BMJ Open

Systematic Review of Prediction Models for Delirium in the Older Adult Inpatient

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019223
Article Type:	Research
Date Submitted by the Author:	18-Aug-2017
Complete List of Authors:	Lindroth, Heidi; University of Wisconsin Madison School of Medicine and Public Health, Department of Anesthesiology; University of Wisconsin Madison Graduate School, School of Nursing Bratzke, Lisa; University of Wisconsin-Madison, School of Nursing Purvis, Suzanne; University of Wisconsin Hospital and Clinics, Nursing Brown, Roger; University of Wisconsin Madison, School of Nursing Coburn, Mark; Uniklinik RWTH Aachen, Anaesthesiology Mrkobrada, Marko; Western University, Medicine Chan, MTV; The Chinese University of Hong Kong, Anesthesia and Intensive Care Davis, Daniel; University College London, MRC Unit for Lifelong Health and Ageing ; University College London Pandharipande, Pratik; Vanderbilt University School of Medicine Carlsson, Cynthia; William S. Middleton Veteran's Administration Geriatric Research Education and Clinical Center (GRECC), GRECC; University of Wisconsin Madison School of Medicine and Public Health Sanders, Robert; University of Wisconsin Madison School of Medicine and Public Health, Department of Anesthesiology
Primary Subject Heading:	Geriatric medicine
Secondary Subject Heading:	Neurology
Keywords:	Delirium, GERIATRIC MEDICINE, Statistic

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 TV Chan, Daniel Davis, Pratik Pandharipande, Cynthia M. Carlsson, Robert D. Sanders
University of Wisconsin School of Medicine and Public Health, Department of Anesthesiology, School of Nursing, 600 Highland Avenue, Madison, WI 53792 Heidi Lindroth BSN PhD Candidate[1]

University of Wisconsin School of Nursing, 701 Highland Avenue, Madison, WI 53705 Lisa Bratzke PhD Assistant Professor[2]

University Hospital, 600 Highland Avenue, Madison, WI 53792 Suzanne Purvis DNP, Clincial Nurse Specialist[3]

University of Wisconsin School of Nursing, 701 Highland Avenue, Madison, WI 53705 Roger L. Brown PhD, Professor[4]

University Hospital RWTH Aachen, Department of Anesthesiology, Pauwelsstrasse 30, 52074 Aachen, Mark Coburn MD, Professor[5]

Western University, Schulich School of Medicine and Dentistry, 339 Windermere Road, Room B9-100, London ON N5X 3W6, Marko Mrkobrada MD Associate Professor[6]

The Chinese University of Hong Kong, Department of Anaesthesia and Intensive Care, 4/F Main Clinical Bock and Trauma Center, Prince of Wales Hospital, Hong Kong Special Administrative Region, China, Matthew TV Chan, MB, BS, PhD, Professor[7]

MRC Unit for Lifelong Health and Ageing at UCL, 33 Bedford Place, London, WC1B 5JU, Daniel Davis PhD MRCP, Senior Clinical Researcher[8]

Vanderbilt University Medical Center, Department of Anesthesiology, 1211 Medical Center Drive, Nashville TN 37322, Pratik Pandharipande MD, Chief, Division of Anesthesiology Critical Care Medicine[9]

University of Wisconsin School of Medicine and Public Health, Madison VA Geriatric Research Education and Clinical Center (GRECC) (D4211), 2500 Overlook Terrace, Madison, WI, 53705, Cynthia M. Carlsson, MD, MS, Associate Professor of Medicine, Division of Geriatrics[10]

University of Wisconsin School of Medicine and Public Health, Department of Anesthesiology, 600 Highland Avenue, Madison, WI 53792, Robert D. Sanders MBBS PhD FRCA, Assistant Professor[11]

[+]**Corresponding author**: *Heidi Lindroth, Department of Anesthesiology, University of Wisconsin School of Medicine and Public Health, 600 Highland Avenue, B6/319 CSC Madison, WI 53792-3272 Telephone: 608-262-6469 Fax: 608-263-0575 Madison, USA. Email:* hlindroth@wisc.edu

Word count: 4043

Objective

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

Design

Systematic review

Data Sources and methods

PubMed, CINAHL, PsychINFO, SocINFO, Cochrane, ISI and EMBASE were searched from 1990/1/1 to 2016/12/31. The PRISMA Statement guided protocol development. Inclusion criteria: Age ≥60, 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. First author synthesized data. Mentoring author resolved disagreement.

Eligibility Criteria

Inclusion criteria: Age ≥60, inpatient, developing or validating an existing prognostic delirium prediction model. Exclusion criteria: alcohol-related delirium, sample size ≤ 50. Data were extracted from published studies. The primary performance measures were calibration and discrimination statistics. Secondary measures included applied statistical methodology.

Results

The initial search resulted in 7,502 studies. Following full-text review of 192 studies, 33 were excluded based on age criteria (<60yrs) and 27 met the defined criteria. Twenty-three delirium prediction models were identified, thirteen were externally validated and three 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. Five models demonstrated an AUROC >0.75, indicating moderate predictive ability. Limitations in design, data collection methods, and calibration 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.

Funding

Department of Anesthesiology, University of Wisconsin-Madison.

R01 NHLBI (HL111111) and research grant from Hospira Inc., NIH collaboration

All authors completed ICMJE disclosure forms and no conflicts of interest declared.

Keywords

Delirium. Aging. Cognition. Prediction. Statistical Models.

Strengths and Limitations of this Study

- The PRISMA Statement and CHARMS checklist were used to develop the protocol for this systematic review.
- Interprofessional authorship providing different perspectives on delirium prediction models.
- Comprehensive search using multiple databases and search terms
- Limited by age (≥60yo)
- Limited to studies developing or validating predictive models, did not include predictive risk factors

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] Delirium has been independently associated with increased mortality, morbidity in terms of impaired cognition and functional disability along with an estimated annual expenditure of $164 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 formalize 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. 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 PRISMA Statement and the CHARMS checklist (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 predetermined modifiable and non-modifiable risk factors of delirium present. This review included studies focused on 1) older adult (≥ 60 years) population, (the U.S. CDC and UN define an older adult as 60 years of age and older)[17 18], 2) inpatient hospital setting, 3) publication dates of 1990/1/1 to 2016/12/31, and 4) developed and/or validated delirium prediction models. Studies were excluded if they 1) studied a different patient population (i.e. emergency department, skilled nursing facilities, palliative care, and hospice) as these are not generalizable to an inpatient hospital setting, 2) related to alcohol withdrawal, or delirium tremens, as the presence of alcohol withdrawal complicates delirium assessment, and 3) had a sample size ≤ 50 for methodological reasons (i.e. underpowered). All study designs were included. Studies were not limited by timeframe 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, PsychINFO, Cochrane Database of Systematic Reviews, SocINDEX, ISI, 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)[19] for case-control and cohort studies. Two authors (HL, SP) independently performed data collection, data extraction, and assessed study quality, with any disagreement resolved by RDS.

