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

Heidi Lindroth, Lisa Bratzke, Suzanne Purvis, Roger Brown, Mark Coburn, Marko Mrkobrada, Matthew T V Chan, Daniel H J Davis, Pratik Pandharipande, Cynthia M Carlsson, Robert D Sanders

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, SociINFO, 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. 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 5 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.15 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...
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). 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 predetermined modifiable and non-modifiable risk factors of delirium present. This review included studies focused on (1) older adult (≥60 years) population, (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, (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) for case-control and cohort studies. Risk of bias was assessed through the Critical Appraisal and Data Extraction for Systematic Reviews (CHARMS) checklist. 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. 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 curve analysis were also collected from each delirium prediction model in reference to the model’s reported cut-off value. Goodness-of-fit statistics including $\chi^2$ 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.

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. 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).

Twenty-three delirium prediction models were developed, 14 were externally validated and three were internally validated. Prospective cohort design was used in 24 studies. Retrospective design was
used in four studies. Nineteen studies used consecutive sampling methods, two of these were part of a randomised control trial. Eleven studies focused on the medical population, and 13 recruited a surgical population (seven orthopaedic, one cardiac, two non-cardiac, one general surgery, and two oncological). None of the identified studies focused on critical care patients. Data collection occurred on admission in 17 studies; participants were approached within 48 hours of admission. Seven studies collected data preoperatively then followed participants postoperatively. Data collection overlapped with delirium assessments in three studies. 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, 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. The frequency of delirium assessment varied from two or more assessments daily (3 studies) to once daily (12 studies) every other day (8 studies), once following surgery and undefined (3 studies). 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. Twenty-one studies used trained research or clinical personnel to conduct the delirium assessments. Three studies relied on delirium diagnosis, or keywords designated as representing delirium, to identify the outcome measure through retrospective chart review. Three studies relied on clinical staff to recognise and chart delirium symptoms. One of these studies retrospectively confirmed the diagnosis of delirium through consensus review of two authors; disagreement was resolved by a psychiatrist. One study did not report details on personnel performing delirium assessments.
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 \(<0.25\)\(^{24,25}\) or a relative risk of \( \geq 1.50\).\(^{30} \) 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.\(^{23-25,29,30,35-37,44,47,48} \) 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.\(^{26} \) 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 studies
Table 1 Displays the 27 studies that were identified for inclusion in this review.

