

BMJ Open Systematic review of prediction models for delirium in the older adult inpatient

Heidi Lindroth,^{1,2} Lisa Bratzke,³ Suzanne Purvis,⁴ Roger Brown,³ Mark Coburn,⁵ Marko Mrkobrada,⁶ Matthew T V Chan,⁷ Daniel H J Davis,⁸ Pratik Pandharipande,⁹ Cynthia M Carlsson,^{1,14,15,16,17} Robert D Sanders¹

To cite: 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. doi:10.1136/bmjopen-2017-019223

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2017-019223>).

Received 18 August 2017
Revised 13 March 2018
Accepted 19 March 2018

ABSTRACT

Objective To identify existing prognostic delirium prediction models and evaluate their validity and statistical methodology in the older adult (≥60 years) acute hospital population.

Design Systematic review.

Data sources and methods PubMed, CINAHL, PsychINFO, SocINFO, Cochrane, Web of Science and Embase were searched from 1 January 1990 to 31 December 2016. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses and CHARMS Statement guided protocol development. Inclusion criteria: age >60 years, inpatient, developed/validated a prognostic delirium prediction model. Exclusion criteria: alcohol-related delirium, sample size ≤50. The primary performance measures were calibration and discrimination statistics. Two authors independently conducted search and extracted data. The synthesis of data was done by the first author. Disagreement was resolved by the mentoring author.

Results The initial search resulted in 7,502 studies. Following full-text review of 192 studies, 33 were excluded based on age criteria (<60 years) and 27 met the defined criteria. Twenty-three delirium prediction models were identified, 14 were externally validated and 3 were internally validated. The following populations were represented: 11 medical, 3 medical/surgical and 13 surgical. The assessment of delirium was often non-systematic, resulting in varied incidence. Fourteen models were externally validated with an area under the receiver operating curve range from 0.52 to 0.94. Limitations in design, data collection methods and model metric reporting statistics were identified.

Conclusions Delirium prediction models for older adults show variable and typically inadequate predictive capabilities. Our review highlights the need for development of robust models to predict delirium in older inpatients. We provide recommendations for the development of such models.

INTRODUCTION

Delirium is an acute disturbance of consciousness and cognition precipitated by an acute event such as sudden illness, infection or surgery. This syndrome is a serious public health concern, as up to 50% of hospitalised older adults will experience delirium in medical and surgical populations.^{1–3}

Strengths and limitations of this study

- This study used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement and the CHARMS checklist to develop a protocol involving comprehensive search terms and databases.
- The assembled interprofessional authorship team contributed different perspectives on delirium prediction models and statistical methodology.
- This review focused on a narrow population and older adult inpatients and could be expanded to include all ages and settings including palliative care, long-term care and the emergency room.

Delirium has been independently associated with increased mortality, morbidity in terms of impaired cognition and functional disability along with an estimated annual US expenditure of \$152 billion.^{4–9} Prediction models allow clinicians to forecast which individuals are at a higher risk for the development of a particular disease process and target specific interventions at the identified risk profile.^{10–13} At present, an extensive list of modifiable and non-modifiable, predisposing and precipitating delirium risk factors encumbers clinicians, hindering the ability to select the most important or contributing risk factor.^{1 14} An accurate and timely delirium prediction model would formalise the highest impact risk factors into a powerful tool, facilitating early implementation of prevention measures.¹¹ This systematic review expands on previous published reviews on delirium prediction models by integrating both medical and surgical populations while examining statistical aspects of each study including reporting metrics and includes recently published models.

Aim

Our aim was to provide important recommendations on study design for future delirium prediction models while integrating knowledge gained from the study of both medical and surgical populations. We conducted a



For numbered affiliations see end of article.

Correspondence to

Heidi Lindroth;
hlindroth@wisc.edu

systematic review of the literature focusing on the identification and subsequent validity of existing prognostic delirium prediction models in the older adult (≥ 60 years old) acute hospital population.

METHODS

This systematic review followed the protocol developed from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement and the CHARMS checklist (online supplementary appendix A).^{15 16} A delirium prediction model was defined as a statistical model that either stratified individuals for their level of delirium risk, or assigned a risk score to an individual based on the number and/or weighted value of pre-determined modifiable and non-modifiable risk factors of delirium present. This review included studies focused on (1) older adult (≥ 60 years) population, (the US Centers for Disease Control and Prevention and United Nations define an older adult as 60 years of age and older),^{17 18} (2) inpatient hospital setting, (3) publication dates of 1 January 1990–31 December 2016 and (4) developed and/or validated delirium prediction models. Studies were excluded if they (1) studied a different patient population (ie, emergency department, skilled nursing facilities, palliative care and hospice) as these are unique patient populations with characteristics requiring specific foci and are not readily generalisable to a medical or surgical inpatient hospital setting. Furthermore, recommended therapies for treatment of delirium symptoms vary between the populations,^{19 20} (2) related to alcohol withdrawal, or delirium tremens, as the presence of alcohol withdrawal complicates delirium assessment and (3) had a sample size of ≤ 50 for methodological reasons (ie, underpowered). All study designs were included. Studies were not limited by time frame of delirium development (prevalent vs incident); however, only prognostic statistics were discussed.

The search terms were as follows: ('Delirium' OR 'postoperative delirium' OR 'ICU delirium' OR 'ICU psychosis' OR 'ICU syndrome' OR 'acute confusional state' OR 'acute brain dysfunction') AND ('inpatient' OR 'hospital*' OR 'postoperative' OR surg* OR 'critical care unit' OR 'intensive care unit' OR CCU OR ICU) AND ('predict*' model OR risk*). Electronic databases of PubMed, CINAHL, PsycINFO, Cochrane Database of Systematic Reviews, SocINDEX, Web of Science and Embase were searched. Studies using a language other than English were included if translation was available through the University of Wisconsin-Madison Health Sciences Librarian. Bibliographies of identified studies were hand-searched for additional references. Study quality was assessed through the Newcastle-Ottawa Scale (NOS)²¹ for case-control and cohort studies. Risk of bias was assessed through the Critical Appraisal and Data Extraction for Systematic Reviews (CHARMS) checklist.¹⁵ Two authors (HL and SP) independently performed data collection, data extraction and assessed study quality, with any disagreement resolved by RDS.

Outcomes

Data extracted included: (1) study characteristics (study design, population and sample size), (2) outcome measure (method of identification and diagnosis, frequency and length of screening), (3) model performance information including the diagnostic accuracy of the delirium prediction models, calibration metrics and events per variable (EPVs), (4) characteristics of the models (variables used in model and scoring/stratification system), (5) cognitive measures used in the study and (6) statistical methods applied for analysis. Five authors were contacted for missing or incomplete data. Four responses were received.

Statistics

Model performance was assessed through calibration and classification metrics.¹⁵ The AUROC was the primary measure collected to evaluate the discriminatory ability of the delirium prediction models. Clinical utility statistics such as sensitivity, specificity, positive predictive values, negative predictive values, ORs, relative risk statistics and use of decision curve analysis or clinical utility cure analysis were also collected from each delirium prediction model in reference to the model's reported cut-off value. Goodness-of-fit statistics including χ^2 and Hosmer-Lemeshow tests were collected to evaluate effective model calibration. Studies were also assessed for the inclusion of calibration plots and slopes. Model calibration refers to the agreement between observed outcomes and predictions.²² Secondary preplanned outcome measures included cognitive assessments and predictive variable use per model.

Role of the funding source

The funding sources named had no role in this study. All authors had full access to all the data in the study and shared responsibility for the decision to submit the publication.

Patient and public involvement

Neither patients nor the public were involved with the development or design of this study.

RESULTS

Twenty-seven studies were identified for inclusion.^{23–47} The initial search resulted in 7,502 citations, with 192 studies chosen for full-text review as detailed in the PRISMA diagram (figure 1). We did not identify any relevant, unpublished studies for this review. The inclusion criteria were modified for two studies that developed models in younger populations, but these models were externally validated in the target population of this review (age ≥ 60 years).^{25 40}

Twenty-three delirium prediction models were developed, 14 were externally validated^{23 27 29–31 33–35 41 43–46 48} and three were internally validated.^{24 37 42} Prospective cohort design was used in 24 studies.^{23 25–31 33–35 37–49} Retrospective design was

