BMJ Open

Head-to-head comparison of 14 prediction models for postoperative delirium in elderly non-ICU patients: an external validation study

Chung Kwan Wong ,[1] Barbara C van Munster,[1] Athanasios Hatseras,[1] Else Huis in 't Veld,[1] Barbara L van Leeuwen,[2] Sophia E de Rooij,[1] Rick G Pleijhuis[3]

To cite: Wong CK, van Munster BC, Hatseras A, *et al*. Head-to-head comparison of 14 prediction models for postoperative delirium in elderly non-ICU patients: an external validation study. *BMJ Open* 2022;**12**:e054023. doi:10.1136/bmjopen-2021-054023

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-054023).

Received 02 June 2021
Accepted 23 February 2022

[1]Department of Geriatrics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[2]Department of Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[3]Department of Internal Medicine, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

Correspondence to
Dr. Rick G Pleijhuis;
r.g.pleijhuis@umcg.nl

ABSTRACT

Objectives Delirium is associated with increased morbidity, mortality, prolonged hospitalisation and increased healthcare costs. The number of clinical prediction models (CPM) to predict postoperative delirium has increased exponentially. Our goal is to perform a head-to-head comparison of CPMs predicting postoperative delirium in non-intensive care unit (non-ICU) elderly patients to identify the best performing models.

Setting Single-site university hospital.

Design Secondary analysis of prospective cohort study.

Participants and inclusion CPMs published within the timeframe of 1 January 1990 to 1 May 2020 were checked for eligibility (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). For the time period of 1 January 1990 to 1 January 2017, included CPMs were identified in systematic reviews based on prespecified inclusion and exclusion criteria. An extended literature search for original studies was performed independently by two authors, including CPMs published between 1 January 2017 and 1 May 2020. External validation was performed using a surgical cohort consisting of 292 elderly non-ICU patients.

Primary outcome measures Discrimination, calibration and clinical usefulness.

Results 14 CPMs were eligible for analysis out of 366 full texts reviewed. External validation was previously published for 8/14 (57%) CPMs. C-indices ranged from 0.52 to 0.74, intercepts from −0.02 to 0.34, slopes from −0.74 to 1.96 and scaled Brier from −1.29 to 0.088. Based on predefined criteria, the two best performing models were those of Dai *et al* (c-index: 0.739; (95% CI: 0.664 to 0.813); intercept: −0.018; slope: 1.96; scaled Brier: 0.049) and Litaker *et al* (c-index: 0.706 (95% CI: 0.590 to 0.823); intercept: −0.015; slope: 0.995; scaled Brier: 0.088). For the remaining CPMs, model discrimination was considered poor with corresponding c-indices <0.70.

Conclusion Our head-to-head analysis identified 2 out of 14 CPMs as best-performing models with a fair discrimination and acceptable calibration. Based on our findings, these models might assist physicians in postoperative delirium risk estimation and patient selection for preventive measures.

Strengths and limitations of this study

- ► This study encompasses the largest head-to-head comparison of clinical prediction models (CPM) for predicting postoperative delirium in elderly non-intensive care unit patients reported to date.
- ► Prospectively collected data were reused for validation purposes.
- ► Model variables not available in the dataset were substituted for using equivalent variables if available.
- ► Evaluated performance measures included both classical statistical metrics and decision curve analysis, providing a solid basis for model comparison.
- ► Identification of eligible CPMs during a 30-year time span was partly based on previously published systematic reviews, which might have resulted in relevant CPMs being overlooked.
- ► The total number of events in the dataset was limited to 25 patients with delirium, resulting in limited power for included prediction models with a large number of variables.

INTRODUCTION

Delirium is a mental disorder, characterised by an acute fluctuating disturbance in awareness and attention accompanied by cognitive deficits such as memory, orientation, language and perception. It is typically caused by an underlying disturbance, such as infection or electrolyte imbalances. Delirium is also common in the postoperative period following (major) surgery and is associated with increased morbidity, mortality and prolonged hospitalisation increasing healthcare costs.[1 2] It is also known to reduce long-term cognitive function, even years after the patient was discharged from the hospital.[2] Incidence of postoperative delirium reported in the literature ranges from 10% to 70%, depending on the type of surgery performed, characteristics of the patient population and criteria used for the diagnosis.[3]

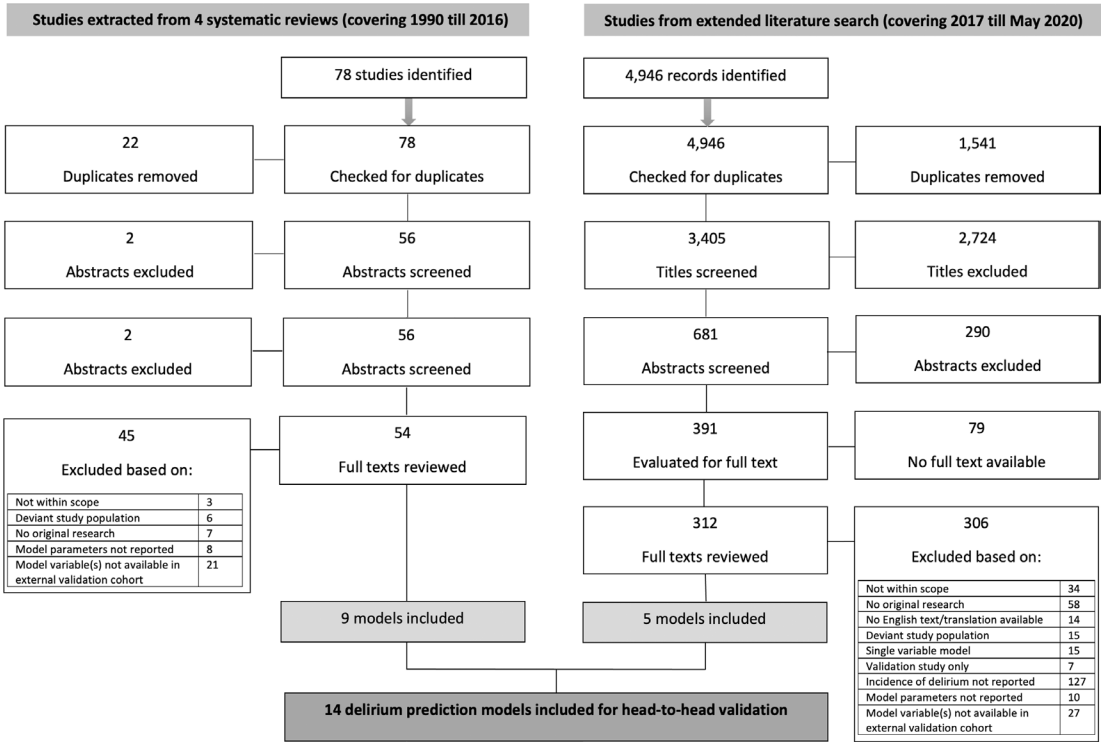

Figure 1 Flow chart indicating the selection process of included delirium prediction models.