<table>
<thead>
<tr>
<th>Author</th>
<th>Study design population sample size sampling method power analysis</th>
<th>Study grade (NOS)</th>
<th>Outcome variable and rate (%)</th>
<th>Delirium measurement and frequency</th>
<th>DPM name and regression model used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrasco et al23</td>
<td>P.Cohort Medical Dev: 374 Val: 104 Consecutive</td>
<td>S: **** C: ** O: ** T: 6 stars</td>
<td>Delirium Dev: 25 (0.08) Val: 12 (12)</td>
<td>CAM Every 48 hours</td>
<td>Predictive Risk Score Forward stepwise</td>
</tr>
<tr>
<td>Dworkin et al47</td>
<td>P.Cohort Elective non-cardiac surg Dev: 76 Consecutive Power analysis</td>
<td>S: **** C: – O: ** T: 6 stars</td>
<td>Delirium Dev: 10 (13)</td>
<td>CAM or FAM-CAM 1x after surgery</td>
<td>Mini-Cog Stratified into a five-point score Stepwise</td>
</tr>
<tr>
<td>Freter et al49</td>
<td>P.Cohort Hip Fx Dev: 100 Consecutive Power analysis</td>
<td>S: ** C: ** O: ** T: 6 stars</td>
<td>Delirium Dev: 24 (24)</td>
<td>CAM Daily</td>
<td>Risk stratification model (DEAR)</td>
</tr>
<tr>
<td>Freter et al27</td>
<td>P.Cohort Hip fracture Dev: 283 Consecutive Power analysis</td>
<td>S: *** C: – O: ** T: 5 stars</td>
<td>Delirium Val: 119 (42)</td>
<td>CAM POD1, 3 and 5</td>
<td>Risk stratification model (DEAR)</td>
</tr>
<tr>
<td>Inouye and Charpentier29</td>
<td>P.Cohort Medical Dev: 196 Val: 312 Consecutive Power analysis</td>
<td>S: **** C: ** O: *** T: 9 stars</td>
<td>Delirium Dev: 35 (18) Val: 47 (15)</td>
<td>CAM Every other day</td>
<td>Risk stratification model based on precipitating factors Backwards and forwards stepwise</td>
</tr>
<tr>
<td>Author</td>
<td>Study design</td>
<td>Population sample size</td>
<td>Sampling method</td>
<td>Power analysis</td>
<td>Study grade (NOS)</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------</td>
<td>------------------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Inouye et al</td>
<td>P.Cohort</td>
<td>Medical</td>
<td>Consecutive</td>
<td></td>
<td>S: **** C: ** O: **** T: 9 stars</td>
</tr>
<tr>
<td>Liang et al</td>
<td>P.Cohort</td>
<td>Elective orthopaedic Surgery</td>
<td>Consecutive</td>
<td></td>
<td>S: *** C: ** O: ** T: 7 stars</td>
</tr>
<tr>
<td>Maekawa et al</td>
<td>P.Cohort</td>
<td>Oncological; gastrointestinal surgery</td>
<td>Consecutive</td>
<td></td>
<td>S: ** C: * O: *** T: 6 stars</td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Author</th>
<th>Study design</th>
<th>Population</th>
<th>Sampling method</th>
<th>Power analysis</th>
<th>Study grade (NOS)</th>
<th>Outcome variable and rate (%)</th>
<th>Delirium measurement and frequency</th>
<th>DPM name and regression model used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martinez et al</td>
<td>P. Cohort</td>
<td>Medical</td>
<td>Dev: 397</td>
<td>Val: 302</td>
<td>Consecutive</td>
<td>S: ***</td>
<td>Delirium</td>
<td>CAM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: **</td>
<td>Dev: 52 (13)</td>
<td>Val: 76 (25)</td>
</tr>
<tr>
<td>O'Keeffe and Lavan</td>
<td>P. Cohort</td>
<td>Acute geriatric unit</td>
<td>Dev: 100</td>
<td>Val: 84</td>
<td>Consecutive</td>
<td>S: ****</td>
<td>Delirium</td>
<td>CAM</td>
</tr>
<tr>
<td>Pendlebury et al</td>
<td>P. Cohort</td>
<td>Medical</td>
<td>Dev: 308</td>
<td>Val: 84</td>
<td>Consecutive</td>
<td>S: ****</td>
<td>Delirium</td>
<td>CAM</td>
</tr>
<tr>
<td>Pendlebury et al</td>
<td>P. Cohort</td>
<td>Medical</td>
<td>Dev: 308</td>
<td>Val: 84</td>
<td>Consecutive</td>
<td>S: ****</td>
<td>Delirium</td>
<td>CAM</td>
</tr>
<tr>
<td>Rudolph et al</td>
<td>P. Cohort</td>
<td>Medical</td>
<td>Dev: 100</td>
<td>Val: 100</td>
<td>Consecutive</td>
<td>S: ****</td>
<td>Delirium</td>
<td>DSM-IV</td>
</tr>
</tbody>
</table>

**Table 1** Continued
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. Two studies additionally evaluated chronic cognitive impairment through family or caregiver interview with the modified Blessed Dementia Rating Scale (mBDRS). Four models combined the cognitive test score derived on enrolment with a history of dementia to define baseline cognitive impairment. History of dementia was defined as follows: two studies: family or caregiver report supplemented with documented history in medical record, one study: medical record review and interview with mBDRS and one study: dementia billing codes or prescription information. One study defined baseline cognitive impairment as a prespecified key term in the electronic health.

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), (1 study) domestic help, help with meals or physical care, and (2 studies) residence in nursing facility or at home with caregivers, and (2 studies) requiring a home care package with professional caregivers or residence in a care home. The latter being obtained on admission from medical records. 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.

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. Two models reported an external validation AUROC above 0.80, indicating moderate predictive ability. 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. The reported $\chi^2$ statistics were significant in five prognostic models and did not reach significance in one model. Four of the 23 studies that developed models...
### Table 2  Detailed description of the externally validated DPMs.