Data extracted included: 1) study characteristics (study design, population, 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 4) characteristics of the models (variables used in model, 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. We chose to designate delirium prediction models with an AUROC greater than 0·75, albeit arbitrary, as clinically relevant.[20] Sensitivity, specificity, positive predictive values and negative predictive values were also collected from each delirium prediction model. Goodness-of-fit statistics including Chi-square (X^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.[21] Secondary pre-planned outcome measures included cognitive assessments, and predictive variable use per model.

Role of the Funding Source

The funding sources named has 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.

RESULTS

Twenty-seven studies were identified for inclusion.[22-46] 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. Two studies that included younger populations in the development cohort for the delirium prediction model were included due to the subsequent older external validation cohort thus meeting our inclusion criteria (age $\geq$ 60).[24 39]

Twenty-three delirium prediction models were developed, thirteen were externally validated [22 26 28-30 32-34 40 42-45] and three were internally validated.[23 36 41] Prospective cohort design was used in 23 studies.[22 24-30 32-34 36-48] Retrospective design was used in four studies.[23 31 35 43] Eleven studies focused on the medical population [22 24 28-32 39 41 44 48], three included medical and surgical [23 42 43] and thirteen recruited a surgical population (seven-orthopaedic [25-27 33 37 40 47], one-cardiac [45], two-noncardiac [36 46], one general surgery[34], two-oncological[35 38]). Data collection occurred upon admission in seventeen studies [22 24 26 28-30 32-34 39-44 47 48]; participants were approached within forty-eight hours of admission. Seven studies collected data pre-operatively then followed participants post-operatively.[25 27 36-38 45 46] The average NOS quality ranking for included cohort studies was seven; five studies received the maximum of nine stars. Further characteristics of studies are listed in Table 1.

Table 1

Author	Study Design Population Sample Size	Study Grade (NOS)	Outcome Variable & Rate (%)	Delirium measurement & frequency	DPM Design & (Name)
Carrasco et al. (2014)[22]	P.Cohort Medical Dev: 374 Val: 104	S: **** C: - O: ** T: 6 stars	Delirium Dev: 25 (.06) Val: 12 (12)	CAM Every 48 h	Predictive Risk Score
de Wit et al. (2016)[23]	Retro All hospital patients Dev: 1291	S: *** C: ** O: *** T: 8 stars	Delirium Dev: 225(17)	Chart abstraction EHR "diagnosis table"	Automated Delirium Prediction Model
Douglas et al.** (2013)[24]	P.Cohort Medical Dev: 209 Val: 165	S: **** C: - O: *** T: 7 stars	Delirium Dev: 25(12) Val: 14(8.5)	CAM-S & CAM Daily	Risk Stratification model (AWOL)
Dworkin et al. (2016)[46]	P.Cohort Elective noncardiac surg Dev: 76	S: **** C: - O: ** T: 6 stars	Delirium Dev: 10(13)	CAM or FAM-CAM 1xafter surgery	Mini-Cog Stratified into a five point score
Fisher and Flowerdew (1995)[25]	P.Cohort Elective Orthopedic Dev: 80	S: ** C: - O: ** T: 4 stars	Delirium Dev: 14(17.5)	CAM 2xDaily	Prediction Model using two variables.
Freter et al. (2005)[27]	P.Cohort Elective Hip surgery Dev: 132	S: ** C: ** O: ** T: 6 stars	Delirium Dev: 18(14)	CAM Daily	Risk Stratification Model (DEAR)
Freter et al. (2005)[47]	P.Cohort Hip Fx Dev: 100	S: ** C: ** O: ** T: 6 stars	Delirium Dev: 24(24)	CAM Daily	Risk Stratification Model (DEAR)
Freter et al. (2015)[26]	P.Cohort Hip Fracture Val: 283	S: *** C: - O: ** T: 5 stars	Delirium Val: 119(42)	CAM POD1, 3 & 5	Risk stratification model (DEAR)
Inouye and Charpentier (1996)[28]	P.Cohort Medical Dev: 196 Val: 312	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 35(18) Val: 47(15)	CAM Every other day	Risk stratification model based on precipitating factors
Inouye et al. (2007)[30]	P.Cohort Medical Dev: 491 Val: 461	S: **** C: ** O: *** T: 9 stars	Delirium/ subsyndrome delirium at discharge Dev: 58(12) Val: 28(6)	CAM Every other day	Risk stratification model
Inouye et la. (1993)[29]	P.Cohort Medical Dev: 107 Val: 174	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 27(25) Val: 29(17)	CAM Daily	Risk stratification model
Isfandiaty et al. (2012)[31]	Retro Medical Dev: 457	S: ** C: - O: *** T: 5 Stars	Delirium Dev: 87(19)	Undefined Daily	Risk stratification model
Kalisvaart et al. (2006)[33]	P.Cohort Hip Surgery & Facture Val: 603	S: *** C: - O: *** T: 6 stars	Delirium Dev: 74(12)	CAM, DRS-98 Daily through POD5	Externally validated Inouye's '93 model.
Kim et al. (2016)[34]	P.Cohort Major General Surgery Dev: 561 Val: 533	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
Korc-Grodzicki et al. (2014)[35]	Retro Oncological Surgery Dev: 416	S: *** C: - O: *** T: 6 stars	Delirium Dev: 79(19)	CAM Daily	Comprehensive Geriatric Assessment (CGA) as model.
Leung et al. (2013)[36]	P.Cohort Noncardiac surgery Dev: 581	S: *** C: - O: ** T: 5 stars	Delirium Dev: 234(40)	CAM Daily	Risk stratification model