Importantly, delirium is considered preventable in up to 40% of the cases.[4 5] For example, non-pharmacological multicomponent interventions were shown effective in reducing the risk of developing delirium in the elderly.[6] Yet, these interventions can be time-consuming and costly, limiting widespread application when resources are scarce. Adequate stratification of delirium risk is of great importance to make sure preventive interventions are provided to those patients expected to benefit most from them. Clinical prediction models (CPMs) can be applied for these purposes.[7] Over the past few decades, the number of CPMs to predict postoperative delirium has steeply increased and continues to do so. CPMs can support patient selection for preventive measures by differentiating low-risk and high-risk patients based on the presence of risk factors associated with delirium. In order for CPMs to be used in a safe and responsible manner, it is of utmost importance that information on the development and overall performance of these models is made available to clinicians. Yet, this information is often lacking.[8] For example, the majority of published CPMs for postoperative delirium have not been externally validated, therewith lacking essential information on model robustness and generalisability. Even when external validations have been performed, they are usually based on different patient cohorts, hampering direct comparison of model performance between CPMs.

The aim of this study is to perform a head-to-head comparison of discriminative power, calibration and clinical utility of previously published CPMs to predict postoperative delirium in non-intensive care unit (non-ICU) elderly patients. For this purpose, we used a single prospectively obtained validation cohort to externally validate multiple CPMs simultaneously.

METHODS

Literature search

An extensive literature search was performed to identify CPMs eligible for external validation. The search comprised two parts, conducted separately.

First, we searched the MEDLINE database for systematic reviews focusing on the prediction of postoperative delirium in elderly non-ICU patients. Systematic reviews published between 1 January 1990 and 1 May 2020 that fulfilled the selection criteria (online supplemental figure 1) were selected. We then extracted all CPMs deemed eligible for validation based on inclusion and exclusion criteria from the systematic reviews, followed by removal of any duplicates.

Second, an extended literature search was performed using the MEDLINE database to cover the time periods not previously taken into account by systematic reviews. Search terms were carefully selected with support of a clinical librarian. A detailed overview of the final search strategy used for the extended search is provided in online supplemental figure 2.

Selection of studies

Studies were selected by two investigators (AH, EHV) who evaluated all search results independently in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. Discrepancies between both investigators were solved by a third and fourth

evaluator (CKW, RP). The same inclusion and exclusion criteria were applied to studies extracted from systematic reviews as well as studies identified through the extended search.

Studies were included when: (i) a prediction model was developed with predicted risk of delirium as the primary outcome and (ii) the primary focus was on elderly hospitalised surgical or mixed (surgical and medical) patients (defined as mean age ≥60 years). Studies were excluded when: (i) the primary focus was on deviant patient populations: delirium tremens, dementia, stroke, psychiatric disorders, acute kidney injury, shock, palliative phase, non-surgical (ie, medical) or intensive care patients, (ii) only external validation of a previously published model was performed, (iii) no prediction model was developed (ie, risk factors reported only or prediction based on a single variable), (iv) the published prediction model could not be reconstructed due to incomplete reporting of model parameters, (v) no full-text article or English translation was available, (vi) the study was considered non-original research and (vii) no incidence of delirium was reported. Finally, CPMs eligible for external validation were selected based on the availability of required model variables in the external validation cohort.

External validation cohort

External validation was performed using an independent dataset (Performance of Risk stratification Instruments for postoperative DElirium; PRIDE cohort) as previously described.[9] In brief, we used an independent dataset comprising 292 elderly hospitalised patients who underwent various surgical procedures between 1 October 2011 and 1 June 2012 in the University Medical Center Groningen, The Netherlands, previously described by Jansen *et al*.[9] This dataset, containing prospectively obtained data, was used to externally validate multiple eligible prediction models simultaneously.

Minor deviations between original model variables and variables available in the dataset were resolved by using substitute parameters. CPMs with more than one variable missing in the validation cohort were excluded. In case of a single missing dichotomous variable, the CPM was still included and a sensitivity analysis was performed.

Patient and public involvement

During this study, there was no direct involvement of the public or patients in the design, conduct or reporting of the research. The results of this study are expected to enhance patient involvement, facilitating shared decision-making.

Head-to-head evaluation of overall model performance

To judge the selected CPMs on their merits and compare them head-to-head, key performance measures were evaluated regarding model discrimination, calibration and clinical usefulness.[10]

Model discrimination

Discrimination refers to the ability of the CPM to distinguish patients who develop delirium from those who do not develop delirium. Model discrimination is expressed as the area under the receiver operating characteristic curve, or 'c-index', which plots the sensitivity (true-positive rate) against 1–specificity (false-positive rate) for consecutive cut-offs of the predicted risk. Perfect discrimination gives a c-index of 1, and no discrimination (no better than the toss of a coin) results in a c-index of 0.50. Prediction models with c-indices between 0.9 and 0.99 are considered to have excellent discrimination, 0.8 and 0.89 good discrimination, 0.7 and 0.79 fair discrimination and 0.51 and 0.69 poor discrimination.[11]

Model calibration

Calibration refers to the agreement between predicted and observed risk.[12] It can be assessed graphically in a plot with predicted probabilities on the x-axis and the proportion of observed risk (delirium present or absent) on the y-axis (figure 1). Perfect predictions should be located on the reference line, described with an intercept of 0 and a slope of 1, indicating that predicted and observed outcomes are alike. The intercept compares the mean of all predicted risks with the mean observed risk. This parameter hence indicates the extent that predictions are systematically too low or too high. The slope is a measure of spread of predicted probabilities.[13]

(Scaled) Brier score

The Brier score is a composite measure based on the mean square error of predictions, assessing both discrimination and calibration.[14] It can be used to compare performance between CPMs predicting binary outcomes (ie, delirium present or absent), with lower scores indicating superior models. A Brier score of 0 represents a perfect model. Scaled Brier scores were calculated to take the baseline incidence of delirium into account, facilitating result interpretation.