<table>
<thead>
<tr>
<th>External validated DPM name</th>
<th>Citation type of validation</th>
<th>Delirium # (%)</th>
<th>Sens Spec PPV NPV (external)</th>
<th>AUROC (95% CI)</th>
<th>Model components</th>
<th>Cog. assess tool and cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWOL tool</td>
<td>Pendlebury et al (2016) Broad val.</td>
<td>1st val: 14 (9) 2nd val: 95 (31) (any delirium) 67: prevalent 28: incident</td>
<td>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</td>
<td>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)</td>
<td>Original AWOL Tool Age &gt;80 years 1 pt Failure to spell WORLD backwards 1 pt Disorientation 1 pt Illness severity 1 pt</td>
<td>MMSE &lt;24 AMTS &lt;9</td>
</tr>
<tr>
<td>Clinical prediction rule: cardiac surgery</td>
<td>Rudolph et al (2009) Narrow val.</td>
<td>Dev: 63 (52) Val: 48 (44) (incident delirium)</td>
<td>Not reported</td>
<td>Dev: 0.74 Val: 0.75 Did not report CI</td>
<td>Weighted points-regression MMSE &lt;23 2 pts MMSE 24–27 1 pt Hx of stroke/TIA 1 pt GDS &gt;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)</td>
<td>MMSE -stratified score</td>
</tr>
<tr>
<td>DEAR</td>
<td>Freter et al (2005) Narrow val.</td>
<td>Dev: (2005) 18 (14) Val: (2015) Preop= 163 (58) Postop= 118 (42)</td>
<td>Sens 0.68 Spec 0.73 PPV 0.65 NPV 0.76 Optimal cut-off score: 3 pts (incident postop delirium)</td>
<td>Dev: (2005) 0.77 (0.64 to 0.87) Val: (2015) AUROC not published</td>
<td>MMSE &lt;23 1 pt Functional dependence 1 pt Sensory impairment 1 pt Substance use 1 pt Age &gt;80 years 1 pt Not weighted, 0–5 score, cut-off of 3 indicating high risk.</td>
<td>MMSE Cut-off ≤23</td>
</tr>
<tr>
<td>External validated DPM name</td>
<td>Citation type of validation</td>
<td>Delirium # (%)</td>
<td>Sens Spec PPV NPV (external)</td>
<td>AUROC (95% CI)</td>
<td>Model components</td>
<td>Cog. assess tool and cut-off</td>
</tr>
<tr>
<td>----------------------------</td>
<td>----------------------------</td>
<td>----------------</td>
<td>-----------------------------</td>
<td>----------------</td>
<td>------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Delirium at discharge prediction model</td>
<td>Narrow val.</td>
<td>Dev: 58 (12) Val: 28 (6) (incident delirium)</td>
<td>Not reported</td>
<td>Dev: 0.80 Val: 0.75 Did not report CI Calibration: χ² trend: p&lt;0.001</td>
<td>Delirium at discharge prediction Dementia diagnosis or mBDRS &gt;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)</td>
<td>MMSE &lt;24 mBDR ≥4</td>
</tr>
<tr>
<td>Delirium Prediction Score (DPS)</td>
<td>Narrow val.</td>
<td>Dev: 25 (0.06) Val: 12 (12) (incident delirium)</td>
<td>Sens 0.88 Spec 0.74 PPV 0.22 NPV 0.99</td>
<td>Dev: 0.86 (0.82 to 0.91) Val: 0.78 (0.66 to 0.90)</td>
<td>DPS=[5×BUN/Cr ratio]−(3× Barthel Index). Cut-off is: &gt;−240=high risk for delirium In conventional units, cut-off is: &gt;−160=high risk for delirium</td>
<td>None. Pfeffer Functional Activities Questionnaire as a proxy for prior dementia</td>
</tr>
<tr>
<td>External validated DPM name</td>
<td>Citation type of validation</td>
<td>Delirium # (%)</td>
<td>Sens Spec PPV NPV (external)</td>
<td>AUROC (95% CI)</td>
<td>Model components</td>
<td>Cog. assess tool and cut-off</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>----------------</td>
<td>----------------------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Delphi score</td>
<td>Kim et al&lt;sup&gt;36&lt;/sup&gt;</td>
<td></td>
<td>Dev: 112 (20) Val: 99 (18)</td>
<td>Sens 0.81 Spec 0.93 PPV 0.70 NPV 0.96</td>
<td>Dev: 0.911 (0.88 to 0.94) Val: 0.938 (0.91 to 0.97)</td>
<td>Age (years) 60–69 0 70–79 1 &gt;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) &lt;10 0 &gt;10 1 Max points: 15 Optimal cut-off: 6.5 High risk: &gt;7 pts</td>
</tr>
</tbody>
</table>