Liang et al. (2015)[37]	P.Cohort Elective Orthopedic Surgery Dev: 461	S: *** C: ** O: ** T: 7 stars	Delirium Dev: 37(8)	CAM Daily Confirmed by psychologist DSM-IV	Built 2 DPMs CGA Risk stratification models
Maekawa et al. (2015)[38]	P.Cohort Oncological; Gastrointestinal Surgery Dev: 517	S: ** C: * O: *** T: 6 stars	Delirium Dev: 124(24)	CAM Unknown frequency	Comprehensive Geriatric Assessment (CGA) as model.
Martinez et al.(2012)[39]**	P.Cohort Medical Dev: 397 Val: 302	S: *** C: - O: ** T: 5 stars	Delirium Dev: 52(13) Val: 76(25)	CAM Undefined	Clinical prediction rule
Moerman et al. (2012)[40]	P.Cohort Hip Fracture Val: 378	S: *** C: - O: *** T: 6 stars	Delirium Val: 102(27)	Ward RN observation, 3xdaily Confirmed by chart review.	Risk stratification model (Risk Model for Delirium, RD)
O'Keeffe and Lavan (1996)[41]	P.Cohort Acute Geriatric Unit Dev: 100 Ival: 84	S: **** C: - O: ** T: 6 stars	Delirium Dev: 28(28) IVal: 25(30)	DAS Every 48 hours DSM III	Risk Stratification model
Pendlebury et al. (2016)[48]	P. Cohort Medical Dev: 308	S: **** C: * O: *** T: 8 stars	Delirium Dev: 95(31)	CAM Every 48-hours Confirmed by DSM-IV interview	Susceptibility Score
Pendlebury et al. (2016)[32]	P.Cohort Medical Val: 308	S: **** C: - O: *** T: 7 stars	Delirium Val: 95(31)	CAM Every 48-hours Confirmed by DSM-IV interview	Externally validated 4 DPMs
Pompei et al. (1994)[42]	P.Cohort Med/surg Dev: 432 V: 323	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 64(14.8) Val: 86(26.3)	CAM 2xweekly. Confirmed with DSM III	Risk stratification model
Rudolph et al. (2009)[45]	P.Cohort Cardiac Surgery Dev: 122 V: 109	S: *** C: * O: ** T: 6 stars	Delirium Dev: 63(52) Val: 48(44)	CAM, MDAS, DSI Daily	Risk stratification model
Rudolph et al. (2011)[44]	P.Cohort Medical V: 100	S: **** C: - O: *** T: 7 stars	Delirium Dev: 23(23)	DSM-IV Daily clinical interview	Externally validated Inouye's '93 model.
Rudolph et al. (2016)[43]	Dev: Retro Val: P.Cohort Med/surg Dev: 27625 Val: 246	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

Key:
**=Models developed in population ≤60 years of age, but validated in population ≥60 years of age.
Study Design: P.Cohort=Prospective Cohort, Retro=Retrospective design. Dev=Development, Val=Validation. Med=Medical, Surg=Surgical.
Study Grade: NOS=Newcastle Ottawa Scale. S=Selection, C=Comparability, O=Ottawa. Max 9 stars.
Outcome Variable: Dev=Development, Val=Validation
Delirium Measurement: CAM=Confusion Assessment Method, DSM=Diagnostic Statistical Manual, POD=Postoperative Day, MDAS=Memorial Delirium Assessment Scale, Nu-Desc=Nursing Delirium Screening Scale, DRS-98=Delirium Rating Scale, EHR=Electronic Health Record
Type of Model: How authors designed their delirium prediction model (DPM)
-Risk stratification model: Points (weighted or un-weighted) assigned per predictive risk factor present.
-CGA=Comprehensive Geriatric Assessment

Delirium assessment

The outcome variable was measured using the Confusion Assessment Method in twenty-one studies.[22 24-30 32-39 42 45-48] The frequency of delirium assessment varied from two or more assessments daily (three studies)[25 34 40], to once daily (twelve studies)[24 27 29 31 33 35-37 43-45 47], every-other day (eight studies)[22 26 28 30 32 41 42 48], once following surgery[46], and undefined (three studies).[23 38 39] 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.[25 34 40] The principal investigator confirmed the presence of delirium following the nurse report of symptoms.[25 34] Twenty-one studies used trained research or clinical personnel to conduct the delirium assessments.[22 24-26 28-30 32-39 42-46 48] Three studies relied on delirium diagnosis, or keywords designated as representing delirium, to identify the outcome measure through retrospective chart review.[23 31 43] Three studies relied on clinical staff to recognize and chart delirium symptoms.[27 40 47] One of these studies retrospectively confirmed the diagnosis of delirium through consensus review of two authors, disagreement was resolved by a psychiatrist.[40] One study did not report details on personnel performing delirium assessments.[41]

Model design and statistical methods

Various statistical techniques were employed by the thirteen externally validated delirium prediction models in the selection of variables for model inclusion. Five used univariate or bivariate analyses and selected variables with a pre-determined statistical value (range for $p<0.05$ to $p<0.25$) for inclusion in the model.[22 24 39 42 45] One of these models paired bivariate analyses with a bootstrapping technique to address lower sample and event size.[45] Three models based their variable selection from a literature review of risk factors for delirium.[26 27 40 43 47] Two used proportional hazards regression modeling paired with bivariate analyses and included variables

with either a *p*-value <0.25[31] or a relative risk of ≥1.5.[29] Five studies published their power analysis.[24 32 34 39 45] To further refine and test the estimated models, the following methods were employed: seven studies-stepwise logistic regression (LR),[22 24 29 34 39 42 45], four studies-multivariate LR[26 31 33 40], one study-continuation ratio model combined with log-binomial regression model[30], one study-multivariable binomial regression.[28]

Variables

Figure two demonstrates the frequency of variable use in the thirteen externally validated delirium prediction models. Baseline cognitive impairment was the most frequently used variable. Five models defined baseline cognitive impairment as a cognitive test score at or below the level of dementia.[26 29 33 42] This cognitive test was administered upon study enrollment. One study additionally evaluated chronic cognitive impairment through family or caregiver interview with the modified Blessed Dementia Rating Scale (mBDRS).[29 30] Four models combined the cognitive test score derived upon enrollment with a history of dementia to define baseline cognitive impairment.[30 32 40 43] History of dementia was defined as follows: Two studies-family or caregiver report supplemented with documented history in medical record [32 40], one study-medical record review and interview with mBDRS[30], and one study-dementia billing codes or prescription information.[43] One study defined baseline cognitive impairment as a pre-specified key term in the electronic health.[44]

Functional impairment was defined as follows: (1) needing assistance with any basic ADL,[26] (1) domestic help, help with meals or physical care[40] and (2) residence in nursing facility or at home with caregivers.[32] Two studies used validated functional assessment tools (iADL and Barthel Index) and evaluated functional status two weeks prior to hospitalization.[22 30]

Externally validated delirium prediction models are detailed in Table 2.