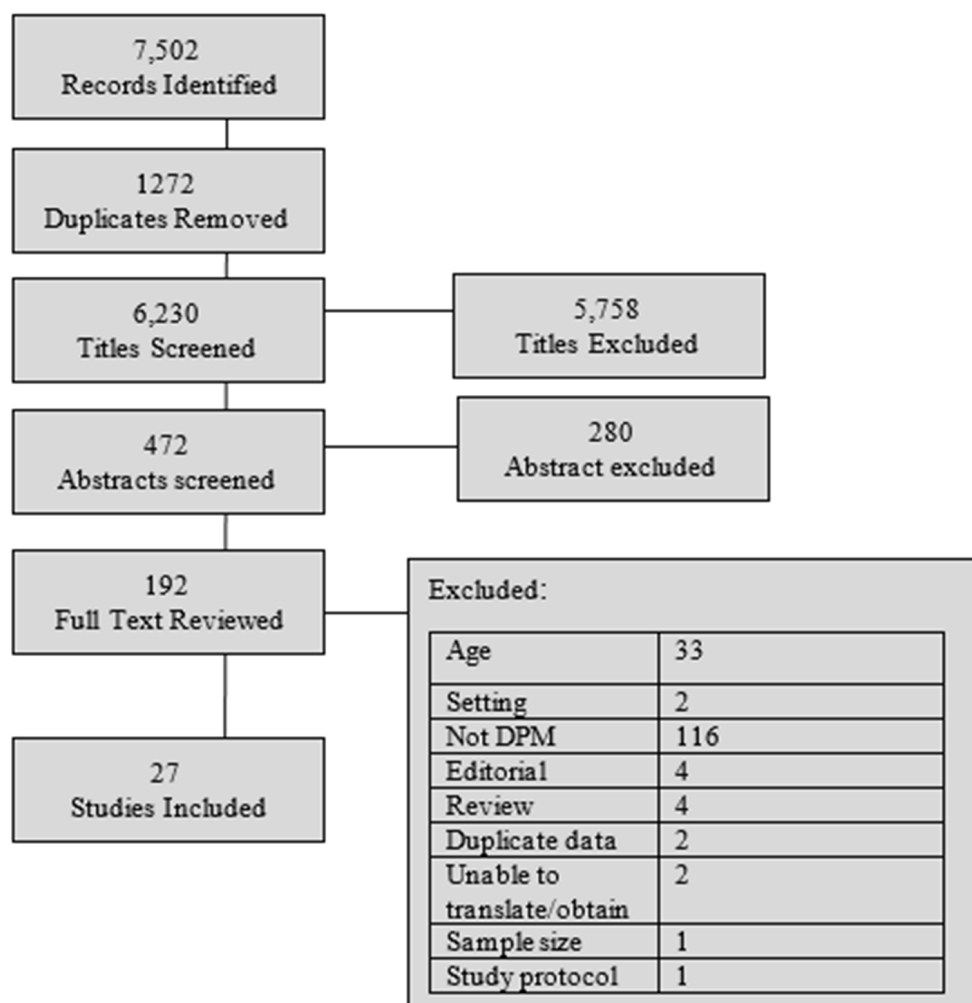


Figure 1 PRISMA diagram: study selection. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

used in four studies.^{24 32 36 44} Nineteen studies used consecutive sampling methods,^{23 25–31 33 34 38 40–42 44 45 47–49} two of these were part of a randomised control trial.^{34 41} Eleven studies focused on the medical population,^{23 25–29–33 40 42 45 48} 3 included medical and surgical^{24 43 44} and 13 recruited a surgical population (seven orthopaedic,^{26–28 34 38 41 49} one cardiac,⁴⁶ two non-cardiac,^{37 47} one general surgery³⁵ and two oncological^{36 39}). None of the identified studies focused on critical care patients. Data collection occurred on admission in 17 studies^{23 25 27 29–31 33–35 40–45 48 49}; participants were approached within 48 hours of admission. Seven studies collected data preoperatively then followed participants postoperatively.^{26 28 37–39 46 47} Data collection overlapped with delirium assessments in three studies.^{27 32 35} The average NOS quality ranking for included cohort studies was seven; six studies received the maximum of nine stars. Risk of bias was assessed using the CHARMS checklist,¹⁵ and results are shown in figure 2. Further characteristics of studies are listed in table 1.

Delirium assessment

The outcome variable was measured using the Confusion Assessment Method in 21 studies.^{23 25–31 33–40 43 46–49}

The frequency of delirium assessment varied from two or more assessments daily (3 studies),^{26 35 41} to once daily (12 studies),^{25 28 30 32 34 36–38 44–46 49} every other day (8 studies),^{23 27 29 31 33 42 43 48} once following surgery⁴⁷ and undefined (3 studies).^{24 39 40} Of the studies that assessed delirium twice or more daily, all of these studies relied on ward nurse observations or telephone interview with the nurse to identify delirium symptoms.^{26 35 41} The principal investigator confirmed the presence of delirium following the nurse report of symptoms.^{26 35} Twenty-one studies used trained research or clinical personnel to conduct the delirium assessments.^{23 25–27 29–31 33–40 43–48} Three studies relied on delirium diagnosis, or keywords designated as representing delirium, to identify the outcome measure through retrospective chart review.^{24 32 44} Three studies relied on clinical staff to recognise and chart delirium symptoms.^{28 41 49} One of these studies retrospectively confirmed the diagnosis of delirium through consensus review of two authors; disagreement was resolved by a psychiatrist.⁴¹ One study did not report details on personnel performing delirium assessments.⁴²

Carrasco 2014	+	+	+	1	1
De Wit 2016	?	1	1	+	+
Douglas	+	+	+	+	+
Dworkin	+	+	+	?	1
Fisher and Flowerdew '95	+	+	+	1	1
Freter 2005	+	+	+	+	1
Freter 2005	+	+	+	1	1
Freter 2015	+	?	+	+	+
Inouye & Charpentier '96	+	+	+	+	+
Inouye 2007	+	+	+	?	+
Inouye 1993	+	+	+	+	+
Isfandiatty 2012	1	1	1	+	?
Kalisvaart 2006	+	+	+	+	+
Kim 2016	+	1	+	+	+
Korc-Grodzicki 2014	1	+	+	+	1
Leung 2013	+	+	+	+	+
Liang 2015	+	+	+	?	+
Maekawa 2015	?	+	+	+	1
Martinez 2012	+	+	+	+	?
Moerman 2012	+	+	?	+	?
O'Keeffe & Lavan 1996	+	+	?	1	?
Pendlebury 2016	+	+	+	+	+
Pendlebury 2016	+	+	+	+	+
Pompei 1994	?	+	+	+	?
Rudolph 2009	?	+	+	+	+
Rudolph 2011	+	+	+	1	+
Rudolph 2016	+	+	+	+	+
	Study Participants	Predictors	Outcome	Sample size and Missing data	Statistical Analysis

Figure 2 This displays the CHARMS risk of bias assessment on all included studies. Study participants: design of included study, sampling method and inclusion/exclusion criteria. Predictors: definition, timing and measurement. Outcome: definition, timing and measurement. Sample size and missing data: number of participants in study, events per variable and missing data. Statistical analysis: selection of predictors, internal validation and type of external validation.

Model design and statistical methods

Various statistical techniques were employed by the 23 included studies. Twelve used univariate or bivariate analyses and selected variables with a predetermined statistical value (range from $p < 0.05$ to $p < 0.25$) for inclusion in the model.^{23–26 32 35–37 40 42 43 46} Five of these models paired bivariate analyses with a bootstrapping technique to address lower sample and event size.^{24 25 37 38 46} Four models based their variable selection from a literature review of risk factors for delirium.^{27 28 41 44 48 49} Two used proportional hazards regression modelling paired with bivariate analyses and included variables with either a p value < 0.25 ³² or a relative risk of ≥ 1.50 .³⁰ Six studies published their power analysis.^{24 25 33 35 40 41} Sixteen studies employed a form of

logistic regression. Twelve of these models applied a stepwise regression approach.^{23 25 26 29 30 35–37 42 43 46 47} Three applied a stepwise forward selection process,^{23 25 30} two employed a stepwise backward selection process^{35 46} and one used a combined approach.²⁹ Statistical methods used for model building are further outlined in table 1.

Per TRIPOD reporting guidelines, validation studies were categorised into type; narrow validation refers to the same investigators subsequently collecting an additional patient cohort, following the development cohort, and broad validation refers to a validation cohort sampled from a different hospital or country.^{50–52} As interpretation of validation studies is dependent on case-mix,⁵³ it is important to note that 8 of the 14 externally validated

Table 1 Displays the 27 studies that were identified for inclusion in this review.

Author	Study design population sample size sampling method power analysis	Study grade (NOS)	Outcome variable and rate (%)	Delirium measurement and frequency	DPM name and regression model used
Carrasco et al ²³	P:Cohort Medical Dev: 374 Val: 104 Consecutive	S: **** C: – O: ** T: 6 stars	Delirium Dev: 25 (0.06) Val: 12 (12)	CAM Every 48 hours	Predictive Risk Score Forward stepwise
	Retro All hospital patients Dev: 1291 Convenience Power analysis	S: *** C: ** O: *** T: 8 stars	Delirium Dev: 225 (17)	Chart abstraction EHR 'diagnosis table'	Automated delirium prediction model Multivariate
Douglas et al ²⁵	P:Cohort Medical Dev: 209 Val: 165 Consecutive Power analysis	S: **** C: – O: *** T: 7 stars	Delirium Dev: 25 (12) Val: 14 (8.5)	Short CAM Daily	Risk stratification model (AWOL) Forward stepwise
Dworkin et al ⁴⁷	P:Cohort Elective non-cardiac surg Dev: 76 Consecutive	S: **** C: – O: ** T: 6 stars	Delirium Dev: 10 (13)	CAM or FAM-CAM 1 x after surgery	Mini-Cog Stratified into a five-point score Stepwise
Fisher and Flowerdew ²⁶	P:Cohort Elective orthopaedic Dev: 80 Consecutive	S: ** C: – O: ** T: 4 stars	Delirium Dev: 14 (17.5)	CAM 2 x Daily	Prediction model using two variables. Stepwise
Freter et al ²⁸	P:Cohort Elective hip surgery Dev: 132 Consecutive	S: ** C: ** O: ** T: 6 stars	Delirium Dev: 18 (14)	CAM Daily	Risk stratification model (DEAR) Built from literature
Freter et al ⁴⁹	P:Cohort Hip Fx Dev: 100 Consecutive	S: ** C: ** O: ** T: 6 stars	Delirium Dev: 24 (24)	CAM Daily	Risk stratification model (DEAR)
Freter et a ²⁷	P:Cohort Hip fracture Val: 283 Consecutive	S: *** C: – O: ** T: 5 stars	Delirium Val: 119 (42)	CAM POD1, 3 and 5	Risk stratification model (DEAR)
Inouye and Charpentier ²⁹	P:Cohort Medical Dev: 196 Val: 312 Consecutive	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 35 (18) Val: 47 (15)	CAM Every other day	Risk stratification model based on precipitating factors Backwards and forwards stepwise