Table 2 Continued
<table>
<thead>
<tr>
<th>External validated DPM name</th>
<th>Citation type of validation</th>
<th>Delirium # (%)</th>
<th>Sens Spec PPV NPV (external)</th>
<th>AUROC (95% CI)</th>
<th>Model components</th>
<th>Cog. assess tool and cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>e-NICE rule</td>
<td>Rudolph et al.</td>
<td>Broad val.</td>
<td>Dev: 2343 (8)</td>
<td>Val: 64 (28) (incident delirium)</td>
<td>Delirium AUROC CI TPR FPR</td>
<td>Dev: 0.81 (0.80 to 0.82)</td>
</tr>
<tr>
<td>Inouye Prediction Rule (IPR)</td>
<td>Inouye et al.</td>
<td>Narrow val.</td>
<td>Dev: 27 (25)</td>
<td>Val: 29 (17) (incident delirium)</td>
<td>Did not report</td>
<td>Dev: 0.74 (0.63 to 0.85)</td>
</tr>
<tr>
<td>IPR</td>
<td>Kalisvaart et al.</td>
<td>Broad val.</td>
<td>Dev: 74 (12)</td>
<td>Did not report</td>
<td>Val: 0.73 (0.65 to 0.78)</td>
<td>Calibration: $X^2$ p&lt;0.05 $X^2$ trend $p&lt;0.002$</td>
</tr>
<tr>
<td>IPR</td>
<td>Rudolph et al.</td>
<td>Broad val.</td>
<td>Dev: 23 (23)</td>
<td>Any delirium 10: prevalent 13: incident</td>
<td>Did not report</td>
<td>Val: 0.56 (0.42 to 0.74)</td>
</tr>
</tbody>
</table>
Table 2  Continued

<table>
<thead>
<tr>
<th>External validated DPM name</th>
<th>Citation type of validation</th>
<th>Delirium # (%)</th>
<th>Sens Spec PPV NPV (external)</th>
<th>AUROC (95% CI)</th>
<th>Model components</th>
<th>Cog. assess tool and cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPR</td>
<td>Pendlebury et al</td>
<td></td>
<td>Cut-off 2 pts All delirium</td>
<td>Val: Incident delirium Cohort 1 (MMSE) 0.73 (0.62 to 0.84) Cohort 2 (AMTS) 0.70 (0.60 to 0.81)</td>
<td>Baseline cognitive impairment 1 pt High BUN/Cr ratio 1 pt Severe illness (SIRS &gt;2) 1 pt Vision impairment 1 pt</td>
<td>Original model: MMSE &lt;24 Modified model: MMSE &lt;24 AMTS &lt;9</td>
</tr>
<tr>
<td></td>
<td>Broad val.</td>
<td>Val: 95 (31)</td>
<td>Any 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</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>67: prevalent</td>
<td>28: incident</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ispandiaty model

| Dev: 87 (19)                | Pendlebury et al            |                | Cut-off 4 pts All delirium  | 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 | Original model: Chart review Modified model: MMSE <24 AMTS <9 |
|                             | Broad val.                  | Val: 95 (31)   | Any 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 |                |                 |              |
|                             |                             | 67: prevalent  | 28: incident                |                |                 |              |

Table 2  Continued

Continued
### Table 2  Continued

<table>
<thead>
<tr>
<th>External validated DPM name</th>
<th>Citation type of validation</th>
<th>Delirium # (%)</th>
<th>Sens PPV NPV (external)</th>
<th>AUROC (95% CI)</th>
<th>Model components</th>
<th>Cog. assess tool and cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martinez et al/ 2012 model</td>
<td>Broad val.</td>
<td>1st Val: 76 (25) 2nd Val: 95 (31) Any delirium 67: prevalent 28: incident</td>
<td>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</td>
<td>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)</td>
<td>Martinez et al/ 2012 original model Age &gt;85 years 1 pt Dependent in &gt;5ADLs 1 pt Drugs on admit: 1 pt/drug 2 pts/antipsych ▶ antidepressants ▶ anticonvulsants ▶ antipsychotics Score 0–3 Score ≥1=high risk for delirium Modified model Age &gt;85 years 1 pt Dependency in &gt;5ADLs 1 pt diag of dementia MMSE &lt;24 AMTS &lt;9 1 pt</td>
<td></td>
</tr>
<tr>
<td>Pompei et al/ 1994 model</td>
<td>Broad val.</td>
<td>Dev: 64 (15) Val: 86 (26) (21=prevalent delirium)</td>
<td>Sens 0.83 Spec 0.50 PPV 0.38 NPV 0.89 *Pts stratified as low or moderate to high risk</td>
<td>Dev: 0.74 ±0.05 Val: 0.64 ±0.05 Calibration: X²trend p&lt;0.0001</td>
<td>Weighted points Baseline cognitive impairment 2 pts Depression 2 pts Alcoholism 3 pts &gt;4 comorbidities 3 pts 0–3 pts=low risk 4–7 pts=moderate risk 8–10 pts=high risk</td>
<td>MMSE Less than high school &lt;21 High school &lt;23 College education &lt;24</td>
</tr>
<tr>
<td>Precipitating risk factors</td>
<td>Narrow val.</td>
<td>Dev: 35 (18) Val: 47 (15) (incident delirium)</td>
<td>Not reported</td>
<td>No AUROC reported Calibration: X²trend p&lt;0.001</td>
<td>Physical restraint use 1 pt Malnutrition 1 pt &gt;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)</td>
<td>None used in model</td>
</tr>
</tbody>
</table>
### Table 2 Continued