Table 2

External Validated DPM Name	Citation	Delirium #(%)	Sens Spec PPV NPV (external)	AUROC (95%CI)	Model Components	Cog. Assess Tool & Cutoff
AWOL Tool	Pendlebury et al. (2016) [34]	1st Val: 14(9) 2nd Val: 95(31) (any delirium) 67-prevalent 28-incident	Mod. AWOL Cutoff - 3 Any Delirium Sens ·7 Spec .66 PPV .55 NPV .79 Incident Del Sens .76 Spec .66 PPV .27 NPV .94	1st Val: 0.69 (0·54-0·83) Incident delirium 2nd Val: Cohort 1 (MMSE) 0·78 (0·68-0·88) Cohort 2 (AMTS) 0·73 (0·63-0·83)	Original AWOL Tool Age >80 — 1 pt Failure to spell WORLD backwards — 1 pt Disorientation — 1 pt Illness Severity — 1 pt Modified AWOL Tool Age >80 — 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 et al. (2009) [47]	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 pt 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	MMSE -Stratified score
DEAR	Freter et al. (2015)[28]	Dev: (2005) 18(14) Val: (2015) Pre-Op= 163(58) Post-op= 118(42)	Sens .68 Spec .73 PPV .65 NPV .76 Optimal cut-off score: 3pts (Incident post-op delirium)	Dev: (2005) 0·77 (0·64-0·87) Val: (2015) AUROC Not published	MMSE ≤ 23 — 1 pt Functional dependence — 1 pt Sensory impairment — 1 pt Substance use — 1 pt Age >80 — 1 pt Not weighted. 0-5 Score, cut-off of 2 or 3 indicating high risk.	MMSE Cut-off ≤ 23
Delirium at Discharge Prediction Model	Inouye et al. (2007)[32]	Dev: 58(12) Val: 28(6) (incident delirium)	Not reported	Dev: 0.80 Val: 0.75 Did not report CI	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 pt = Intermediate Risk 4-5 pt = High Risk	MMSE < 24 mBDR ≥ 4
Delirium Prediction Score (DPS)	Carrasco et al. (2014)[24]	Dev: 25(.06) Val: 12(12)	Sens .88 Spec .74 PPV .22	Dev: 0.86 (0.82-0.91)	DPS=[5xBUN/Cr ratio]-(3xBarthel Index). Cut off is: > -240 = High risk for Delirium	None. Pfeffer

<table>
<tr><td></td><td></td><td>(incident delirium)</td><td>NPV .99</td><td>Val: 0.78 (0.66-0.90)</td><td>In conventional units, cut-off is:
> -160 = High Risk for Delirium</td><td>Functional Activities Questionnaire as a proxy for prior dementia</td></tr>
<tr><td>Delphi Score</td><td>Kim et al. (2016)[36]</td><td>Dev: 112(20)
Val: 99(18)

(incident delirium)</td><td>Sens .81
Spec .93
PPV .70
NPV .96

Optimal cut-off score: 6.5pts</td><td>Dev: 0.911 (0.88-0.94)
Val: 0.938 (0.91-0.97)</td><td><table><tr><td>Age (years)</td><td></td></tr><tr><td>60-69</td><td>0</td></tr><tr><td>70-79</td><td>1</td></tr><tr><td>≥80</td><td>2</td></tr><tr><td>Low Physical Activity</td><td></td></tr><tr><td>Self-sufficient</td><td>0</td></tr><tr><td>Need assist.</td><td>2</td></tr><tr><td>Heavy ETOH</td><td></td></tr><tr><td>No</td><td>0</td></tr><tr><td>Yes</td><td>1</td></tr><tr><td>Hearing Impairment</td><td></td></tr><tr><td>No</td><td>0</td></tr><tr><td>Yes</td><td>1</td></tr><tr><td>History of delirium</td><td></td></tr><tr><td>No</td><td>0</td></tr><tr><td>Yes</td><td>2</td></tr><tr><td>Emergency Surgery</td><td></td></tr><tr><td>No</td><td>0</td></tr><tr><td>Yes</td><td>1</td></tr><tr><td>Open Surgery</td><td></td></tr><tr><td>No</td><td>0</td></tr><tr><td>Yes</td><td>2</td></tr><tr><td>ICU Admission</td><td></td></tr><tr><td>No</td><td>0</td></tr><tr><td>Yes</td><td>3</td></tr><tr><td>Pre-Op CRP (mg/dL)</td><td></td></tr><tr><td><10</td><td>0</td></tr><tr><td>≥10</td><td>1</td></tr></table>Max points: 15
Optimal cut-off: 6.5
High Risk: ≥7 points</td><td>No measure of cognition.