Clinical usefulness

When assessing predictive value, although of importance, traditional statistical metrics as discrimination and calibration are not directly informative with regard to clinical value. As a means to overcome these limitations, Vickers and Elkin introduced the concept of decision curve analysis, providing a more holistic understanding of the clinical relevance of CPMs.[15] In brief, decision curve analysis calculates a clinical 'net benefit' for CPMs in comparison to default strategies of imposing an intervention for all or no patients.[16] Net benefit is calculated across a range of threshold probabilities, defined as the minimum probability of disease at which further intervention would be warranted.

Statistical analysis

Continuous baseline characteristics are presented as mean and SD in the case of normally distributed data, whereas skewed data are presented as median and IQR.

Table 1 Baseline characteristics of external validation cohort

Characteristic	No. of patients (n=292)
Gender, n (%)	
Men	175 (60)
Age, mean (SD), years	66 (8)
Age category (years), n (%)	
50–59	75 (26)
60–69	128 (44)
70–79	69 (24)
>80	20 (7)
APACHE II, median (IQR), points	5 (4–7)
Number of comorbidities, n (%)	
0–1	82 (28)
>2	210 (72)
Type of comorbidities, n (%)	
Diabetes mellitus	54 (19)
Hypertension	109 (37)
Other cardiovascular disease	87 (30)
Cerebrovascular disease	28 (10)
Other neurological disease	12 (4)
Chronic pulmonary disease	35 (12)
Chronic renal disease	20 (7)
Medication use, n (%)	
0–1	144 (49)
>4	148 (51)
History of delirium, n (%)	41 (14)
Cognitive impairment*, n (%)	40 (14)
Admission type, n (%)	
Elective	264 (90)
Emergency	28 (10)
Type of surgery, n (%)	
Oncological	93 (32)
General	84 (29)
Vascular	54 (18)
Hepatobiliary	39 (13)
Other	22 (8)
Length of stay, median (IQR), days	8 (4–14)
Postoperative delirium, n (%)	25 (9)

Categorical variables are expressed as the total number of patients with corresponding percentages between brackets. Continuous variables are expressed as median values with IQR unless specified otherwise.
APACHE, Acute Physiology, Age, Chronic Health Evaluation.

We used multiple performance measures to evaluate model performance based on previously published recommendations for reporting on external validation studies.[10] These included: calibration plot (calibration-in-the-large) and model intercept, calibration slope, discrimination with concordance statistic and clinical usefulness with decision curve analysis.

As recommended by Steyerberg *et al*,[12] we used the scaled Brier score as a combined measure of model discrimination and calibration instead of the goodness-of-fit (Hosmer-Lemeshow) test.[17 18]

Sensitivity and specificity rates were calculated for all models. Negative and positive predictive values strongly depend on delirium incidence and were therefore not reported.

Calculations were performed semi-automatically using R-based validation software V.2.18 (available at https://www.evidencio.org).[19] Differences in discriminative power between CPMs were assessed by comparing area under the curves using MedCalc V.20.015.

Handling of missing data

Missing data were reported separately for each model. A complete case analysis was performed without using imputation techniques. CPMs were excluded if >30% of the patients in the validation cohort had missing data for either one or more of the variables included in the CPM.

Sensitivity analysis

To evaluate how sensitive CPM outputs are to changes in inputs, a sensitivity analysis was performed for variables in the validation cohort that could not be implemented exactly as described in the original studies. Analyses were repeated by changing the specific variables to extreme values, to investigate their impact on model performance.

Assessment of quality

For quality assessment of included articles, we used the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) 20-item checklist for prediction model development.[20] The following checklist items were deemed not applicable and were therefore excluded: details of treatments received, actions to blind assessment of outcome and actions to blind assessment of predictors.

RESULTS

Literature search

A total of four systematic reviews were identified that reported on CPMs for postoperative delirium in elderly non-ICU patients.[9 21–23] Altogether, the systematic reviews covered a time period from 1 January 1990 to 1 January 2017 with partial overlap. After removal of duplicates, 56 unique studies were further evaluated, resulting in 30 CPMs for validation after application of exclusion criteria (figure 1). Model variables required for validation were available in the validation cohort for 9 out of 30 (30%) CPMs, leaving 9 CPMs extracted from systematic reviews eligible for external validation.

The extended systematic search, covering a time period from 1 January 2017 to 1 May 2020 not covered by previously published systematic reviews, resulted in 3405 titles

Table 2 Overview of study populations, diagnostic instruments and model variables for all models included for analysis