<table>
<thead>
<tr>
<th>External validated DPM name</th>
<th>Citation type of validation</th>
<th>Delirium # (%)</th>
<th>Sens Spec PPV NPV (external)</th>
<th>AUROC (95% CI)</th>
<th>Model components</th>
<th>Cog. assess tool and cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Model for Delirium (RD)</td>
<td>Narrow val.</td>
<td>Val: 102 (27)  (incident delirium)</td>
<td>Sens 0.81 Spec 0.56 PPV 0.41 NPV 0.89 Optimal cut-off score: 4 pts</td>
<td>Val: 0.73 (0.68 to 0.77)</td>
<td>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 ► &gt;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 &gt;4 alcohol 2 pts ≥5 pts=high risk</td>
<td>CDT − 11:10 Two categories 1: small mistakes 2: big mistakes</td>
</tr>
<tr>
<td>Susceptibility score</td>
<td>Broad val.</td>
<td>Val: 308 (28)  (incidence delirium)</td>
<td>Sens 0.71 Spec 0.88 PPV 0.5 NPV 0.95 Cut-off score: 5 pts</td>
<td>Val: 0.81 (0.70 to 0.92) Improved with age eliminated to 0.84 (0.77 to 0.92)</td>
<td>Weighted points Dementia/cog impair 2 pts Age &gt;80 years 2 pts Severe illness (SIRS+) 1 pt Infection-working diagnosis 1 pt Vision impairment 1 pt ORs for &gt;5 risk score: 25.0 (3.0 to 208.9) RR for &gt;5 risk score: 5.4</td>
<td>Known diagnosis of dementia or MMSE &lt;24 AMTS &lt;9</td>
</tr>
</tbody>
</table>

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.
reported calibration statistics. 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. 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. Five had fewer than optimum events in the external validation stage. Two models did not reach optimum events for the number of parameters in either the development or the external validation studies. 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. 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.

### 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, R² and integrated discrimination improvement indices as well as the clinical utility curve statistic and the decision curve analysis. Six externally validated delirium prediction model studies reported ORs or relative risk statistics evaluating the highest risk stratification cut-off point. Seven studies reported sensitivity and specificity, and one study reported the rate of true positives and false positives. 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. 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. Moerman et al reported feasibility and reliability statistics following the incorporation of the risk prediction tool into practice.

---

**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.
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.\cite{15,27,32,35} Second, broad variable definitions, particularly in functional and cognitive abilities, may have led to overlapping data capture. For example, Pendlebury \textit{et al} demonstrated this possible effect in the development of the \textit{Susceptibility Score}, model performance did not improve with the addition of functional impairment to a model that already included cognitive impairment.

**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.
and age. 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, 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 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. 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). 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.

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. 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. 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. 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. 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. 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. 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 delirium; 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). 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. 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). Cognitive impairment was the most frequently used variable and is a known risk factor for delirium development. Prior research demonstrates individuals with mild cognitive impairment (MCI) are at a significantly higher risk of delirium development. 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 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 Akaïke 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.

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

1. Department of Anesthesiology, University of Wisconsin Madison School of Medicine and Public Health, Madison, Wisconsin, USA
2. School of Nursing, University of Wisconsin Madison, Madison, Wisconsin, USA
3. School of Nursing, University of Wisconsin-Madison, Madison, Wisconsin, USA
4. Department of Nursing, University Hospital, Madison, Wisconsin, USA
5. Department of Anesthesiology, University Hospital RWTH Aachen, Aachen, Germany
6. Department of Medicine, Western University, London, Ontario, Canada
7. Anesthesia and Intensive Care, The Chinese University of Hong Kong, Shatin, Hong Kong
8. MRC Unit for Lifelong Health and Ageing, University College London, London, UK
9. Division of Anesthesiology Critical Care Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
10. Department of Medicine, Division of Geriatrics, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA
11. Geriatric Research, Education, and Clinical Center (GRECC), William S. Middleton Memorial Veterans Hospital, Madison, Wisconsin, USA
12. Wisconsin Alzheimer’s Disease Research Center, Madison, Wisconsin, USA
13. Wisconsin 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 AG057500. PP acknowledges funding support from RO1 NHLBIHL111111 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