Excluded participants if MMSE <24</td></tr>
<tr><td>e-NICE Rule</td><td>Rudolph et al. (2016)[45]</td><td>Dev: 2343(8)
Val:64(26)

(incident delirium)</td><td><table><tr><td>Cohort</td><td>AUROC</td><td>CI</td></tr><tr><td>Dev</td><td>0.81</td><td>(0.80-0.82)</td></tr><tr><td colspan="3">Validation AUROCs*</td></tr><tr><td>Original</td><td>0.69</td><td>(0.61-0.77)</td></tr><tr><td>mRASS</td><td>0.72</td><td>(0.65-0.79)</td></tr><tr><td>TMYB</td><td>0.73</td><td>(0.66-0.80)</td></tr><tr><td>MoCA</td><td>0.74</td><td>(0.66-0.81)</td></tr></table>*Any delirium
Original model-AUROC of 0.68 (95%CI0.59-0.77) in incident delirium.
Did not report sens, spec, PPV, NPV</td><td><table><tr><td colspan="2">Weighted Points/OR</td></tr><tr><td>Cog impair -Medications, diagnosis or both</td><td>4 pt</td></tr><tr><td>Age ≥ 65 y</td><td>2 pt</td></tr><tr><td>Age ≥ 80 y</td><td>3 pt</td></tr><tr><td>Infection</td><td>2 pt</td></tr><tr><td>Fracture</td><td>4 pt</td></tr><tr><td>Vision</td><td>1 pt</td></tr><tr><td>Severe Illness</td><td>2 pt</td></tr></table>0-2 pts = Low Risk
2-5 pts = Intermediate Risk
6-8 pts = High Risk
≥ 9 pts = Very High Risk</td><td>e-NICE Tool
Diagnosis of dementia, medications for dementia or both qualified as “cognitive impairment” in model.

Prospective cohort, additional:
-mRASS
-TMYB
-MoCA ≤ 18</td></tr>
<tr><td>Inouye Prediction Rule (IPR)</td><td>Inouye et al. (1993)[31]</td><td>Dev: 27(25)
Val: 29(17)

(incident delirium)</td><td>Did not report</td><td>Dev: 0.74 (0.63-0.85)
Val: 0.66 (0.55-0.77)</td><td><table><tr><td>Baseline cognitive impairment</td><td>1 pt</td></tr><tr><td>High BUN/Cr ratio</td><td>1 pt</td></tr><tr><td>Severe illness (Composite score: APACHE II >16 + RN rating)</td><td>1 pt</td></tr><tr><td>Vision impairment</td><td>1 pt</td></tr></table>Not weighted.
0 pts = Low risk</td><td>MMSE Cut-off < 24

Family/care giver bDRS

Excluded</td></tr>
</table>

					1-2 pts = Intermediate risk 3-4 pts = High risk	those w/history of severe dementia
IPR	Kalisvaart et al. (2006)[35]	Val: 74(12)	Did not report	Val: 0.73 (0.65-0.78)	Externally validated IPR in surgical hip fracture population.	MMSE Cut-off < 24
IPR	Rudolph et al. (2011)[46]	Val: 23(23) Any delirium 10-Prevalent 13-Incident	Did not report	Val: 0.56 (0.42-0.74) Incident delirium	Externally validated IPR in medical VA population.	MMSE Cut-off < 24
IPR	Pendlebury et al. (2016)[34]	Val: 95(31) Any delirium 67-prevalent 28-incident	Cutoff 2pts All Delirium Sens .57 Spec .80 PPV .64 NPV .76 Incident D Sens .52 Spec .80 PPV .31 NPV .91	Val: Incident delirium Cohort 1 (MMSE) 0.73 (0.62-0.84) Cohort 2-(AMTS) 0.70 (0.60-0.81)	Baseline cognitive impairment: 1 pt High BUN/Cr ratio: 1 pt Severe illness (SIRS ≥ 2): 1 pt Vision impairment: 1 pt 4pts=Incident delirium	Original Model: MMSE <24 Modified Model: MMSE < 24 AMTS < 9
Isfandiaty model	Pendlebury et al. (2016)[34]	Dev: 87(19) Val: 95 (31) Any delirium 67-prevalent 28-incident	Cutoff 4pts Any Delirium Sens .74 Spec .71 PPV .60 NPV .82 Incident Del Sens .81 Spec .71 PPV .31 NPV .96	Dev: 0.82 (0.77-0.88) Val: Incident delirium Cohort 1 (MMSE) 0.83 (0.74-0.91) Cohort 2 (AMTS) 0.77 (0.67-0.86)	Baseline cognitive impairment: 3 pt Functional dependency: 2 pt Infection w/sepsis: 2 pt Infection w/out 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
Martinez et al. 2012 model	Pendlebury et al. (2016)[34]	1st Val: 76(25) 2nd Val: 95(31) Any delirium 67-prevalent 28-incident	Modified Model Cutoff 2pts Any Delirium Sens .62 Spec .68 PPV .54 NPV .75 Incident Del Sens .81 Spec .68 PPV .29 NPV .96	1st Val: 0.85 (0.80-0.88) Incident delirium 2nd Val: Cohort 1 (MMSE) 0.78 (0.68-0.88) Cohort 2 (AMTS) 0.75 (0.65-0.84)	Martinez et al. 2012 Original Model Age >85: 1 pt Dependent in ≥5 ADLs: 1 pt Drugs on admit: -Antidepressants -Antidementia -anticonvulsants -antipsychotics: 1pt/drug 2pt/ antipsych Score 0-3 Score >1 = high risk for delirium Modified Model Age >85: 1 pt Dependency in ≥ 5 ADLs: 1 pt Diag of Dementia MMSE<24 AMTS<9: 1 pt	Original Model: -No cognitive measure Modified Model: MMSE < 24 AMTS < 9

Pompei et al. 1994 model	Pompei et al. (1994)[44]	Dev: 64(15) Val: 86(26) (21=prevalent delirium)	Sens .83 Spec .50 PPV .38 NPV .89 *Pts stratified as low or moderate to high-risk	Dev: 0.74 +/- 0.05 Val: 0.64 +/- 0.05	Weighted Points Baseline cognitive impairment: 2 pt Depression: 2 pt Alcoholism: 3 pt ≥ 4 comorbidities: 3 pt 0-3 pts = Low risk 4-7 pts = Moderate risk 8-10 pts = High risk	MMSE Less than HS <21 High school <23 College edu < 24
Precipitating Risk Factors	Inouye and Charpentier (1996)[30]	Dev: 35(18) Val: 47(15) (incident delirium)	Not reported	No AUROC reported	Physical restraint use: 1 pt Malnutrition: 1 pt ≥3 medications added: 1 pt Bladder catherization: 1 pt Any iatrogenic event: 1 pt Not weighted. 0 pt = Low Risk 1-2 pt = Intermediate ≥ 3 pt = High Risk	None used in model
Risk Model for Delirium (RD)	Moerman et al. (2012)[42]	Val: 102(27) (incident delirium)	Sens .81 Spec .56 PPV .41 NPV .89 Optimal cut-off score: 4 pts	Val: 0.73 (0.68-0.77)	Weighted Points Delirium-previous hospitalization: 5 pt Dementia: 5 pt Clock Drawing: -Sm mistake: 1 pt -big mistake: 2 pt Age: -70 to 85 years old: 1 pt - >85 years: 2 pt Impaired hearing: 1 pt Impaired vision: 1 pt Problems w/ADL: -Help w/meal prep: .5p -help w/physical: .5p Use of heroin, methadone, morphine: 2 pt Daily >4 alcohol: 2 pt ≥ 5 pts = High risk	CDT -11:10 -Two Categories 1 Small mistakes 2 Big mistakes