Study	Population	Mean age	Diagnostic instrument	Variables
Ten Broeke *et al*.,[32]	Cardiac surgery	68	DOSS	Age (continuous variable) Comorbidity SMMSE score <23 History of delirium
Carrasco *et al*.,[27]	Mixed	78	CAM	BUN/Creatinine ratio Barthel index
Dai *et al*[34]	Orthopaedic/Urological surgery	72.7	DSM-IV, CMMSE (MMSE translated to Chinese)	Old age (>80 years) Cognitive impairment (MMSE <24) Psychoactive drugs used
Ettema *et al*[45]	Mixed	81	DOSS	Previous confusion Needed help in self-care Memory problems (three question Dutch screening for delirium)
Freter *et al*,[46]	Orthopaedic surgery	76.8	CAM	Cognitive impairment (MMSE <24) Substance use Age >80 years Dependence in >1 ADL Sensory impairment
Halladay *et al*.,[31]	Mixed	75	DSM-IV	Cognitive impairment (prior diagnosis of dementia in the EMR or outpatient prescription of a medication for dementia at admission) Age (continuous variable) Infection Fracture Visual impairment
Kim *et al*,[28]	Mixed surgical (>50% open)	NR	ICDSC	Age (continuous variable) Physical activity Alcoholism Hearing impairment History of delirium Emergency surgery Open surgery ICU admission CRP (mg/dL)
Litaker *et al*,[35]	Mixed surgical (>50% orthopaedic)	67	CAM, DSM-IV	Age >70 years History of delirium Pre-existing cognitive impairment (TICS <30) Self-report: alcohol affected health Preoperative use of narcotic analgesics Admission to neurosurgery service
Pendlebury *et al*.,[30]	Mixed	81	CAM, DSM-V	Dementia/Cognitive impairment (AMTS <9 or MMSE <24) Age (continuous variable) Severe illness Infection Vision impairment
Pompei *et al*,[24]	Mixed	74.3	CAM, DSM-III	Cognitive impairment (MMSE cut-off 21–24, based on education level) Number of MDCs (comorbidity) Depression Alcoholism
Rudolph *et al*,[25]	Cardiac surgery	74.7	CAM	MMSE <23 History of stroke/TIA GDS >4 Abnormal albumin
Rudolph *et al*.,[26]	Mixed	72.1	DSM-IV-TR	Cognitive impairment (MOCA ≤18) Age >65 years Infection Fracture Vision Severe illness

Continued

Table 2 Continued

Study	Population	Mean age	Diagnostic instrument	Variables
de Wit *et al*,[29]	Mixed	76.9	NS	Age (continuous variable) Polypharmacy ATC-5th Anxiolytics Antidementia Antidepressants Antiparkinson medication Antidiabetics Psychotropics Analgetics Sleep medication CRP (mg/L) Urea (mmol/L)
Zhang *et al*,[33]	Orthopaedic surgery	79	DSM-V	Preoperative cognitive impairment (not defined in original article) Number of medical comorbidities ASA class Transfusion >2 units of RBCs ICU stay

ADL, activities of daily life; AMTS, Abbreviated Mental Test Score; ASA, American Association of Anesthesiologists; CAM(-ICU), confusion assessment method (for the intensive care unit); DOSS, Delirium Observation Screening Scale; DSM, Diagnostic and Statistical Manual for Mental Disorders; EMR, electronic medical record; FE, femoral endarterectomy; GDS, Geriatric Depression Scale; ICDSC, Intensive Care Delirium Screening Checklist; MDC, major diagnostic categories; RBCs, red blood cells; SMMSE, Standardised Mini Mental State Examination; TIA, transient ischaemic attack; TICS, Telephone Interview For Cognitive Status.

after removal of duplicates. A total of 391 abstracts were selected for review of full texts. No full text was available for 79 abstracts, leaving 312 articles for further evaluation. After application of exclusion criteria (detailed in figure 1), the extended search resulted in 32 additional CPMs suitable for validation. Model variables required for validation were available in the validation cohort for 5 out of 32 (16%) CPMs, leaving 5 CPMs identified through the extended search eligible for external validation.

External validation

Baseline characteristics of the external validation cohort are shown in table 1. In brief, the cohort consisted of 292 elderly hospitalised patients with a mean age of 66 years (SD ±8 years). All patients underwent surgery (general, oncological, vascular, hepatobiliary or 'other'), of which the vast majority (90%) concerned elective procedures. A total of 25 patients (9%) developed delirium postoperatively.

An overview of all CPM variables (table 2) and their matched variables from the validation cohort (online supplemental table 1) is provided. Risk factors most frequently used in the included CPMs were increasing age and pre-existing impaired cognition. External validation was previously published for 8 out of 14 (57%) CPMs (table 3). In all cases, c-indices of previously externally validated CPMs were higher compared with our findings.

Overall performance of clinical prediction models for postoperative delirium

Head-to-head evaluation of overall model performance was assessed for 14 included CPMs simultaneously (table 3). Calculated c-indices ranged from 0.52 to 0.74, intercepts from −0.02 to 0.34, slopes from −0.74 to 1.96, Brier scores from 0.07 to 0.22 and scaled Brier scores from −1.29 to 0.088.

For the vast majority (12 out of 14) of included CPMs, model discrimination was considered poor with corresponding c-indices <0.70 (figure 2). Model calibration and clinical usefulness for all included CPMs are represented graphically as calibration plots and clinical decision curves, respectively, in online supplemental figure 3. A positive net benefit was observed in the 5%–20% and 10%–30% threshold probability range for CPMs developed by Dai *et al* and Litaker *et al*, respectively, suggesting superiority to the 'treat none' strategy at these thresholds. For Ettema *et al*, positive net benefit was observed in the 10%–15% threshold probability. CPMs developed by Pompei *et al*,[24] Rudolph *et al*,[25 26] Carrasco *et al*,[27] Kim *et al*,[28] de Wit *et al*,[29] Pendlebury *et al*,[30] Halladay *et al*,[31] Ten Broeke *et al*[32] and Zhang *et al*[33] showed limited net benefit.

The two best performing CPMs were Dai *et al* (c-index: 0.739; 95% CI: 0.664 to 0.813; intercept: −0.018; slope: 1.96; Brier score: 0.077, scaled Brier score: 0.049) and Litaker *et al* (c-index: 0.706; 95% CI: 0.590 to 0.823; intercept: −0.015; slope: 0.995; Brier score: 0.074, scaled Brier score: 0.088) (table 3). Graphical representations of discrimination, calibration and clinical usefulness of both models are shown in figure 3.

On secondary analysis, there was no significant difference in model performance between the CPMs developed by Dai *et al* and Litaker *et al*. Yet, the discriminative power of Dai *et al* significantly differed from almost 50% of all included CPMs. Direct comparison of model discrimination between the remaining 12 CPMs showed no significant difference (online supplemental table 2).

