-39.
107. Krzych LJ, Wybraniec MT, Krupka-Matuszczyk I, Skrzypek M, Bochenek AA. Delirium Screening in Cardiac Surgery (DESCARD): a useful tool for nonpsychiatrists. *Can J Cardiol*. 2014;30:932-939.
108. Lapp L, Roper M, KavanaghK, Schraag S. Predicting the onset of delirium on hourly basis in an intensive care unit following cardiac surgery. In: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE; 2022:234–239.
109. Lee DY, Oh AR, Park J, et al. Machine learning-based prediction model for postoperative delirium in non-cardiac surgery. *BMC Psychiatry*. 2023;23:317.
110. Li GH, Zhao L, Lu Y, et al. Development and validation of a risk score for predicting postoperative delirium after major abdominal surgery by incorporating preoperative risk factors and surgical Apgar score. *J Clin Anesth*. 2021;75:110408.
111. Li B, Ju J, Zhao J, Qin Y, Zhang Y. A nomogram to predict delirium after hip replacement in elderly patients with femoral neck fractures. *Orthop Surg*. 2022;14:3195-3200.
112. Li Q, Zhao Y, Chen Y, Yue J, Xiong Y. Developing a machine learning model to identify delirium risk in geriatric internal medicine inpatients. *Eur Geriatr Med*. 2022;13:173-183.
113. Ling YT, Guo QQ, Wang SM, et al. Nomogram for prediction of postoperative delirium after deep brain stimulation of subthalamic nucleus in Parkinson's disease under general anesthesia. *Parkinsons Dis*. 2022;2022:6915627.
114. Liu S, Schlesinger JJ, McCoy AB, et al. New onset delirium prediction using machine learning and long short-term memory (LSTM) in electronic health record. *J Am Med Inform Assoc*. 2022;30:120-131.
115. Lucini FR, Stelfox HT, Lee J. Deep learning-based recurrent delirium prediction in critically ill patients. *Crit Care Med*. 2023;51:492-502.
116. Lucini FR, Fiest KM, Stelfox HT, Lee J. Delirium prediction in the intensive care unit: a temporal approach. *Ann Int Conf IEEE Eng Med Biol Soc (EMBC),* 2020;2020:5527-5530.
117. Marra A, Pandharipande PP, Shotwell MS, et al. Acute brain dysfunction: development and validation of a daily prediction model. *Chest*. 2018;154:293-301.
118. Martinez JA, Belastegui A, Basabe I, et al. Derivation and validation of a clinical prediction rule for delirium in patients admitted to a medical ward: an observational study. *BMJ Open*. 2012;2: e001599.
119. Matsumoto K, Nohara Y, Sakaguchi M, et al. Temporal generalizability of machine learning models for predicting postoperative delirium using electronic health record data: model development and validation study. *JMIR Perioper Med* 2023;6:e50895.
120. Matsuoka A, Miike T, Miyazaki M, et al. Development of a delirium predictive model for adult trauma patients in an emergency and critical care center: a retrospective study. *Trauma Surg Acute Care Open*. 2021;6:e000827.
121. Menzenbach J, Kirfel A, Guttenthaler V, PROPDESC Collaboration Group, et al. PRe-Operative prediction of postoperative DElirium by appropriate SCreening (PROPDESC) development and validation of a pragmatic POD risk screening score based on routine preoperative data. *J Clin Anesth*. 2022;78:110684.
122. Moon KJ, Jin Y, Jin T, Lee SM. Development and validation of an automated delirium risk assessment system (Auto-DelRAS) implemented in the electronic health record system. *Int J Nurs Stud*. 2018;77:46-53.
123. Mueller B, Street WN, Carnahan RM, Lee S. Evaluating the performance of machine learning methods for risk estimation of delirium in patients hospitalized from the emergency department. *Acta Psychiatr Scand*. 2023;147:493-505.
124. Mufti HN, Hirsch GM. Perioperative prediction of agitated (hyperactive) delirium after cardiac surgery in adults – the development of a practical scorecard. *J Crit Care*. 2017;42:192-199.
125. Nagata C, Hata M, Miyazaki Y, et al. Development of postoperative delirium prediction models in patients undergoing cardiovascular surgery using machine learning algorithms. *Sci Rep*. 2023;13:21090.
126. Nakamizo T, Kanda T, Kudo Y, et al. Development of a clinical score, PANDA, to predict delirium in stroke care unit. *J Neurol Sci*. 2020;415:116956.
127. Neto PCS, Rodrigues AL, Stahlschmidt A, Helal L, Stefani LC. Developing and validating a machine learning ensemble model to predict postoperative delirium in a cohort of high-risk surgical patients: a secondary cohort analysis. *Eur J Anaesthesiol*. 2023;40:356-364.
128. Oberai T, Oosterhoff JHF, Woodman R, Doornberg JN, Kerkhoffs G, Jaarsma R. Development of a postoperative delirium risk scoring tool using data from the australian and New Zealand hip fracture registry: an analysis of 6672 patients 2017-2018. *Arch Gerontol Geriatr*. 2021;94:104368.
129. Oldenbeuving AW, de Kort PL, van Eck van der Sluijs JF, Kappelle LJ, Roks G. An early prediction of delirium in the acute phase after stroke. *J Neurol Neurosurg Psychiatry*. 2014;85:431-434.
130. Oliveira J E Silva L, Stanich JA, Jeffery MM, et al. REcognizing DElirium in geriatric emergency medicine: the REDEEM risk stratification score. *Acad Emerg Med*. 2022;29:476-485.
131. Oosterhoff JHF, Karhade AV, Oberai T, Franco-Garcia E, Doornberg JN, Schwab JH. Prediction of postoperative delirium in geriatric hip fracture patients: a clinical prediction model using machine learning algorithms. *Geriatr Orthop Surg Rehabil*. 2021;12:21514593211062277.
132. Pasinska P, Kowalska K, Klimiec E, et al. Poststroke delirium clinical motor subtypes: the PRospective observational POLIsh study (PROPOLIS). *J Neuropsychiatry Clin Neurosci*. 2019;31:104-111.
133. Racine AM, Tommet D, D'Aquila ML, the RISE Study Group, et al. Machine learning to develop and internally validate a predictive model for post-operative delirium in a prospective, observational clinical cohort study of older surgical patients. *J Gen Intern Med*. 2021;36:265-273.
134. Ren Y, Zhang Y, Zhan J, et al. Machine learning for prediction of delirium in patients with extensive burns after surgery. *CNS Neurosci Ther* 2023;29:2986-2997.
135. Rössler J, Shah K, Medellin S, et al. Development and validation of delirium prediction models for noncardiac surgery patients. *J Clin Anesth*. 2024;93:111319.
136. Schulthess-Lisibach AE, Gallucci G, Benelli V, et al. Predicting delirium in older non-intensive care unit inpatients: development and validation of the DELIrium risK tool (DELIKT). *Int J Clin Pharm*. 2023;45:1118-1127.