Key:
Dev=Development, Val=Validation
Sens=Sensitivity, Spec=Specificity, PPV=Positive Predictive Value, NPV=Negative Predictive Value
Area Under the Receiver Operating Curve Statistic, Dev=Development, Val=Validation, mRASS=Modified Richmond Agitation-Sedation Scale, TMTYB=The Months of the Year Backwards
ADL=Activities of Daily Living
MMSE=Mini Mental Status Exam, AMTS=Abbreviated Mental Test Score, CDT=Clock Drawing Test, mBDR=Modified Blessed Dementia Rating, MoCA=Montreal Cognitive Assessment.

Predictive ability

Reported AUROC in externally validated delirium prediction models ranged from 0.52-0.94 (Figure three). Five models attained an AUROC above 0.75 indicating potential clinical

relevance and moderate predictive ability.[22 32 34] Of these five models, the highest performing model (AUROC 0.94, CI 0.91-0.97) was developed and validated in a surgical population.[34] Carrasco et al. (2014) was developed and validated in a medical population (AUROC 0.78, CI 0.66-0.90).[22] The remaining three models were developed in separate medical cohort populations[24 31 39] but, were externally validated within the same cohort of medical patients and modified to share similar variable measures of cognition, functional status and illness severity (AUROC 0.78-0.83).[32] These five models share similarities with variable use, as seen in Figure two.

Model calibration

Four of the thirteen externally validated delirium prediction models reported calibration metrics.[28 29 33 44] The reported chi-square statistics were significant in three models[28 29 33] and did not reach significance in one model.[44] None of the included studies reported Hosmer-Lemeshow test statistics, calibration plots or slopes.

Risk of overfitting

Events per variable (EPV) were examined in each of the thirteen externally validated models. Models estimating more parameters than events in a 1:10 ratio are at risk of statistical overfitting.[15 49 50] In 13 models with external validation, four had fewer than optimum events for the number of parameters estimated in the development stage of the models.[24 28 29 47] Five had fewer than optimum events in the external validation stage.[22 28-30 44] Two models did not reach optimum events for the number of parameters in either the development or the external validation studies.[28 29] Of the five models with an AUROC greater than 0.75, one of these models did not obtain sufficient EPV in the development stage[24] and another did not attain sufficient EPV in the external validation study, likely impacting the model's predictive ability

(development – AUROC 0.86, CI-0.82-0.91, external validation – AUROC 0.78, CI-0.66-0.90).[22]

DISCUSSION

This review identified moderate predictive ability in five of the thirteen externally validated delirium prediction models, however, three main limitations were identified. First, assessment of the outcome variable, delirium, was largely non-systematic, once daily, and avoided weekends. This is a major limitation for an acute condition that fluctuates and may occur suddenly. In the highest performing model, a major limitation was identified: data collection overlapped with the initial diagnosis of delirium, likely exaggerating model performance.[15 34] Second, model performance may be influenced by inadequate EPV leading to statistical overfitting and exaggerated model performance. Overall reporting of model performance measures was inconsistent with only four models reporting calibration statistics. Finally, variable definition was heterogeneous and often indistinct, making comparisons between models difficult and decreasing the ability to generalise models across populations. Further, broad variable definitions, particularly in functional and cognitive abilities, may have led to overlapping data capture. Pendlebury et al. (2016) facilitated comparisons between three of the moderately performing models by externally validating these in the same cohort.[32] These models were re-developed to best fit the validation cohort. While this provides an opportunity to compare models, it is not known how these models will generalise to subsequent patient populations. Re-development is not equal to model validation.[15] Taken together, these findings suggest that current delirium prediction is limited by moderately performing, heterogeneous, non-generalizable models that may be improved with the application of frequent, systematic delirium assessments and the use of applicable statistical methods to evaluate and build clinical prediction

models.

As delirium is a multifactorial syndrome representing an interrelationship between premorbid and precipitating factors,[28] the time course of data collection is important. Eight of the thirteen externally validated delirium prediction models incorporate precipitating factors into their predictive model; two models [28 30] 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 (1993).[29] However, if variables are collected after the onset of delirium this would exaggerate model performance (e.g. ICU admission). As an example, one delirium prediction model has a robust AUROC of 0.94 (CI 0.91-0.97).[34] This study excluded those with a 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. The remaining three models with AUROCs greater than 0.75 included data about the precipitating factor present upon admission and either excluded those with prevalent delirium or calculated separate AUROCs for prevalent delirium versus incident delirium.

Model underperformance may be explained through low powered studies leading to insufficient events per variable (EPV) resulting in statistical overfitting.[49 50] As overfitting of a model leads to an underestimation of event probability in low risk patients and overstates the probability in high risk patients, it is an important consideration when evaluating the predictive performance of delirium prediction models.[51] This effect is highlighted in the Carrasco et al.(2014) model as the AUROC decreased from the development study (0.82) to the external validation study (0.78).

Future studies should attain adequate EPV to avoid overfitting. Further, past models validated with insufficient EPV should be interpreted with caution.

The identified studies largely used univariate or bivariate analysis then stepwise logistic regression to develop the delirium prediction models. Although these are common methods to use for model development and may counter the effects of insufficient EPV, each approach has significant drawbacks.[51] 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.[21 52] Statistical methods to counter low EPV could include penalised regression using either ridge or lasso regression and bootstrapping.[21 51]

Increasing age, pre-existing cognitive impairment, 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. Interestingly, these variables were also not consistently included in the five highest performing delirium prediction models, questioning their potential role in delirium prediction. 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[53] 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.