Table 3 Performance of the included clinical prediction models on external model validation

Study	Sample size derivation cohort	Reported TRIPOD items	Validation previously performed	C-statistic previous validation	Sample size validation cohort	C-statistic current validation (95% CI)	Δ C-statistic (%)	Intercept	Slope	Brier score	Scaled Brier score
Pompei *et al*.[24]	432	17/20	Yes	0.64	281	0.543 (0.441 to 0.645)	0.097 (15)	0.069	0.211	0.086	−0.059
Dai *et al*.[34]	701	12/20	No	NA	283	0.739 (0.664 to 0.813)	NA	−0.018	1.96	0.077	0.049
Litaker *et al*.[35]	500	14/20	No	NA	282	0.706 (0.590 to 0.823)	NA	−0.015	0.995	0.074	0.088
Freter *et al*.[46]	132	16/20	No	NA	282	0.576 (0.472 to 0.680)	NA	0.045	0.267	0.093	−0.148
Rudolph *et al*.[25]	122	16/20	Yes	0.75	167	0.610 (0.485 to 0.734)	0.14 (19)	0.002	0.249	0.220	−1.289
Carrasco *et al*.[27]	374	18/20	Yes	0.78	268	0.563 (0.435 to 0.692)	0.217 (28)	0.340	−0.743	0.144	−0.834
Kim *et al*.[28]	561	17/20	Yes	0.94	206	0.610 (0.505 to 0.715)	0.33 (35)	0.018	0.309	0.124	−0.410
Rudolph *et al*.[26]	27 625	17/20	Yes	0.74	231	0.624 (0.504 to 0.743)	0.116 (16)	0.054	0.638	0.080	−0.018
de Wit *et al*.*[29]	1291	18/20	Yes	0.77	206	0.635 (0.501 to 0.769)	0.135 (18)	0.076	0.592	0.092	−0.045
Pendlebury et al.[30]	308	19/20	Yes	0.81	219	0.539 (0.424 to 0.654)	0.271 (33)	0.039	0.329	0.088	−0.062
Ettema *et al*.**[45]	3786	17/20	No	NA	281	0.580 (0.478 to 0.683)	NA	0.070	0.225	0.08	−0.020
Halladay *et al*.[31]	27 625	18/20	Yes	0.91	227	0.519 (0.412 to 0.626)	0.391 (43)	0.063	0.295	0.095	−0.181
Ten Broeke *et al*.[32]	329	18/20	No	NA	283	0.635 (0.521 to 0.749)	NA	0.037	0.219	0.114	−0.419
Zhang *et al*.[33]	825	16/20	No	NA	282	0.650 (0.541 to 0.759)	NA	0.024	0.258	0.117	−0.45

Overview of model performance on external validation of all 14 eligible clinical prediction models for delirium, expressed as model discrimination (C-statistic with corresponding 95% CI), calibration (model intercept and slope) and a composite measure of discrimination and calibration (Brier score and scaled Brier score).
*Full model as reported by de Wit *et al*.[29]
†Three-question model as reported by Ettema *et al*.[45]
NA, not available; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

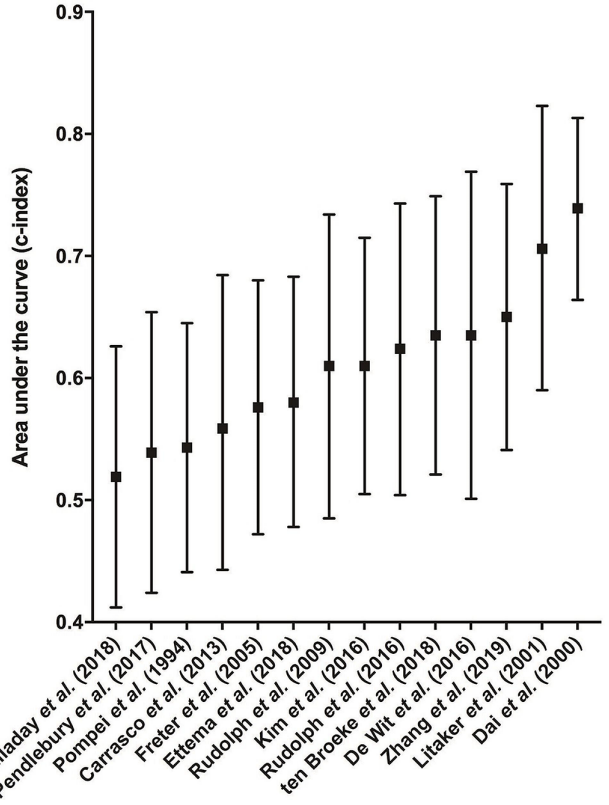

Figure 2 Head-to-head comparison of discriminative power of delirium prediction models. Discriminative power of externally validated delirium prediction models is reported as c-indices with associated 95% CIs, ranked from low to high. A c-index of 0.5 resembles a situation in which the model has no discriminative power, that is, the model predicts no better than flipping a coin. Only 2 out of 14 validated models showed fair discrimination with c-indices >0.70 (0.71 and 0.74 for the models developed by Litaker *et al* and Dai *et al*, respectively) and 95% CIs with lower bounds >0.50. Discriminative power of the remaining 12 models was considered poor.

Sensitivity analysis

In case no matching variable was available in our database, analyses were repeated by using substitute variables to inquire the possible dependency of results on the definition of the risk factors. This was performed for the following variables: activities of daily living, infection, comorbidities, severity of illness and memory problems. No significant differences in CPM performance were observed for any of the variables (data not shown). In case of minute differences in CPM performance between different substitute variables, the variable resulting in the best overall CPM performance was ultimately selected.

DISCUSSION

The goal of this study was to identify clinical prediction models for delirium developed and published since 1990 and to compare their performance head-to-head. Overall, we identified 62 CPMs that were developed for predicting postoperative delirium risk over the last 30 years, of which 14 (23%) could be externally validated using our independent cohort. As studies comparing similar models head-to-head are lacking, caregivers find themselves confronted with the difficult task to select the best-suited model from the great variety of models available. In our study, the two best performing models were those of Dai *et al* and Litaker *et al*, with c-indices of 0.739 and 0.706, respectively, regarded as adequate discrimination.[34 35] Both models showed acceptable calibration, sufficiently stratifying patients in different risk groups. Based on these findings, these two models were considered most promising for guiding patient selection for preventive measures out of 14 evaluated models.