Continued

Table 1 Continued

Author	Study design population sample size sampling method power analysis	Study grade (NOS)	Outcome variable and rate (%)	Delirium measurement and frequency	DPM name and regression model used
Inouye <i>et al</i> ³¹	P:Cohort Medical Dev: 491 Val: 461 Consecutive	S: **** C: ** O: *** T: 9 stars	Delirium/ subsyndromedelirium at discharge Dev: 58 (12) Val: 28 (6)	CAM Every other day	Risk stratification model Log-binomial regression
Inouye <i>et al</i> ³⁰	P:Cohort Medical Dev: 107 Val: 174 Consecutive	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 27 (25) Val: 29 (17)	CAM Daily	Risk stratification model Forward stepwise
Isfandiyyat <i>et al</i> ³²	Retro Medical Dev: 457 Convenience	S: ** C: – O: *** T: 5 stars	Delirium Dev: 87 (19)	Undefined Daily	Risk stratification model Cox's proportional hazard
Kalisvaart <i>et al</i> ³⁴	P:Cohort Hip surgery and fracture Val: 603 Consecutive	S: *** C: – O: *** T: 6 stars	Delirium Val: 74 (12)	CAM, DRS-98 Daily through POD5	Externally validated Inouye's 1993 model.
Kim <i>et al</i> ³⁵	P:Cohort Major general surgery Dev: 561 Val: 533 Not stated Power analysis	S: *** C: ** O: *** T: 8 stars	Delirium Dev: 112 (20) Val: 99 (18)	Nu-Desc: every shift by RNs Confirmed with CAM	Risk stratification model Backwards stepwise
Korc-Grodzicki <i>et al</i> ³⁶	Retro Oncological surgery Dev: 416 Convenience	S: *** C: – O: *** T: 6 stars	Delirium Dev: 79 (19)	CAM Daily	Comprehensive Geriatric Assessment (CGA) as model Stepwise
Leung <i>et al</i> ³⁷	P:Cohort Non-cardiac surgery Dev: 581 Not stated	S: *** C: – O: ** T: 5 stars	Delirium Dev: 234 (40)	CAM Daily	Risk stratification model Stepwise
Liang <i>et al</i> ³⁸	P:Cohort Elective orthopaedic Surgery Dev: 461 Consecutive	S: *** C: ** O: *** T: 7 stars	Delirium Dev: 37 (8)	CAM Daily	Built two DPMs using CGA Risk stratification models Backward stepwise
Maekawa <i>et al</i> ³⁹	P:Cohort Oncological; gastrointestinal surgery Dev: 517 Consecutive	S: ** C: * O: *** T: 6 stars	Delirium Dev: 124 (24)	CAM Unknown frequency	CGA as model Proportional hazards

Continued

Table 1 Continued

Author	Study design population sample size sampling method power analysis	Study grade (NOS)	Outcome variable and rate (%)	Delirium measurement and frequency	DPM name and regression model used
Martinez <i>et al</i> ^{40*}	P:Cohort Medical Dev: 397 Val: 302 Consecutive Power analysis	S: *** C: – O: ** T: 5 stars	Delirium Dev: 52 (13) Val: 76 (25)	CAM Undefined	Clinical prediction rule Multivariate Recursive partitioning
Moerman <i>et al</i> ⁴¹	P:Cohort Hip fracture Val: 378 Consecutive Power analysis	S: *** C: – O: *** T: 6 stars	Delirium Val: 102 (27)	Ward RN observation, 3× daily Confirmed by chart review	Risk stratification model (Risk Model for Delirium, RD) Built from literature
O'Keeffe and Lavan ⁴²	P:Cohort Acute geriatric unit Dev: 100 Ival: 84 Consecutive	S: **** C: – O: ** T: 6 stars	Delirium Dev: 28 (28) IVal: 25 (30)	DAS Every 48 hours DSM III	Risk stratification model Stepwise
Pendlebury <i>et al</i> ⁴⁸	P:Cohort Medical Val: 308 Consecutive	S: **** C: ** O: *** T: 9 stars	Delirium Val: 95 (31)	CAM Every 48 hours Confirmed by DSM-IV interview	Susceptibility Score Built from literature
Pendlebury <i>et al</i> ³³	P:Cohort Medical Val: 308 Consecutive Power analysis	S: **** C: – O: *** T: 7 stars	Delirium Val: 95 (31)	CAM Every 48 hours Confirmed by DSM-IV interview	Externally validated four DPMs
Pompei <i>et al</i> ⁴³	P:Cohort Med/surg Dev: 432 Val: 323 Not stated	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 64 (14.8) Val: 86 (26.3)	CAM 2× weekly, Confirmed with DSM III	Risk stratification model Stepwise
Rudolph <i>et al</i> ⁴⁶	P:Cohort Cardiac surgery Dev: 122 Val: 109 Not stated	S: *** C: * O: ** T: 6 stars	Delirium Dev: 63 (52) Val: 48 (44)	CAM, MDAS, DSI Daily	Risk stratification model Backward stepwise
Rudolph <i>et al</i> ⁴⁵	P:Cohort Medical Val: 100 Consecutive	S: **** C: – O: *** T: 7 stars	Delirium Dev: 23 (23)	DSM-IV Daily clinical interview	Externally validated Inouye's 1993 model

Continued

Table 1 Continued

Author	Study design population sample size sampling method power analysis	Study grade (NOS)	Outcome variable and rate (%)	Delirium measurement and frequency	DPM name and regression model used
Rudolph <i>et al</i> ⁴⁴	Dev: Retro Val: P Cohort Med/surg Dev: 27 625 Val: 246 Consecutive	S: *** C: – O: ** T: 6 stars	Delirium Dev: 2343 (8) Val: 64 (26)	Dev: chart audit Val: DSM-IV Daily clinical interview	Risk stratification model Built from literature

Study design: Dev, development; Med, medical; P Cohort, prospective cohort; Retro, retrospective design; Surg, surgical; Val, validation; Power analysis, reported in identified study.
 Study grade: NOS, Newcastle Ottawa Scale; C, comparability; O, Ottawa; S, Selection; T, Total; Max 9 stars. Outcome variable: CAM, Confusion Assessment Method; DRS-98, Delirium Rating Scale-R-98; DSM, Diagnostic Statistical Manual; EHR, Electronic Health Record; MDAS, Memorial Delirium Assessment Scale; Nu-Desc, Nursing Delirium Screening Scale; POD, postoperative day; DSI, Delirium Symptom Interview; DAS, Delirium Assessment Scale; FAM-CAM, Family Confusion Assessment Method; RNs, Registered Nurses; IVAL, Internal Validation
 Type of model: how authors designed their delirium prediction model (DPM), statistical method used.
 Risk stratification model: points (weighted or unweighted) assigned per predictive risk factor present.
 Built from literature: authors selected risk factors for DPM based on literature review.
 AWOL, DEAR, and RD are the names of the prediction models given by the developing authors.
 *Models developed in population ≤ 60 years of age but validated in population ≥ 60 years of age.
 CGA, Comprehensive Geriatric Assessment.

models are categorised as narrow validations.^{23 27 29–31 35 41 46}
 Further information is outlined in [table 2](#).

Variables

Figure 3 demonstrates the frequency of variable use in the 14 externally validated delirium prediction models. Baseline cognitive impairment was the most frequently used variable. Six models defined baseline cognitive impairment as a cognitive test score at or below the level of dementia.^{27 30 34 43 48} This cognitive test was administered on study enrolment or extracted from past medical records.⁴⁸ Two studies additionally evaluated chronic cognitive impairment through family or caregiver interview with the modified Blessed Dementia Rating Scale (mBDRS).^{30 31} Four models combined the cognitive test score derived on enrolment with a history of dementia to define baseline cognitive impairment.^{31 33 41 44} History of dementia was defined as follows: two studies: family or caregiver report supplemented with documented history in medical record,^{33 41} one study: medical record review and interview with mBDRS³¹ and one study: dementia billing codes or prescription information.⁴⁴ One study defined baseline cognitive impairment as a prespecified key term in the electronic health.⁴⁵ **Table 2** details cognitive tests used in the externally validated delirium prediction models.

Functional impairment was defined as follows: (1 study) needing assistance with any basic activities in daily living (ADL),²⁷ (1 study) domestic help, help with meals or physical care⁴¹ and (2 studies) residence in nursing facility or at home with caregivers,³³ and (2 studies) requiring a home care package with professional caregivers or residence in a care home.^{33 48} The latter being obtained on admission from medical records.^{33 48} Two studies used validated functional assessment tools (Instrumental Activities of Daily Living (iADL) and Barthel Index) and evaluated functional status 2 weeks prior to hospitalisation.^{23 31}

Externally validated delirium prediction models are detailed in [table 2](#).

Predictive ability

Reported AUROC in externally validated delirium prediction models ranged from 0.52 to 0.94 ([figure 4](#)). Of these models, the highest performing model (AUROC 0.94, 95% CI 0.91 to 0.97) was developed and validated in a surgical population.³⁵ Two models reported an external validation AUROC above 0.80, indicating moderate predictive ability.^{33 48} Both were developed and validated in medical populations and share similarities with variable use including pre-existing cognitive impairment and presence of infection.

Model calibration

Six of the 14 externally validated delirium prediction models reported calibration metrics.^{29–31 34 43 45} The reported χ^2 statistics were significant in five prognostic models^{29–31 34 43} and did not reach significance in one model.⁴⁵ Four of the 23 studies that developed models

Table 2 Detailed description of the externally validated DPMs.