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).[34] Cognitive impairment was the most frequently used variable and is a known risk factor for delirium development.[2 53] Prior research demonstrates individuals with Mild Cognitive Impairment (MCI) are at a significantly higher risk of delirium development.[54 55] 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. (2016) demonstrated a strong linear relationship between risk of delirium and all levels of cognitive function, even those considered unimpaired through formal testing.[53] 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. (2015) found associations between baseline executive functioning, complex attention and semantic networks to be associated with subsequent delirium development[56]. The inclusion of MCI, or simple cognitive tests as employed by Fong et al. (2015), 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.

Four of the best-performing models contained a measurement of functional or physical impairment.[22 32 34] This measurement may be representative of numerous underlying factors working to inhibit a biological compensatory mechanism and serve as a marker for a vulnerable individual.[57 58] Carrasco et al. (2014) used the Barthel Index, which evaluates basic functioning in ten different areas. A proxy completed this measure instead of self-report which has been

shown to improve accuracy.[59] Kim et al. (2016) did not report use of a standardized measurement tool, but defined impaired physical status as the inability to be self-sufficient. Pendlebury et al. (2016) defined functional impairment as an individual residing in a care home or receiving care at their home and applied this definition in two of the four models validated within that patient cohort. These broad definitions lead to the inability to discern whether the functional impairment was due to a physical or cognitive mechanism.

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 twenty-seven studies with thirteen externally validated delirium prediction models. Our author team is interprofessional, providing the opportunity for different perspectives on model evaluation. Further, this review synthesizes 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 along with studies identifying predictive risk factors, not exclusively predictive models. This limitation could narrow the generalisability of the results of this systematic review to the broader population however delirium predominantly affects older adults.

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.[60-62] 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. Further 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

Future studies should focus on the development and validation of delirium prediction models using the following broad principles: (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 explore the use of further cognitive variables to enhance current model performance and should distinguish functional impairment due to physical conditions, cognitive impairment or both, (3) should include structured, twice daily assessment (regardless of weekends) using validated tools and trained research staff to identify delirium, (4) adhere to strict guidelines for both statistical methodology and metric reporting, (5) Delirium prediction model variables should have sufficient prevalence along with the number of events within the population studied to optimize model performance and (6) consider development of dynamic predictive models using AI methods and machine learning. In addition, rigorous statistical methods would improve the development and validation of models and avoid issues of under and overfitting of models. An example of this would be to employ Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) in stepwise selection. This would avoid exclusion of variables that may not be statistically significant in standard hypothesis testing, yet may yield important variable prediction in model estimations.[21] Standardized metric reporting would augment model

development and validation, facilitating the ability to compare model across populations and settings. Recommendations for future statistical reporting of delirium prediction models include sensitivity, specificity, positive predictive value, negative predictive value, Nagelkerke's R^2, area under the receiver operating curve (AUROC), and goodness-of-fit measures. Further, calculating and reporting statistical metrics on the calibration and clinical usefulness of models would benefit delirium prediction.[21]

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. Thirteen of these were externally validated and three were internally validated. Of the thirteen validated delirium prediction models, the overall predictive ability is moderate with only five models achieving an AUROC above 0.75.[22 32 34] Assessment of the outcome variable, delirium, is often non-systematic and future studies would be improved with more standardized 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 generalizable to several populations within the remit of understanding the relationship with the precipitating event.

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 synthesized 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. DD 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.

Declaration of Interests

All authors have completed ICMJE disclosure forms and no conflicts of interest are declared.

Acknowledgements

HL and RDS acknowledge funding support from the Department of Anesthesiology at University of Wisconsin-Madison. PP acknowledges funding support from R01 NHLBI (HL111111) and research grant from Hospira Inc. in collaboration with NIH. 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 and Lily Turner for their assistance with proofreading the manuscript.

References

1. Sanders RD, Pandharipande PP, Davidson AJ, et al. Anticipating and managing postoperative delirium and cognitive decline in adults. *BMJ: British Medical Journal (Overseas & Retired Doctors Edition)* 2011;343(7816):197-200 4p. doi: 10.1136/bmj.d4331

2. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet* 2014;383(9920):911-22. doi: 10.1016/s0140-6736(13)60688-1

10.1016/S0140-6736(13)60688-1. Epub 2013 Aug 28.</p>

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(1) doi: 10.1136/bmjopen-2012-001772

4. Pandharipande PP, Girard TD, Jackson JC, et al. Long-Term Cognitive Impairment after Critical Illness. *The New England Journal Of Medicine* 2013;369(14):1306-16. doi: 10.1056/NEJMoa1301372

5. Leslie DL, Inouye SK. The importance of delirium: economic and societal costs. *Journal of the American Geriatrics Society* 2011;59 Suppl 2:S241-3. doi: 10.1111/j.1532-5415.2011.03671.x [published Online First: 2011/12/07]

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. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* 2016 doi: 10.1016/j.jagp.2016.05.016 [published Online First: 2016/09/14]

7. Pollard C, Fitzgerald M, Ford K. Delirium: The lived experience of older people who are delirious post-orthopaedic surgery. *International Journal of Mental Health Nursing* 2015;24(3):213-21. doi: 10.1111/inm.12132

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 Surgery* 2015;150(12):1134-40. doi: 10.1001/jamasurg.2015.2606

9. Robinson TN, Raeburn CD, Tran ZV, et al. Postoperative delirium in the elderly: risk factors and outcomes. *Annals of surgery* 2009;249(1):173-8. doi: 10.1097/SLA.0b013e31818e4776 [published Online First: 2008/12/25]

10. Adams ST, Leveson SH. Clinical prediction rules. *BMJ (Clinical research ed)* 2012;344

11. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of internal medicine* 2006;144(3):201-09.