Risk factors for delirium have been studied extensively in the past few decades, hence a multitude of identified risk factors exist for delirium in hospitalised elderly.[36–38] Clinical prediction models based on these risk factors provide an integrated approach in delirium risk estimation. Many CPMs for delirium were developed in specific niche populations which may hamper their generalisability and applicability in daily practice in the overall hospitalised populations. For example, many models contained highly specific biomarkers that are not readily available in most hospitals. In addition, instruments used to determine cognitive impairment often differed between models, making a direct comparison challenging. This was reflected by our finding that only 14 out of 62 studies could be validated despite the fact that we made use of a prospectively collected patient database containing over 200 distinct variables.

Preoperative stratification of patients based on estimated risk for postoperative delirium could identify those patients expected to benefit most from preventive measures. Although there is no conclusive evidence that different drugs are effective in preventing delirium,[39] the evidence for non-pharmacological multicomponent interventions is considered sufficiently robust for clinical practice recommendations in elderly non-ICU patients.[6] In healthcare institutions, applying multicomponent non-pharmacological measures to *all* patients would result in a high burden on scarce human and material resources. The labour-intensive and costly nature of multicomponent interventions requires appropriate selection of patients who are expected to benefit most from such interventions or for whom certain interventions could be omitted. In addition, prediction models can be used to inform patients regarding their individual risk to develop postoperative delirium, providing a solid basis for shared decision-making.

Assessment of model performance in external validation cohort

There is broad consensus that CPMs must be validated in independent patient cohorts prior to clinical application. In reality, external validation studies are often lacking, as was also the case for 6 out of 14 (43%) CPMs included in our study. A possible explanation might be that the information provided in model development

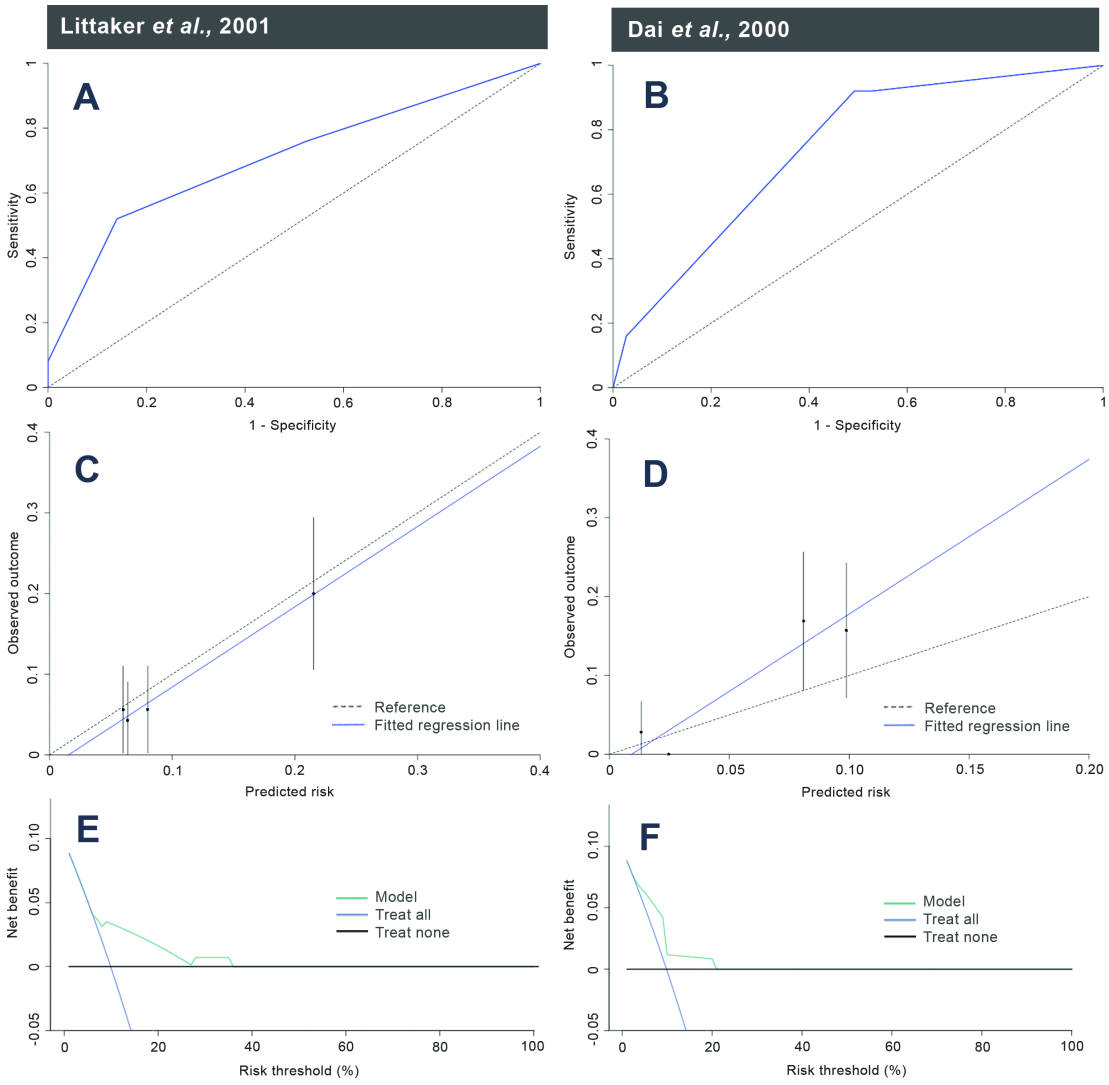

Figure 3 Discrimination, calibration and clinical utility of best performing models. Panels A and B show the receiver operating characteristic (ROC) curve of the delirium prediction models by Litaker *et al* and Dai *et al*, respectively, with the area under the ROC curve (c-index) indicating the discriminative power of the model. A graphical representation of the calibration of both models is shown in panels C and D, plotting the predicted probability (x-axis) with corresponding 95% CI against the actually observed occurrence of delirium in the validation cohort (y-axis). The model by Litaker *et al* showed adequate calibration (panel C), correctly differentiating patients at low risk of delirium (20%). The model by Dai *et al* correctly identified patients at low risk (20%). Panels E and F show decision curve analyses as a measure of clinical utility of both models. For the models by Litaker *et al* and Dai *et al*, a positive net benefit was observed in the 10%–35% threshold probability range (panel E) and the 5%–20% threshold probability range (panel F), respectively.

studies is often lacking specific details (eg, an intercept in the case of logistic regression analysis) to reproduce the original model formula. Other reasons might be the time-intensive nature of external validation and the apparent tendency to develop new CPMs rather than evaluating existing ones.