External validated DPM name	Citation type of validation	Delirium # (%)	Sens Spec PPV NPV (external)	AUROC (95% CI)	Model components	Cog. assess tool and cut-off
AWOL tool	Pendlebury <i>et al</i> (2016) Broad val.	1st val: 14 (9) 2nd val: 95 (31) (any delirium) 67: prevalent 28: incident	Mod. AWOL Cut-off: 3 Any delirium Sens 0.7 Spec 0.66 PPV 0.55 NPV 0.79 Incident delirium Sens 0.76 Spec 0.66 PPV 0.27 NPV 0.94	1st val: 0.69 (0.54 to 0.83) Incident delirium 2nd val: Cohort 1 (MMSE) 0.78 (0.68 to 0.88) Cohort 2 (AMTS) 0.73 (0.63 to 0.83)	Original AWOL Tool Age >80 years 1 pt Failure to spell WORLD backwards 1 pt Disorientation 1 pt Illness severity 1 pt Modified AWOL Tool Age >80 years 1 pt Diag of dementia 1 pt MMSE <24, AMTS <9 1 pt Illness severity 1 pt	MMSE <24 AMTS <9
Clinical prediction rule: cardiac surgery	Rudolph <i>et al</i> (2009) Narrow val.	Dev: 63 (52) Val: 48 (44) (incident delirium)	Not reported	Dev: 0.74 Val: 0.75 Did not report CI	Weighted points-regression MMSE <23 2 pts MMSE 24–27 1 pt Hx of stroke/TIA 1 pt GDS >4 1 pt Abnormal albumin 1 pt Stratified into point categories 0 pt 1 pt 2 pts ≥3 pts – high-risk group RR in high-risk group: 4.9 (3.8–6.2)	MMSE -stratified score
DEAR	Freter <i>et al</i> ²⁷ Narrow val.	Dev: (2005) 18 (14) Val: (2015) Preop= 163 (58) Postop= 118 (42)	Sens 0.68 Spec 0.73 PPV 0.65 NPV 0.76 Optimal cut-off score: 3 pts (incident postop delirium)	Dev: (2005) 0.77 (0.64 to 0.87) Val: (2015) AUROC not published	MMSE <23 1 pt Functional dependence 1 pt Sensory impairment 1 pt Substance use 1 pt Age >80 years 1 pt Not weighted. 0–5 score, cut-off of 3 indicating high risk.	MMSE Cut-off ≤23

Continued

Table 2 Continued

External validated DPM name	Citation type of validation	Delirium # (%)	Sens Spec PPV NPV (external)	AUROC (95% CI)	Model components	Cog. assess tool and cut-off
Delirium at discharge prediction model	Inouye <i>et al</i> ³¹ Narrow val.	Dev: 58 (12) Val: 28 (6) (incident delirium)	Not reported	Dev: 0.80 Val: 0.75 Did not report CI Calibration: χ^2 trend: $p < 0.001$	Delirium at discharge prediction Dementia diagnosis or mBDRS > 4 1 pt Vision impairment 1 pt ADL impairment 1 pt Charlson score 1 pt Restraint use during delirium 1 pt Not weighted. 0–1 pt=low risk 2–3 pts=intermediate risk 4–5 pt=high risk RR in high-risk group: 10.2 (3.2–32.7)	MMSE <24 mBDR ≥ 4
Delirium Prediction Score (DPS)	Carrasco <i>et al</i> ²³ Narrow val.	Dev: 25 (0.06) Val: 12 (12) (incident delirium)	Sens 0.88 Spec 0.74 PPV 0.22 NPV 0.99	Dev: 0.86 (0.82 to 0.91) Val: 0.78 (0.66 to 0.90)	DPS=[5×BUN/Cr ratio]–(3× Barthel Index). Cut-off is: > 240 =high risk for delirium In conventional units, cut-off is: > 160 =high risk for delirium	None. Pfeffer Functional Activities Questionnaire as a proxy for prior dementia

Continued

Table 2 Continued

External validated DPM name	Citation type of validation	Delirium # (%)	Sens Spec NPV (external)	AUROC (95% CI)	Model components	Cog. assess tool and cut-off
Delphi score	Kim <i>et al</i> . ³⁵ Narrow val.	Dev: 112 (20) Val: 99 (18) (incident delirium)	Sens 0.81 Spec 0.93 PPV 0.70 NPV 0.96 Optimal cut-off score: 6.5 pts	Dev: 0.911 (0.88 to 0.94) Val: 0.938 (0.91 to 0.97)	Age (years) 60–69 0 70–79 1 >80 2 Low physical activity Self-sufficient 0 Need assist. 2 Heavy ETOH No 0 Yes 1 Hearing impairment No 0 Yes 1 History of delirium No 0 Yes 2 Emergency surgery No 0 Yes 1 Open surgery No 0 Yes 2 ICU admission No 0 Yes 3 Preop CRP (mg/dL) <10 0 >10 1 Max points: 15 Optimal cut-off: 6.5 High risk: > 7 pts	No measure of cognition. Excluded participants if MMSE <24

Continued

Table 2 Continued

External validated DPM name	Citation type of validation	Delirium # (%)	Sens Spec PPV NPV (external)	AUROC (95% CI)	Model components	Cog. assess tool and cut-off
e-NICE rule	Rudolph <i>et al</i> ⁴⁴ Broad val. Dev: 2343 (8) Val: 64 (26) (incident delirium)	Cohort AUROC CI TPR FPR Dev: 0.81 (0.80 to 0.82) Val: AUROC* Original 0.69 (0.61 to 0.77) 64%–33% mRASS 0.72 (0.65 to 0.79) 69%–35% TMYB 0.73 (0.66 to 0.80) 78%–43% MoCA 0.74 (0.66 to 0.81) 75%–43% *Any delirium Original model: AUROC of 0.68 (95% CI 0.59 to 0.77) in incident delirium. Did not report sens, spec, PPV and NPV			Weighted points/OR Cog. impair Medications, diagnosis or both 4 pts Age >65 years 2 pts Age >80 years 3 pts Infection 2 pts Fracture 4 pts Vision 1 pt Severe illness 2 pts 0–2 pts=low risk 2–5 pts=intermediate risk 6–8 pts=high risk ≥9 pts=very high risk	e-NICE Tool Diagnosis of dementia, medications for dementia or both qualified as 'cognitive impairment' in model. Prospective cohort, additional: mRASS TMYB MoCA <18
Inouye Prediction Rule (IPR)	Inouye <i>et al</i> ³⁰ Narrow val.	Dev: 27 (25) Val: 29 (17) (incident delirium)	Did not report	Dev: 0.74 (0.63 to 0.85) Val: 0.66 (0.55 to 0.77) Calibration below: Dev: X ² trend p<0.00001 Val: X ² trend p<0.002	Baseline cognitive impairment 1 pt High BUN/Cr ratio 1 pt Severe illness (Composite score: APACHE II >16+RN rating) 1 pt Vision impairment 1 pt Not weighted. 0 pt=low risk 1–2 pts=intermediate risk 3–4 pts=high risk RR in high risk group: 9.5 (no CI)	MMSE cut-off <24 Family/caregiver bDRS Excluded those with history of severe dementia
IPR	Kalisvaart <i>et al</i> ³⁴ Broad val.	Val: 74 (12)	Did not report	Val: 0.73 (0.65 to 0.78) Calibration: X ² p<0.05 X ² trend p<0.002	Externally validated IPR in surgical hip fracture population. ► Addition of age and type of admission improved model performance, R ² =0.20 RR of high risk group: 9.8	MMSE cut-off <24
IPR	Rudolph <i>et al</i> ⁴⁵ Broad val.	Val: 23 (23) Any delirium 10: prevalent 13: incident	Did not report	Val: 0.56 (0.42 to 0.74) Incident delirium Calibration: X ² 1.3, p=0.53	Externally validated IPR in medical VA population, investigated feasibility of chart abstraction tool.	MMSE cut-off <24

Continued

Table 2 Continued

External validated DPM name	Citation type of validation	Delirium # (%)	Sens Spec PPV NPV (external)	AUROC (95% CI)	Model components	Cog. assess tool and cut-off
IPR	Pendlebury <i>et al</i> ³³ Broad val.	Val: 95 (31) Any delirium 67: prevalent 28: incident	Cut-off 2 pts All delirium Sens 0.57 Spec 0.80 PPV 0.64 NPV 0.76 Incident delirium Sens 0.52 Spec 0.80 PPV 0.31 NPV 0.91	Val: Incident delirium Cohort 1 (MMSE) 0.73 (0.62 to 0.84) Cohort 2 (AMTS) 0.70 (0.60 to 0.81)	Baseline cognitive impairment 1 pt High BUN/Cr ratio 1 pt Severe illness (SIRS >2) 1 pt Vision impairment 1 pt 4 pts=incident delirium	Original model: MMSE <24 Modified model: MMSE <24 AMTS <9
Istandiati model	Pendlebury <i>et al</i> ³³ Broad val.	Dev: 87 (19) Val: 95 (31) Any delirium 67: prevalent 28: incident	Cut-off 4 pts All delirium Sens 0.74 Spec 0.71 PPV 0.60 NPV 0.82 Incident delirium Sens 0.81 Spec 0.71 PPV 0.31 NPV 0.96	Dev: 0.82 (0.77 to 0.88) Val: Incident delirium Cohort 1 (MMSE) 0.83 (0.74 to 0.91) Cohort 2 (AMTS) 0.77 (0.67 to 0.86)	Baseline cognitive impairment 3 pts Functional dependency 2 pts Infection with sepsis 2 pts Infection without sepsis 1 pt Weighted score Score=7 for incident delirium Cohort 1: MMSE Cohort 2: AMTS	Original model: Chart review Modified model: MMSE >24 AMTS <9