137. Sheikhalishahi S, Bhattacharyya A, Celi LA, Osmani V. An interpretable deep learning model for time-series electronic health records: case study of delirium prediction in critical care. *Artif Intell Med*. 2023;144:102659.

138. Shen J, An Y, Jiang B, Zhang P. Derivation and validation of a prediction score for postoperative delirium in geriatric patients undergoing hip fracture surgery or hip arthroplasty. *Front Surg*. 2022;9:919886.

139. Shi Y, Wang H, Zhang L, et al. Nomogram models for predicting delirium of patients in emergency intensive care unit: a retrospective cohort study. *Int J Gen Med*. 2022;15:4259-4272.

140. Song YX, Yang XD, Luo YG, et al. Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: a retrospective study. *CNS Neurosci Ther*. 2023;29:158-167.

141. Spiller TR, Tufan E, Petry H, et al. Delirium screening in an acute care setting with a machine learning classifier based on routinely collected nursing data: a model development study. *J Psychiatr Res*. 2022;156:194-199.

142. Sweerts L, Hoogeboom TJ, van Wessel T, van der Wees PJ, van de Groes SAW. Development of prediction models for complications after primary total hip and knee arthroplasty: a single-centre retrospective cohort study in The Netherlands. *BMJ Open*. 2022;12:e062065.

143. Tamai K, Terai H, Nakamura H, et al. Delirium risk score in elderly patients with cervical spinal cord injury and/or cervical fracture. *J Clin Med*. 2023;12:2387.

144. Tian Y, Liang Y, Chen Y, Bian H. Analysis of delirium prediction in the ICU based on the hybrid SGDCS-ANFIS approach. *Med Biol Eng Comput*. 2023;61:673-683.

145. Tian Y, Ji B, Diao X, et al. Dynamic predictive scores for cardiac surgery-associated agitated delirium: a single-center retrospective observational study. *J Cardiothorac Surg*. 2023;18:219.

146. Tyson B, Shahein A, Erdodi L, et al. Delirium as a presenting symptom of COVID-19. *Cogn Behav Neurol*. 2022;35:123-129.

147. Vacas S, Grogan T, Cheng D, Hofer I. Risk factor stratification for postoperative delirium: a retrospective database study. *Medicine (Baltimore)*. 2022;101:e31176.

148. Veeranki SPK, Hayn D, Jauk S, et al. An improvised classification model for predicting delirium. *Stud Health Technol Inform*. 2019;264:1566-1567.

149. Visser L, Prent A, van der Laan MJ, et al. Predicting postoperative delirium after vascular surgical procedures. *J Vasc Surg*. 2015;62:183-189.

150. Wang J, Ji Y, Wang N, et al. Establishment and validation of a delirium prediction model for neurosurgery patients in intensive care. *Int J Nurs Pract*. 2020;26:e12818.