Methodological guidelines recommend external validation in terms of discrimination and calibration to assess model robustness and generalisability.[10]

We found that the discriminative power determined in the original studies was higher in all cases (Δ c-indices ranging from 0.116 to 0.391) compared with our validation despite the nature of our study population, consisting solely of postoperative patients. A possible explanation is the tendency of overfitting in the case of narrow validation when the same database is (partly) used for model derivation and validation purposes.

Although assessment of calibration performance is an important measure to interpret CPM performance in addition to model discrimination, it has generally received little attention. As shown by Calster *et al*, poorly calibrated CPMs can be misleading and potentially harmful for clinical decision-making.[40] In our current study, model calibration was assessed for all included CPMs and compared head-to-head.

In addition to conventional statistical performance measures, there is a growing interest in the use of decision curve analysis to evaluate net clinical benefit of CPMs

in clinical practice.[15] Decision curve analysis incorporates the consequences of the decisions made on the basis of a CPM, regarding impact on utilities, costs and harms. It is therefore considered a direct measure of clinical value.[16] In the case of delirium risk prediction, a false-positive result (ie, patient falsely stratified as high-risk) is usually not harmful to the patient. False-negative outcomes (ie, patient falsely stratified as low-risk), however, could result in withholding adequate preventive measures for delirium development, exposing the patient to an increased risk for medical complications, prolonged hospitalisation and long-term adverse effects.[2] The medical, psychological and economic effects of false-negative results are therewith considered to outweigh those of false-positive results.

Guidelines for transparent reporting

To be able to adequately assess potential usefulness and risk of bias of prediction models, full and clear reporting of information on all aspects are a prerequisite.[41] Yet, multiple reviews concluded that reporting of model development is poor with insufficient information described in all aspects, from descriptions of patient data to statistical modelling methods.[41–43] In response, a collaborative network of international experts developed methodological guidelines, like the TRIPOD checklist, to facilitate accurate, complete and transparent reporting.[44] In this study, we noticed an improving trend of overall quality of reporting of studies since the introduction of the TRIPOD statement in 2015, although this finding was only based on 14 publications.

Enhancing clinical applicability of CPMs by using an online platform

To facilitate clinical application, published prediction models are sometimes made available as a digital calculator through a website or mobile device app, usually dedicated to a single model. As a result, the landscape of digital calculators is highly fragmented. This confronts healthcare professionals with new challenges. An example includes the current lack of transparency, that is, lack of insight in underlying model formulas, source codes or characteristics of the derivation cohort, turning digital calculators into a 'black box'. Another challenge is how to ascertain their quality and performance when external validations or head-to-head comparisons are not available. The current lack of standardisation results in limited scalability as well as relatively high costs for hosting and updating digital calculators. Indeed, multiple examples exist of websites and apps that are no longer supported or even withdrawn several years after the initial project funding ceased to exist. Even when a prediction model meets all the above-mentioned requirements, this is still no guarantee that the model is actually applied in clinical practice. To facilitate clinical implementation, prediction models should be easily accessible in the clinical workflow, that is, integrated in the electronic health record system, digital protocols or decision support systems. The current variation in prediction models made available through different websites and apps, however, hampers (scalable) possibilities for integration.

To address abovementioned issues, we made use of an existing cloud-based platform that facilitates the standardised creation, head-to-head comparison and integration of CPMs (https://www.evidencio.org). After identifying the best performing CPMs in a given target population, an intuitive user interface can be added automatically to facilitate CPM use (online supplemental figure 4). In addition, direct integration of CPMs in the clinical workflow (ie, electronic health record system) is expected to further increase their impact on clinical decision-making.[7] Before the CPMs evaluated in the current study can be generally applied in a clinical setting, however, further validations in different cohorts are encouraged to further consolidate our findings in terms of model robustness and generalisability in non-surgical populations.

CONCLUSION

Over the last few decades, the number of CPMs developed to predict postoperative delirium has increased exponentially. Overall reproducibility was limited due to the requirement of specific variables not commonly available in daily practice and a lack of reported details to reconstruct model formulas. Nearly half of the CPMs included in our study had previously not been validated in an independent cohort. Our head-to-head analysis of 14 CPMs identified two best-performing models with a fair discrimination and acceptable calibration. Corresponding clinical usefulness was considered promising based on decision curve analysis. Based on our findings, these models might assist physicians in postoperative delirium risk estimation and selection of elderly non-ICU patients for preventive measures, although further validations in different cohorts are encouraged.

Acknowledgements The authors are thankful to Carolien Jansen and clinical librarian Sjoukje van der Werf for granting permission to use the PRIDE validation cohort and providing support in conducting the systematic search, respectively.

Contributors BCvM, SDR and RGP designed the study. AH and EHiV conducted the systematic search supervised by CKW and RGP. CKW and RGP externally validated selected clinical prediction models. BLvL provided data for external validation. CKW, RGP, AH and EHiV wrote the manuscript. BCvM, BLvL and SDR critically revised the manuscript. CKW submitted the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. RGP accepts full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The PRIDE database is available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been