Continued

Table 2 Continued

External validated DPM name	Citation type of validation	Delirium # (%)	Sens Spec PPV NPV (external)	AUROC (95% CI)	Model components	Cog. assess tool and cut-off
Martinez <i>et al</i> 2012 model	Pendlebury <i>et al</i> ³³ Broad val.	1st Val: 76 (25) 2nd Val: 95 (31) Any delirium 67: prevalent 28: incident	Modified model Cut-off 2 pts All delirium Sens 0.62 Spec 0.68 PPV 0.54 NPV 0.75 Incident delirium Sens 0.81 Spec 0.68 PPV 0.29 NPV 0.96	1st Val: 0.85 (0.80 to 0.88) Incident delirium 2nd Val: Cohort 1 (MMSE) 0.78 (0.68 to 0.88) Cohort 2 (AMTS) 0.75 (0.65 to 0.84)	Martinez <i>et al</i> 2012 original model Age >85 years 1 pt Dependent in >5ADLs 1 pt Drugs on admit: 1 pt/drug 2 pts/antipsych 2 pts/antidepressants 2 pts/antidementia 2 pts/anticonvulsants 2 pts/antipsychotics Score 0-3 Score >1=high risk for delirium Modified model Age >85 years 1 pt Dependency in >5ADLs 1 pt Diag of dementia MMSE <24 AMTS <9 1 pt	Original model: No cognitive measure Modified model: MMSE <24 AMTS <9
Pompei <i>et al</i> 1994 model	Pompei <i>et al</i> ⁴³ Broad val.	Dev: 64 (15) Val: 86 (26) (21=prevalent delirium)	Sens 0.83 Spec 0.50 PPV 0.38 NPV 0.89 *Pts stratified as low or moderate to high risk	Dev: 0.74 ±0.05 Val: 0.64 ±0.05 Calibration: X ² trend p<0.0001	Weighted points Baseline cognitive impairment 2 pts Depression 2 pts Alcoholism 3 pts >4 comorbidities 3 pts 0-3 pts=low risk 4-7 pts=moderate risk 8-10 pts=high risk	MMSE Less than high school <21 High school <23 College education <24
Precipitating risk factors	Inouye and Charpentier ²⁹ Narrow val.	Dev: 35 (18) Val: 47 (15) (incident delirium)	Not reported	No AUROC reported Calibration: X ² trend p<0.001	Physical restraint use 1 pt Malnutrition 1 pt >3 medications added 1 pt Bladder catheterisation 1 pt Any iatrogenic event 1 pt Not weighted. 0 pt=low risk 1-2 pts=intermediate ≥3 pts=high risk RR of high risk: 17.5 (8.1 to 37.4)	None used in model

Continued

Table 2 Continued

External validated DPM name	Citation type of validation	Delirium # (%)	Sens Spec PPV NPV (external)	AUROC (95% CI)	Model components	Cog. assess tool and cut-off
Risk Model for Delirium (RD)	Moerman <i>et al</i> ⁴¹ Narrow val.	Val: 102 (27) (incident delirium)	Sens 0.81 Spec 0.56 PPV 0.41 NPV 0.89 Optimal cut-off score: 4 pts	Val: 0.73 (0.68 to 0.77)	Weighted points Delirium: previous hospitalisation 5 pts Dementia 5 pts Clock drawing ▶ Small mistake 1 pt ▶ Big mistake 2 pts Age ▶ 70–85 years old 1 pt ▶ >85 years 2 pts Impaired hearing 1 pt Impaired vision 1 pt Problems with ADL ▶ Help with meal prep 0.5 pt ▶ help with physical 0.5 pt Use of heroin, methadone, morphine 2 pts Daily >4 alcohol 2 pts ≥5 pts=high risk	CDT –11:10 Two categories 1: small mistakes 2: big mistakes
Susceptibility score	Pendlebury <i>et al</i> ⁴⁸ Broad val.	Val: 308 (28) (incidence delirium)	Sens 0.71 Spec 0.88 PPV 0.5 NPV 0.95 Cut-off score: 5 pts	Val: 0.81 (0.70 to 0.92) Improved with age eliminated to 0.84 (0.77 to 0.92)	Weighted points Dementia/cog impair 2 pts Age >80 years 2 pts Severe illness (SIRS+) 1 pt Infection-working diagnosis 1 pt Vision impairment 1 pt >5 pts=high risk ORs for >5 risk score: 25.0 (3.0 to 208.9) RR for >5 risk score: 5.4	Known diagnosis of dementia or MMSE <24 AMTS <9

ADL, activities of daily living; AMTS, Abbreviated Mental Test Score; AUROC, area under the receiver operating curve statistic; CI, Confidence Intervals; RR, Relative Risk; TPR, True Positive Rate; FPR, False Positive Rate; BUN/CR, Blood Urea Nitrogen/Creatinine ratio; CDT, Clock Drawing Test; CRP, C reactive protein; ETOH, alcohol use; Dev, development; DPM, delirium prediction model; GDS, Geriatric Depression Score; Hx, History; ICU, intensive care unit; IPR, Inouye Prediction Rule; mBDR, Modified Blessed Dementia Rating; bDRS, Blessed Dementia Rating Scale; MMSE, Mini-Mental Status Exam; MoCA, Montreal Cognitive Assessment; mRASS, Modified Richmond Agitation-Sedation Scale; NPV, negative predictive value; PPV, positive predictive value; RN, Registered Nurse; Sens, Sensitivity; Spec, Specificity; SIRS, Systemic Inflammatory Response Syndrome; TIA, Transient Ischemic Attack; TMTYB, the months of the year backwards; VA, Veterans Administration; val, validation.

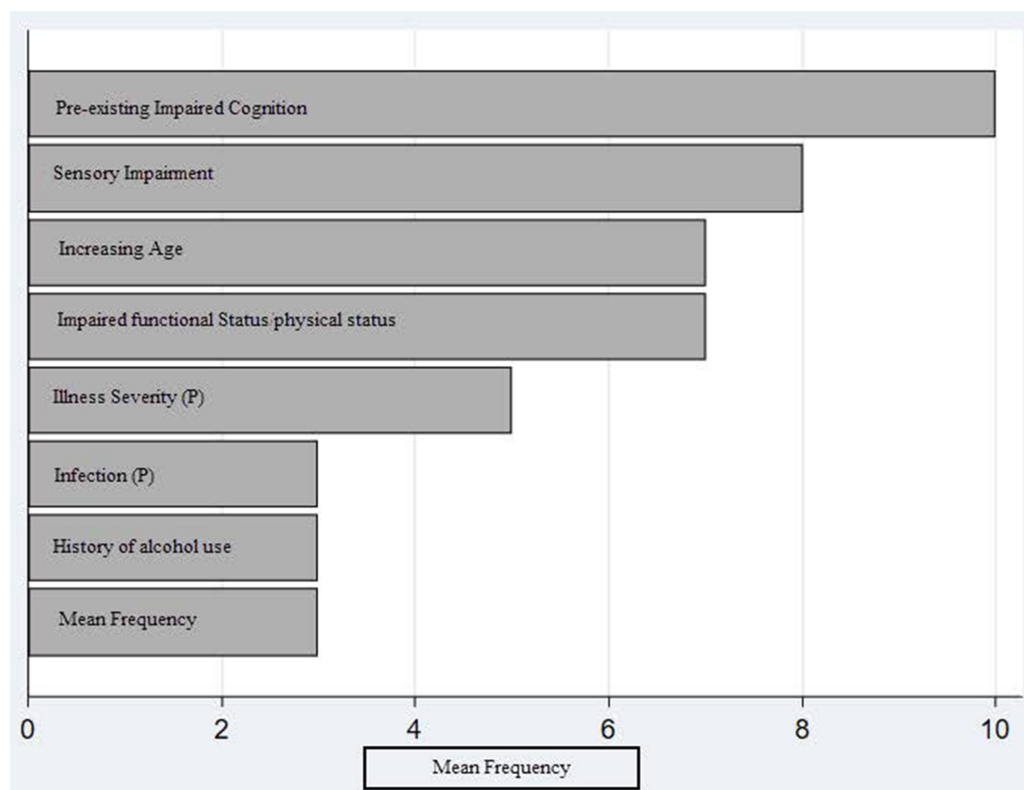


Figure 3 This displays the mean frequency of variable use in the 14 externally validated delirium prediction models. ‘(P)’ indicated a precipitating risk factor used in a delirium prediction model. The following variables were used twice and are not represented in the figure: BUN/Cr ratio (Blood Urea Nitrogen/Creatinine ratio), comorbidities, history of delirium, depression, medications (1: upon admission, 1: added during hospital stay), restraint use and malnutrition (1: altered albumin level, 1: malnutrition scale). The following variables were used once and are not represented in the figure: bladder catheter use, C reactive protein, emergency surgery, presence of fracture on admission, history of cerebrovascular accident, iatrogenic event, intensive care unit admission and open surgery.

reported calibration statistics.^{32 37 40 42} None of the included studies reported calibration plots or slopes.

Risk of overfitting

EPVs were examined in each of the 14 externally validated models. Models estimating more parameters than events in a 1:10 ratio are at risk of statistical overfitting, potentially leading to overly optimistic model performance.^{22 54–57} In 14 models with external validation, four had fewer than optimum events for the number of parameters estimated in the development stage of the models.^{25 29 30 49} Five had fewer than optimum events in the external validation stage.^{23 29–31 45} Two models did not reach optimum events for the number of parameters in either the development or the external validation studies.^{29 30} Various statistical techniques such as shrinkage procedures, the use of lasso or penalised regression and internal validation methods are suggested to counter the effects of lower EPV.^{15 54 58} None of the identified studies report use of statistical shrinkage procedures. Five studies applied internal validation techniques in the development stage of their model to account for stability within their model.^{24 25 37 38 46}

Clinical utility

Clinical utility of a prediction model may be evaluated through several different statistical metrics including ORs, relative risk, sensitivity and specificity, receiver operator curves, R^2 and integrated discrimination improvement indices as well as the clinical utility curve statistic and the decision curve analysis.^{57 59} Six externally validated delirium prediction model studies reported ORs or relative risk statistics evaluating the highest risk stratification cut-off point.^{29–31 34 46 48} Seven studies reported sensitivity and specificity,^{23 27 33 35 41 43 48} and one study reported the rate of true positives and false positives.⁴⁴ None of the identified studies reported decision curve analysis or clinical utility curve analysis. While the majority of studies selected variables that were either routinely used in practice or were feasible to administer, two studies developed delirium prediction models based on data routinely entered into the electronic health record to increase feasibility of use.^{24 44} Pendlebury *et al* adapted variable definition and use to match routine clinical assessment while externally validating four delirium prediction models and creating an additional risk stratification tool.^{33 48} Moerman *et al* reported feasibility and reliability statistics following the incorporation of the risk prediction tool into practice.⁴¹