151. Wang L, Chignell M, Zhang Y, et al. Physician experience design (PXD): more usable machine learning prediction for clinical decision making. *AMIA Summits Transl Sci Proc*. 2022;2022:476-485.

152. Wang XQ, Zhuang HX, Zhang LX, Chen X, Niu CS, Zhao M. Nomogram for predicting postoperative delirium after deep brain stimulation surgery for Parkinson's disease. *World Neurosurg* 2019;130:e551-e557.

153. Wang Y, Lei L, Ji M, Tong J, Zhou CM, Yang JJ. Predicting postoperative delirium after microvascular decompression surgery with machine learning. *J Clin Anesth*. 2020;66:109896.

154. Wang ML, Kuo YT, Kuo LC, et al. Early prediction of delirium upon intensive care unit admission: model development, validation, and deployment. *J Clin Anesth*. 2023;88:111121.

155. Wassenaar A, van den Boogaard M, van Achterberg T, et al. Multinational development and validation of an early prediction model for delirium in ICU patients. *Intensive Care Med*. 2015;41:1048-1056.

156. Whitlock EL, Braehler MR, Kaplan JA, et al. Derivation, validation, sustained performance, and clinical impact of an electronic medical record-based perioperative delirium risk stratification tool. *Anesth Analg*. 2020;131:1901-1910.

157. de Wit HA, Winkens B, Mestres Gonzalvo C, et al. The development of an automated ward independent delirium risk prediction model. *Int J Clin Pharm*. 2016;38:915-923.

158. 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:e181018.

159. Xiang D, Xing H, Zhu Y. A predictive nomogram model for postoperative delirium in elderly patients following laparoscopic surgery for gynecologic cancers. *Support Care Cancer*. 2022;31:24.

160. Xing H, Zhou W, Fan Y, Wen T, Wang X, Chang G. Development and validation of a postoperative delirium prediction model for patients admitted to an intensive care unit in China: a prospective study. *BMJ Open*. 2019;9:e030733.

161. Xu Y, Meng Y, Qian X, et al. Prediction model for delirium in patients with cardiovascular surgery: development and validation. *J Cardiothorac Surg*. 2022;17:247.

162. Xue X, Chen W, Chen X. A novel radiomics-based machine learning framework for prediction of acute kidney injury-related delirium in patients who underwent cardiovascular surgery. *Comput Math Methods Med*. 2022;2022:4242069.

163. Xue B, Li D, Lu C, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. *JAMA Netw Open*. 2021;4:e212240.

164. Zhan L, Wang XQ, and Zhang LX. Nomogram model for predicting risk of postoperative delirium after deep brain stimulation surgery in patients older than 50 years with Parkinson disease. *World Neurosurgery*. 2020;139:e127-e135.

165. Zhang Y, Wan DH, Chen M, et al. Automated machine learning-based model for the prediction of delirium in patients after surgery for degenerative spinal disease. *CNS Neurosci Ther*. 2023;29:282-295.

166. Zhang Y, Hu J, Hua T, Zhang J, Zhang Z, Yang M. Development of a machine learning-based prediction model for sepsis-associated delirium in the intensive care unit. *Sci Rep*. 2023;13:12697.

167. Zhang H, Yuan J, Chen Q, et al. Development and validation of a predictive score for ICU delirium in critically ill patients. *BMC Anesthesiol*. 2021;21:37.

168. Zhang S, Ji MH, Ding S, et al. Inclusion of interleukin-6 improved performance of postoperative delirium prediction for patients undergoing coronary artery bypass graft (POD-CABG): a derivation and validation study. *J Cardiol*. 2022;79:634-641.

169. Zhang X, Tong DK, Ji F, et al. Predictive nomogram for postoperative delirium in elderly patients with a hip fracture. *Injury*. 2019;50:392-397.

170. Zhao X, Li J, Xie X, et al. Online interpretable dynamic prediction models for postoperative delirium after cardiac surgery under cardiopulmonary bypass developed based on machine learning algorithms: a retrospective cohort study. *J Psychosom Res*. 2024;176:111553.

171. Zhao H, You J, Peng Y, Feng Y. Machine learning algorithm using electronic chart-derived data to predict delirium after elderly hip fracture surgeries: a retrospective case-control study. *Front Surg*. 2021;8:634629.

172. Zheng YB, Ruan GM, Fu JX, Su ZL, Cheng P, Lu JZ. Postoperative plasma 8-iso-prostaglandin f2α levels are associated with delirium and cognitive dysfunction in elderly patients after hip fracture surgery. *Clinica Chimica Acta*. 2016;455:149-153.

173. Zucchelli A, Apuzzo R, Paolillo C, et al. Development and validation of a delirium risk assessment tool in older patients admitted to the emergency department observation unit. *Aging Clin Exp Res*. 2021;33:2753-2758.

JAMIA Open, 2025, 8, 1–14
https://doi.org/10.1093/jamiaopen/ooaf037
Review