ORCID iD

Chung Kwan Wong http://orcid.org/0000-0001-5746-2888

REFERENCES

1 Haley MN, Casey P, Kane RY, *et al*. Delirium management: let's get physical? A systematic review and meta-analysis. *Australas J Ageing* 2019;38:231–41.
2 Maldonado JR. Acute brain failure: pathophysiology, diagnosis, management, and sequelae of delirium. *Crit Care Clin* 2017;33:461–519.
3 Schenning KJ, Deiner SG. Postoperative delirium in the geriatric patient. *Anesthesiol Clin* 2015;33:505–16.
4 Inouye SK, Bogardus ST, Charpentier PA, *et al*. A multicomponent intervention to prevent delirium in hospitalized older patients. *N Engl J Med* 1999;340:669–76.
5 Marcantonio ER, Flacker JM, Wright RJ, *et al*. Reducing delirium after hip fracture: a randomized trial. *J Am Geriatr Soc* 2001;49:516–22.
6 Salvi F, Young J, Lucarelli M, *et al*. Non-pharmacological approaches in the prevention of delirium. *Eur Geriatr Med* 2020;11:71–81.
7 Adibi A, Sadatsafavi M, Ioannidis JPA. Validation and utility testing of clinical prediction models: time to change the approach. *JAMA* 2020;324:235-236.
8 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2014;350:g7594.
9 Jansen CJ, Absalom AR, de Bock GH, *et al*. Performance and agreement of risk stratification instruments for postoperative delirium in persons aged 50 years or older. *PLoS One* 2014;9:e113946.
10 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–31.
11 Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied logistic regression*. New York: John Wiley & Sons, 2013.
12 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
13 Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? *J Clin Epidemiol* 2020;118:93–9.
14 Rufibach K. Use of Brier score to assess binary predictions. *J Clin Epidemiol* 2010;63:938–9.
15 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565–74.
16 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:18.
17 Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer, 2001.
18 Peek N, Arts DGT, Bosman RJ, *et al*. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epidemiol* 2007;60:491.e1–491.e13.
19 van Steenbeek CD, van Maaren MC, Siesling S, *et al*. Facilitating validation of prediction models: a comparison of manual and semi-automated validation using registry-based data of breast cancer patients in the Netherlands. *BMC Med Res Methodol* 2019;19:117.
20 Tripod checklist: prediction model development. Available: https://www.tripod-statement.org/wp-content/uploads/2020/01/Tripod-Checlist-Prediction-Model-Development.pdf [Accessed 24 Feb 2021].
21 van Meenen LCC, van Meenen DMP, de Rooij SE, *et al*. Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *J Am Geriatr Soc* 2014;62:2383–90.
22 Kalimisetty S, Askar W, Fay B, *et al*. Models for predicting incident delirium in hospitalized older adults: a systematic review. *J Patient Cent Res Rev* 2017;4:69–77.
23 Lindroth H, Bratzke L, Purvis S, *et al*. Systematic review of prediction models for delirium in the older adult inpatient. *BMJ Open* 2018;8:e019223–019223.
24 Pompei P, Foreman M, Rudberg MA, *et al*. Delirium in hospitalized older persons: outcomes and predictors. *J Am Geriatr Soc* 1994;42:809–15.
25 Rudolph JL, Jones RN, Levkoff SE, *et al*. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. *Circulation* 2009;119:229–36.
26 Rudolph JL, Doherty K, Kelly B, *et al*. Validation of a delirium risk assessment using electronic medical record information. *J Am Med Dir Assoc* 2016;17:244–8.
27 Carrasco MP, Villarroel L, Andrade M, *et al*. Development and validation of a delirium predictive score in older people. *Age Ageing* 2014;43:346–51.
28 Kim MY, Park UJ, Kim HT, *et al*. Delirium prediction based on hospital information (Delphi) in general surgery patients. *Medicine* 2016;95:e3072.
29 de Wit HAJM, Winkens B, Mestres Gonzalvo C, *et al*. The development of an automated ward independent delirium risk prediction model. *Int J Clin Pharm* 2016;38:915–23.
30 Pendlebury ST, Lovett NG, Smith SC, *et al*. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of a susceptibility score based on factors identified externally in pooled data for use at entry to the acute care pathway. *Age Ageing* 2017;46:226–31.
31 Halladay CW, Sillner AY, Rudolph JL. Performance of electronic prediction rules for prevalent delirium at hospital admission. *JAMA Netw Open* 2018;1:e181405.
32 Ten Broeke M, Koster S, Konings T, *et al*. Can we predict a delirium after cardiac surgery? A validation study of a delirium risk checklist. *Eur J Cardiovasc Nurs* 2018;17:255–61.
33 Zhang X, Tong D-K, Ji F, *et al*. Predictive nomogram for postoperative delirium in elderly patients with a hip fracture. *Injury* 2019;50:392–7.
34 Dai YT, Lou MF, Yip PK, *et al*. Risk factors and incidence of postoperative delirium in elderly Chinese patients. *Gerontology* 2000;46:28–35.
35 Litaker D, Locala J, Franco K, *et al*. Preoperative risk factors for postoperative delirium. *Gen Hosp Psychiatry* 2001;23:84–9.
36 Wang C-G, Qin Y-F, Wan X, *et al*. Incidence and risk factors of postoperative delirium in the elderly patients with hip fracture. *J Orthop Surg Res* 2018;13:186.
37 Raats JW, van Eijsden WA, Crolla RMPH, *et al*. Risk factors and outcomes for postoperative delirium after major surgery in elderly patients. *PLoS One* 2015;10:e0136071.
38 Setters B, Solberg LM. Delirium. *Prim Care* 2017;44:541–59.
39 Chen Z, Chen R, Zheng D, *et al*. Efficacy and safety of haloperidol for delirium prevention in adult patients: an updated meta-analysis with trial sequential analysis of randomized controlled trials. *J Clin Anesth* 2020;61:109623.
40 Van Calster B, McLernon DJ, van Smeden M, *et al*. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230–7.
41 Bouwmeester W, Zuithoff NPA, Mallett S, *et al*. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1001221.
42 Collins GS, de Groot JA, Dutton S, *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.
43 Collins GS, Omar O, Shanyinde M, *et al*. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66:268–77.
44 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
45 Ettema R, Heim N, Hamaker M, *et al*. Validity of a screening method for delirium risk in older patients admitted to a general Hospital in the Netherlands. *Gen Hosp Psychiatry* 2018;55:44–50.
46 Freter SH, Dunbar MJ, MacLeod H, *et al*. Predicting post-operative delirium in elective orthopaedic patients: the delirium elderly at-risk (dear) instrument. *Age Ageing* 2005;34:169–71.