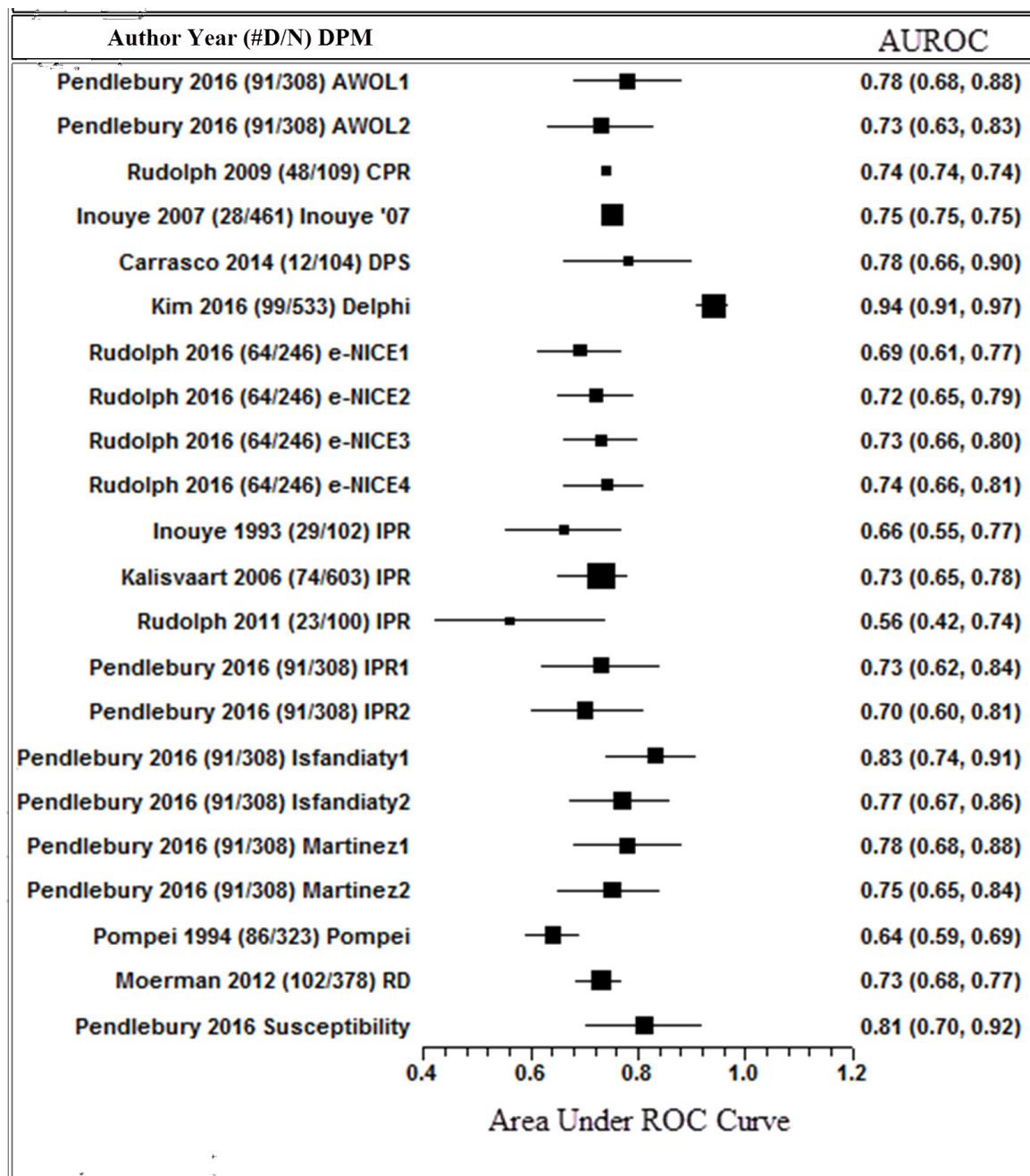


Figure 4 This shows the published AUROC statistic for the 14 externally validated delirium prediction models. #D/N: number of confirmed delirium in study/overall sample size. DPM: delirium prediction model name. The corresponding number of references the different AUROCs calculated based on different cognitive tests applied to the model by the authors. Squares with error bars: size of square corresponds to sample size of study. AUROC: reported area under the receiver curve statistic, 95% CIs.

DISCUSSION

This review identified moderate predictive ability (AUROC 0.52–0.94) in 14 externally validated delirium prediction models with 8 out of 14 models using narrow validation. However, three main limitations were identified. First, study design, application and reporting of statistical methods appear inadequate. Data collection overlapped with the initial diagnosis of delirium in the highest performing model as well as in two other included studies, likely exaggerating model performance.^{15 27 32 35}

Low EPV combined with limited application of internal validation techniques contributed to an increased risk of bias and likely the creation of overly optimistic models.^{15 50–52} Second, broad variable definitions, particularly in functional and cognitive abilities, may have led to overlapping data capture. For example, Pendlebury *et al* demonstrated this possible effect in the development of the *Susceptibility Score*, model performance did not improve with the addition of functional impairment to a model that already included cognitive impairment

and age.⁴⁸ Lastly, assessment of the outcome variable, delirium, was largely non-systematic, once daily and avoided weekends. In the studies that assessed delirium more than once per day, the assessment was performed by routine clinical staff, decreasing consistency. This is a major limitation for an acute condition that fluctuates, may occur suddenly and is dependent on precise, objective assessment. While case-mix between populations may impact observed delirium rates, we believe it would be advantageous for future studies to incorporate systematic, frequent and consistent delirium assessments.

As delirium is a multifactorial syndrome representing an inter-relationship between premorbid and precipitating factors,²⁹ the time course of data collection is important. Nine of the 14 externally validated delirium prediction models incorporate precipitating factors into their predictive model; two models^{29 31} are intentionally constructed in this manner. The inclusion of a precipitating factor into a premorbid delirium prediction model may provide important predictive power if designed in the appropriate manner, as demonstrated by Inouye *et al.*³⁰ However, if variables are collected after the onset of delirium, this would exaggerate model performance (eg, ICU admission). As an example, one delirium prediction model has a robust AUROC of 0.94 (95% CI 0.91 to 0.97).³⁵ This study excluded those with an MMSE <23 and prevalent delirium. Data collection occurred within the first 24 hours following surgery; however, delirium assessment began immediately after surgery, with a 50% delirium prevalence on the day of surgery. This overlap of data collection and delirium assessment likely exaggerated model performance for this outlier study. Seven externally validated models included data about the precipitating factor present on admission and either excluded those with prevalent delirium or calculated separate AUROCs for prevalent delirium versus incident delirium.^{23 30 33 44 48}

Model underperformance may be explained by low powered studies, insufficient EPV as well as the use of univariate analyses and stepwise regression to select predictive variables for inclusion into models. Although these are common methods to use for model development and may counter the effects of insufficient EPV, each approach has significant drawbacks.⁶⁰ Univariate analysis may reduce predictive ability by inclusion of variables that are not independent of each other, and stepwise regression disadvantages include conflation of p values and a biased estimation of coefficients.^{15 22 50 61} While EPV was originally adapted to ensure stability in regression covariates, it has been identified as an important component to predictive model stability and reproducibility due to the result of overfitting.^{15 50 62} Ogundimu *et al* demonstrate this effect by simulating models with EPV of 2, 5, 10, 15, 20, 25 and 50. Stability of models increased as the EPV increased and models including predictors with low population prevalence required >20 EPV.⁶³ The degree of model overfitting should be assessed through calibration statistics and forms of internal validation such as bootstrapping. Future

studies should consider the use of statistical methods to counter low EPV including the application of statistical shrinkage techniques and penalised regression using ridge or lasso regression.^{15 22 56 60 64} Furthermore, future studies may benefit from the incorporation of advanced statistical techniques such as Bayesian Networks and machine learning that have shown to improve the performance of previous prediction models that were built using standard logistic regression.^{65 66} These methods facilitate the exploration of complex interactions between risk factors as well as adapt to changing patient conditions, allowing for a dynamic model.

Increasing age, pre-existing cognitive impairment and functional and sensory impairments were the most frequently used variables in the externally validated delirium prediction models. However, many studies employed different definition for these variables, making comparisons difficult between models and limiting generalisability across populations. Functional and physical impairments were broadly defined resulting in the inability to discern whether impairments resulted from truly physical origins or if the noted decrease in function was related to cognitive impairment leading to an overlap in data collection. Age may not be a relevant risk factor when considering an older cohort of patients; for example, a recent study found that global cognition may mediate the relationship between age and postoperative delirium⁶⁷; therefore, the inclusion of age in a delirium prediction model may not add to the overall performance of the model if cognition is adequately captured or if only elderly patients are included in the study. This effect was demonstrated by Pendlebury *et al*, an improved AUROC resulted when age was removed from the prediction model (0.81 to 0.84).⁴⁸ As the inclusion of age, functional, physical and cognitive impairments may result in an overlap of data collection, future models may want to explore variables that have not been frequently used in delirium prediction yet are highly predictive of mortality, surgical complications and depression. An example would be the self-rated health question. This is a single-item question evaluating an individual's perception of their own health and has been found to be a significant predictor of subjective memory complaints, depression and mortality.^{68–74} Furthermore, this variable is feasible as it takes minimal time and no training. Incorporation of variables such as self-rated health may increase both predictive ability and feasibility, thus improving clinical utility.

The highest performing delirium prediction model excluded those with pre-existing cognitive impairment, did not incorporate a cognitive variable and used hearing impairment as a predictive variable (note the methodological concerns of this study were discussed above).³⁵ Cognitive impairment was the most frequently used variable and is a known risk factor for delirium development.^{2 67} Prior research demonstrates individuals with mild cognitive impairment (MCI) are at a significantly higher risk of delirium development.^{75 76} All models used cut-off

scores on cognitive tests that would indicate dementia, providing no evaluation of subtler cognitive decline such as MCI. Furthermore, Jones *et al* demonstrated a strong linear relationship between risk of delirium and all levels of cognitive function, even those considered unimpaired through formal testing.⁶⁷ In this study, a general cognitive performance score was developed using a complex battery of neuropsychological tests. Unfortunately, the neuropsychological battery is too complex to be practical for the clinical setting. Fong *et al* found associations between baseline executive functioning, complex attention and semantic networks to be associated with subsequent delirium development.⁷⁷ The inclusion of MCI, or simple cognitive tests as employed by Fong *et al*, as a variable may increase the detection and prevalence of cognitive impairment as a variable thus increasing its predictive power. Further exploration into isolated cognitive tests that are feasible to administer in a clinical setting as well as sensitive to the spectrum of cognitive impairment may enhance delirium prediction.

The clinical utility of a prediction model is dependent on both its efficacy at predicting those at risk and feasibility, hence both must be considered when building and validating a model. Clinical utility is compromised by efficacious models that are not feasible. Conversely, a feasible model that is not effective at identifying those at risk also lacks clinical utility. To this end, model derivation must focus on building an effective model. The next aspect that must be considered is the ability to enhance clinical care. Predicting individuals at high risk is clearly important, but to an experienced clinician, delirium may already be anticipated. Maximum value may be obtained by aiding in prediction of moderate risk patients, where the risk of delirium may be more ambiguous.

Strengths and weaknesses of this study

This systematic review benefitted from a prospectively developed protocol. A comprehensive literature search from multiple databases using broad search terms yielded 27 studies with 14 externally validated delirium prediction models. Our author team is interprofessional, providing the opportunity for different perspectives on model evaluation. Furthermore, this review synthesises evidence from both medical and surgical populations while providing statistical-based recommendations for study and model design for future delirium prediction model studies.

The limitations of this systematic review may be that articles focused on a younger population were not included. This limitation could narrow the generalisability of the results of this systematic review to the broader population; however, delirium predominantly affects older adults. Furthermore, this review is limited by population focus. We did not include prediction models built-in palliative care, long-term care facilities or the emergency department.

Strengths and weaknesses in relation to other studies

Past systematic reviews concluded that the identified delirium prediction models were largely heterogeneous in variable inclusion and were not sufficiently developed for incorporation into practice.⁷⁸⁻⁸⁰ Recommendations include further testing on existing delirium prediction models followed by integration in practice as well as further exploration into measurements that are feasible clinically. This review included eight models not previously identified in past systematic reviews of delirium prediction models. Furthermore, this review is the first to identify study and model design issues and discusses the paucity of measurements sensitive to the spectrum of cognitive impairment.

Implications and future research

Two avenues may be pursued for future studies. The first avenue involves model aggregation; currently available delirium prediction models would be combined into a meta-model through stacked regression in a new cohort of participants. This method would update currently published models to a new population, furthering generalisability and bolstering broad external validation.⁸¹ Variable definition could be harmonised in the meta-model with the intention to use variables that are readily available and feasible for routine practice. This method would further delirium prediction for those with dementia-level pre-existing cognitive impairment as well as examine the individual contributions of functional impairment due to physical conditions, cognitive impairment or age through model refitting. Nonetheless, a future meta-model would continue presently identified limitations such as exclusion of the spectrum of cognition. The second avenue should focus on the development and broad validation of delirium prediction models exploring the use of simple cognitive tests that would be inclusive to MCI and sensitive to the spectrum of cognition. Furthermore, future models should consider development of dynamic predictive models using advanced statistical methods such as Bayesian Networks, artificial intelligence and machine learning as these methods have shown to improve models built using standard logistic regression.^{66 82}

We suggest the following broad principles for use in future studies: (1) delirium prediction models should be developed only using data available prior to the onset of delirium and likely should be focused in specific populations depending on whether the precipitating event has occurred or not; (2) should include structured, twice daily assessment (regardless of weekends) using validated tools and trained research staff to identify delirium; (3) should consider inclusion of variables and assessments that are readily available in clinical practice and are feasible to administer without extensive training or interpretation where possible and not to exclude a more informative variable; (4) model development and validation should follow rigorous methods outlined by Steyerberg²² and Steyerberg and Vergouwe⁵⁶ including strategies to counter low sample size and overly optimistic model

performance, the use of Akaike information criterion and Bayesian information criterion to assess model fit and consider broad validations to expand case-mix and generalisability; and (5) adhere to strict guidelines as outlined by the TRIPOD Statement for statistical performance reporting including calibration and clinical utility statistics.^{22 50–52 56 59}

Two classes of delirium prediction models may be required based on the acuity of the admission (elective or emergency). If precipitating factors are included in an elective admission delirium prediction model, where the patient is yet to incur the delirium provoking event, an individual's delirium risk may be overestimated. In the second option, inclusion of only premorbid factors may underestimate delirium risk given the emergency clinical scenario.

CONCLUSION

Twenty-three delirium prediction models were identified. Fourteen of these were externally validated, and three were internally validated. Of the fourteen validated delirium prediction models, the overall predictive ability is moderate with an AUROC range from 0.52 to 0.94. Assessment of the outcome variable, delirium, is often non-systematic, and future studies would be improved with more standardised and frequent assessment. Overall, the variable inclusion and applied definitions in delirium prediction models are heterogeneous, making comparisons difficult. To improve delirium prediction models, future models should consider using standard variables and definitions to work towards a prediction tool that is generalisable to several populations within the remit of understanding the relationship with the precipitating event.

Author affiliations

¹Department of Anesthesiology, University of Wisconsin Madison School of Medicine and Public Health, Madison, Wisconsin, USA

²School of Nursing, University of Wisconsin Madison, Madison, Wisconsin, USA

³School of Nursing, University of Wisconsin-Madison, Madison, Wisconsin, USA

⁴Department of Nursing, University Hospital, Madison, Wisconsin, USA

⁵Department of Anesthesiology, University Hospital RWTH Aachen, Aachen, Germany

⁶Department of Medicine, Western University, London, Ontario, Canada

⁷Anesthesia and Intensive Care, The Chinese University of Hong Kong, Shatin, Hong Kong

⁸MRC Unit for Lifelong Health and Ageing, University College London, London, UK

⁹Division of Anesthesiology Critical Care Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

¹⁴Department of Medicine, Division of Geriatrics, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA

¹⁵Geriatric Research, Education, and Clinical Center (GRECC), William S. Middleton Memorial Veterans Hospital, Madison, Wisconsin, USA

¹⁶Wisconsin Alzheimer's Disease Research Center, Madison, Wisconsin, USA

¹⁷Wisconsin Alzheimer's Institute, Madison, Wisconsin, USA

Acknowledgements We would like to express our sincere gratitude for the assistance of Mary Hitchcock in Ebling Health Sciences Library in the design of the search strategy and Dave Dwyer, Lily Turner, and Casandra Stanfield for their assistance with proofreading the manuscript.

Contributors HL and SP with the mentorship of RDS formulated the aim, developed the study protocol, completed the search and extracted the data. HL and

RDS synthesised the data. HL with the mentorship of RDS drafted the manuscript and designed the tables. RB designed the figures and assisted with statistical interpretation. LB provided expertise on content related to cognition and reviewed the manuscript. DHJD and CMC assisted with synthesis and interpretation of results and discussion in relation to their expertise in geriatrics, cognition and delirium. MC, MM, MTVC and PP assisted with synthesis of results and discussion section, providing expertise in delirium in its respective settings.

Funding HL and RDS acknowledge funding support from the Department of Anesthesiology at University of Wisconsin-Madison. RDS acknowledges funding support from K23 AG055700. PP acknowledges funding support from R01 NHLBI(HL111111) and research grant from Hospira Inc in collaboration with National Institutes of Health.

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Complete search results including excluded studies and CHARMS Risk of Bias checklist decision tree available from corresponding author upon request.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2018. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

REFERENCES

1. Sanders RD, Pandharipande PP, Davidson AJ, *et al.* Anticipating and managing postoperative delirium and cognitive decline in adults. *BMJ* 2011;343:d4331.
2. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet* 2014;383:911–22.
3. Ryan DJ, O'Regan NA, Caoimh RÓ, *et al.* Delirium in an adult acute hospital population: predictors, prevalence and detection. *BMJ Open* 2013;3:e001772.
4. Pandharipande PP, Girard TD, Jackson JC, *et al.* Long-term cognitive impairment after critical illness. *N Engl J Med* 2013;369:1306–16.
5. Leslie DL, Inouye SK. The importance of delirium: economic and societal costs. *J Am Geriatr Soc* 2011;59(Suppl 2):S241–3.
6. Han JH, Brummel NE, Chandrasekhar R, *et al.* Exploring delirium's heterogeneity: association between arousal subtypes at initial presentation and 6-month mortality in older emergency department patients. *Am J Geriatr Psychiatry* 2017;25.
7. Pollard C, Fitzgerald M, Ford K. Delirium: the lived experience of older people who are delirious post-orthopaedic surgery. *Int J Ment Health Nurs* 2015;24:213–21.
8. 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;150:1134–40.
9. Robinson TN, Raeburn CD, Tran ZV, *et al.* Postoperative delirium in the elderly: risk factors and outcomes. *Ann Surg* 2009;249:173–8.
10. Adams ST, Leveson SH. Clinical prediction rules. *BMJ* 2012;344:d8312.
11. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;144:201–9.
12. Moons KG, Kengne AP, Woodward M, *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98:683–90.
13. Debray TP, Moons KG, Ahmed I, *et al.* A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;32:3158–80.
14. Hayhurst CJ, Pandharipande PP, Hughes CG. Intensive care unit delirium: a review of diagnosis, prevention, and treatment. *Anesthesiology* 2016;125:1229–41.
15. Moons KG, de Groot JA, Bouwmeester W, *et al.* Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744.

16. Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg* 2010;8:336–41.
17. Nations U. Ageing. 2016 <http://www.un.org/en/sections/issues-depth/ageing/index.html> (accessed 7 Mar 2017).
18. Centers for Disease C. Elder abuse: definitions. 2016 <https://www.cdc.gov/violenceprevention/elderabuse/definitions.html> (accessed 7 Mar 2017).
19. Barr J, Fraser GL, Puntillo K, *et al.* Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. *Crit Care Med* 2013;41:263–306.
20. Hui D, Frisbee-Hume S, Wilson A, *et al.* Effect of Lorazepam With Haloperidol vs Haloperidol alone on agitated delirium in patients with advanced cancer receiving palliative care: a randomized clinical trial. *JAMA* 2017;318:1047–56.
21. Well GA, Shea B, O'Connell JP, *et al.* The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital, 2014.
22. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. New York: Springer, 2009.
23. Carrasco MP, Villarreal L, Andrade M, *et al.* Development and validation of a delirium predictive score in older people. *Age Ageing* 2014;43:346–51.
24. 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;38:915–23.
25. Douglas VC, Hessler CS, Dhaliwal G, *et al.* The AWOL tool: derivation and validation of a delirium prediction rule. *J Hosp Med* 2013;8:493–9.
26. Fisher BW, Flowerdew G. A simple model for predicting postoperative delirium in older patients undergoing elective orthopedic surgery. *J Am Geriatr Soc* 1995;43:175–8.
27. Freter S, Dunbar M, Koller K, *et al.* Risk of Pre-and post-operative delirium and the Delirium Elderly At Risk (DEAR) tool in hip fracture patients. *Can Geriatr J* 2015;18:212–6.
28. 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.
29. Inouye SK, Charpentier PA. Precipitating factors for delirium in hospitalized elderly persons. Predictive model and interrelationship with baseline vulnerability. *JAMA* 1996;275:852–7.
30. Inouye SK, Viscoli CM, Horwitz RI, *et al.* A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. *Ann Intern Med* 1993;119:474–81.
31. Inouye SK, Zhang Y, Jones RN, *et al.* Risk factors for delirium at discharge: development and validation of a predictive model. *Arch Intern Med* 2007;167:1406–13.
32. Isfandiati R, Harimurti K, Setiati S, *et al.* Incidence and predictors for delirium in hospitalized elderly patients: a retrospective cohort study. *Acta Med Indones* 2012;44:290–7.
33. Pendlebury ST, Lovett NG, Smith SC, *et al.* Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of externally derived risk scores. *Age Ageing* 2016;45:60–5.
34. Kalisvaart KJ, Vreeswijk R, de Jonghe JF, *et al.* Risk factors and prediction of postoperative delirium in elderly hip-surgery patients: implementation and validation of a medical risk factor model. *J Am Geriatr Soc* 2006;54:817–22.
35. Kim MY, Park UJ, Kim HT, *et al.* DELirium prediction based on hospital information (delphi) in general surgery patients. *Medicine* 2016;95:e3072.
36. Korc-Grodzicki B, Sun SW, Zhou Q, *et al.* Geriatric assessment as a predictor of delirium and other outcomes in elderly patients with cancer. *Ann Surg* 2015;261:1085–90.
37. Leung JM, Sands LP, Lim E, *et al.* Does preoperative risk for delirium moderate the effects of postoperative pain and opiate use on postoperative delirium? *Am J Geriatr Psychiatry* 2013;21:946–56.
38. Liang CK, Chu CL, Chou MY, *et al.* Developing a prediction model for post-operative delirium and long-term outcomes among older patients receiving elective orthopedic surgery: a prospective cohort study in Taiwan. *Rejuvenation Res* 2015;18:347–55.
39. Maekawa Y, Sugimoto K, Yamasaki M, *et al.* Comprehensive geriatric assessment is a useful predictive tool for postoperative delirium after gastrointestinal surgery in old-old adults. *Geriatr Gerontol Int* 2016;16.
40. 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.
41. Moerman S, Tuinebreijer WE, de Boo M, *et al.* Validation of the risk model for delirium in hip fracture patients. *Gen Hosp Psychiatry* 2012;34:153–9.
42. O'Keeffe ST, Lavan JN. Predicting delirium in elderly patients: development and validation of a risk-stratification model. *Age Ageing* 1996;25:317–21.
43. Pompei P, Foreman M, Rudberg MA, *et al.* Delirium in hospitalized older persons: outcomes and predictors. *J Am Geriatr Soc* 1994;42:809–15.
44. 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.
45. Rudolph JL, Harrington MB, Lucatorto MA, *et al.* Validation of a medical record-based delirium risk assessment. *J Am Geriatr Soc* 2011;59(Suppl 2):S289–94.
46. 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.
47. Dworkin A, Lee DS, An AR, *et al.* A simple tool to predict development of delirium after elective surgery. *J Am Geriatr Soc* 2016;64:e149–53.
48. 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.
49. Freter SH, George J, Dunbar MJ, *et al.* Prediction of delirium in fractured neck of femur as part of routine preoperative nursing care. *Age Ageing* 2005;34:387–8.
50. 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* 2015;350:g7594.
51. Moons KG, Altman DG, Reitsma JB, *et al.* New Guideline for the Reporting of Studies Developing, Validating, or Updating a Multivariable Clinical Prediction Model: The TRIPOD Statement. *Adv Anat Pathol* 2015;22:303–5.
52. Moons KG, 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.
53. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172:971–80.
54. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res* 2017;26:796–808.
55. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245–7.
56. 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.
57. 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.
58. Peduzzi P, Concato J, Kemper E, *et al.* A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373–9.
59. Campbell DJ. The clinical utility curve: a proposal to improve the translation of information provided by prediction models to clinicians. *BMC Res Notes* 2016;9:219.
60. Pavlou M, Ambler G, Seaman SR, *et al.* How to develop a more accurate risk prediction model when there are few events. *BMJ* 2015;351:h3868.
61. Grobman WA, Stamilio DM. Methods of clinical prediction. *Am J Obstet Gynecol* 2006;194:888–94.
62. Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? *Contemp Clin Trials* 2013;36:636–41.
63. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol* 2016;76:175–82.
64. Steyerberg EW, Eijkemans MJ, Harrell FE, *et al.* Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001;21:45–56.
65. Strobl AN, Vickers AJ, Van Calster B, *et al.* Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators. *J Biomed Inform* 2015;56:87–93.
66. Weng SF, Reps J, Kai J, *et al.* Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;12:e0174944.

67. Jones RN, Marcantonio ER, Saczynski JS, *et al.* Preoperative cognitive performance dominates risk for delirium among older adults. *J Geriatr Psychiatry Neurol* 2016;320–7.
68. Christian LM, Glaser R, Porter K, *et al.* Poorer self-rated health is associated with elevated inflammatory markers among older adults. *Psychoneuroendocrinology* 2011;36:1495–504.
69. Haring R, Feng YS, Moock J, *et al.* Self-perceived quality of life predicts mortality risk better than a multi-biomarker panel, but the combination of both does best. *BMC Med Res Methodol* 2011;11:103.
70. Moore RC, Eyster LT, Mausbach BT, *et al.* Complex interplay between health and successful aging: role of perceived stress, resilience, and social support. *Am J Geriatr Psychiatry* 2015;23:622–32.
71. Arnberg FK, Lekander M, Morey JN, *et al.* Self-rated health and interleukin-6: Longitudinal relationships in older adults. *Brain Behav Immun* 2016;54:226–32.
72. Falk H, Skoog I, Johansson L, *et al.* Self-rated health and its association with mortality in older adults in China, India and Latin America—a 10/66 Dementia research group study. *Age Ageing* 2017;46:932–9.
73. Woo H, Zajacova A. Predictive strength of self-rated health for mortality risk among older adults in the United States: does it differ by race and ethnicity? *Res Aging* 2017;39:879–905.
74. Yates JA, Clare L, Woods RT. What is the relationship between health, mood, and mild cognitive impairment? *J Alzheimers Dis* 2017;55:1183–93.
75. Kazmierski J, Banys A, Latek J, *et al.* Mild cognitive impairment with associated inflammatory and cortisol alterations as independent risk factor for postoperative delirium. *Dement Geriatr Cogn Disord* 2014;38:65–78.
76. Veliz-Reissmüller G, Agüero Torres H, van der Linden J, *et al.* Pre-operative mild cognitive dysfunction predicts risk for post-operative delirium after elective cardiac surgery. *Aging Clin Exp Res* 2007;19:172–7.
77. Fong TG, Hsieh TT, Wong B, *et al.* Neuropsychological profiles of an elderly cohort undergoing elective surgery and the relationship between cognitive performance and delirium. *J Am Geriatr Soc* 2015;63:977–82.
78. van Meenen LC, van Meenen DM, 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.
79. Newman MW, O'Dwyer LC, Rosenthal L. Predicting delirium: a review of risk-stratification models. *Gen Hosp Psychiatry* 2015;37:408–13.
80. Campaner CP. Valoración del riesgo de delirium en pacientes mayores hospitalizados. Delirium risk assessment in elderly hospitalized patients. *Rev Esp Geriatr Geront* 2010;45:285–90.
81. Debray TP, Koffijberg H, Nieboer D, *et al.* Meta-analysis and aggregation of multiple published prediction models. *Stat Med* 2014;33:2341–62.
82. Kim SY, Moon SK, Jung DC, *et al.* Pre-operative prediction of advanced prostatic cancer using clinical decision support systems: accuracy comparison between support vector machine and artificial neural network. *Korean J Radiol* 2011;12:588–94.