12. Moons KGM, 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(9):683-90. doi: 10.1136/heartjnl-2011-301246

13. Debray TPA, Moons KGM, Ahmed I, et al. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in medicine* 2013;32(18):3158-80. doi: 10.1002/sim.5732

14. Hayhurst CJ, Pandharipande PP, Hughes CG. Intensive Care Unit DeliriumA Review of Diagnosis, Prevention, and Treatment. *Anesthesiology* 2016;125(6):1229-41. doi: 10.1097/ALN.0000000000001378

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 medicine* 2014;11(10):e1001744. doi: 10.1371/journal.pmed.1001744 [published Online First: 2014/10/15]

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(5):336-41. doi: 10.1016/j.ijsu.2010.02.007 [published Online First: 2010/02/23]

17. Nations U. Ageing 2016 [Available from: http://www.un.org/en/sections/issues-depth/ageing/index.html accessed March 7th 2017.

18. Centers for Disease C. Elder Abuse: Definitions 2016 [Available from: https://www.cdc.gov/violenceprevention/elderabuse/definitions.html accessed March 7th 2017.

19. Well GA, Shea, B., O'Connell, J.P., Welch, V., Losos, M., Tugwell, P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses Ottawa: Ottawa Hospital; 2014 [Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp accessed 05-31 2016.

20. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine* 2015;8(1):19-20. doi: 10.1017/S1481803500013336

21. Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating: New York : Springer, [2009] ©2009 2009.

22. Carrasco MP, Villarroel L, Andrade M, et al. Development and validation of a delirium predictive score in older people. *Age and ageing* 2014;43(3):346-51. doi: 10.1093/ageing/aft141 [published Online First: 2013/09/26]

23. de Wit HAJM, Winkens B, Mestres Gonzalvo C, et al. The development of an automated ward independent delirium risk prediction model. *Int J Clin Pharm* 2016 doi: 10.1007/s11096-016-0312-7

24. Douglas VC, Hessler CS, Dhaliwal G, et al. The AWOL tool: derivation and validation of a delirium prediction rule. *J Hosp Med* 2013;8(9):493-9. doi: 10.1002/jhm.2062 [published Online First: 2013/08/08]

25. Fisher BW, Flowerdew G. A simple model for predicting postoperative delirium in older patients undergoing elective orthopedic surgery. *Journal of the American Geriatrics Society* 1995;43(2):175-8. [published Online First: 1995/02/01]

26. 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(4):212-16. doi: 10.5770/cgj.18.185

27. Freter SH, Dunbar MJ, MacLeod H, et al. Predicting post-operative delirium in elective orthopaedic patients: the Delirium Elderly At-Risk (DEAR) instrument. *Age and ageing* 2005;34(2):169-71.

28. Inouye SK, Charpentier PA. Precipitating factors for delirium in hospitalized elderly persons. Predictive model and interrelationship with baseline vulnerability. *Jama* 1996;275(11):852-7. [published Online First: 1996/03/20]

29. Inouye SK, Viscoli CM, Horwitz RI, et al. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. *Annals of internal medicine* 1993;119(6):474-81. [published Online First: 1993/09/15]

30. Inouye SK, Zhang Y, Jones RN, et al. Risk factors for delirium at discharge: development and validation of a predictive model. *Archives of internal medicine* 2007;167(13):1406-13. doi: 10.1001/archinte.167.13.1406 [published Online First: 2007/07/11]

31. Isfandiaty R, Harimurti K, Setiati S, et al. Incidence and predictors for delirium in hospitalized elderly patients: a retrospective cohort study. *Acta medica Indonesiana* 2012;44(4):290-7. [published Online First: 2013/01/15]

32. Pendlebury ST, Lovett N, Smith SC, et al. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of externally derived risk scores. *Age and ageing* 2016;45(1):60-65. doi: 10.1093/ageing/afv177

33. 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. *Journal of the American Geriatrics Society* 2006;54(5):817-22. doi: 10.1111/j.1532-5415.2006.00704.x [published Online First: 2006/05/16]

34. Kim MY, Park UJ, Kim HT, et al. DELirium Prediction Based on Hospital Information (Delphi) in General Surgery Patients. *Medicine (Baltimore)* 2016;95(12):e3072. doi: 10.1097/MD.0000000000003072

35. Korc-Grodzicki B, Sun SW, Zhou Q, et al. Geriatric Assessment as a Predictor of Delirium and Other Outcomes in Elderly Patients With Cancer. *Annals of surgery* 2014 doi: 10.1097/sla.0000000000000742 [published Online First: 2014/06/03]

36. 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? *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* 2013;21(10):946-56. doi: 10.1016/j.jagp.2013.01.069 [published Online First: 2013/05/11]

37. Liang C-K, Chu C-L, Chou M-Y, 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(4):347-55. doi: 10.1089/rej.2014.1645

38. 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* 2015 doi: 10.1111/ggi.12587

39. 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(5) doi: 10.1136/bmjopen-2012-001599 [published Online First: 2012/09/18]

40. 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(2):153-9. doi: 10.1016/j.genhosppsych.2011.11.011 [published Online First: 2012/01/10]

41. O'Keeffe ST, Lavan JN. Predicting delirium in elderly patients: development and validation of a risk-stratification model. *Age and ageing* 1996;25(4):317-21. [published Online First: 1996/07/01]

42. Pompei P, Foreman M, Rudberg MA, et al. Delirium in hospitalized older persons: outcomes and predictors. *Journal of the American Geriatrics Society* 1994;42(8):809-15. [published Online First: 1994/08/01]

43. Rudolph JL, Doherty K, Kelly B, et al. Validation of a Delirium Risk Assessment Using Electronic Medical Record Information. *Journal of the American Medical Directors Association* 2016;17(3):244-48 5p. doi: 10.1016/j.jamda.2015.10.020

44. Rudolph JL, Harrington MB, Lucatorto MA, et al. Validation of a medical record-based delirium risk assessment. *Journal of the American Geriatrics Society* 2011;59 Suppl 2:S289-94. doi: 10.1111/j.1532-5415.2011.03677.x

45. Rudolph JL, Jones RN, Levkoff SE, et al. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. *Circulation* 2009;119(2):229-36 8p. doi: 10.1161/CIRCULATIONAHA.108.795260

46. Dworkin A, Lee DS, An AR, et al. A Simple Tool to Predict Development of Delirium After Elective Surgery. *Journal of the American Geriatrics Society* 2016;64(11):e149-e53. doi: 10.1111/jgs.14428 [published Online First: 2016/09/22]

47. Freter SH, George J, Dunbar MJ, et al. Prediction of delirium in fractured neck of femur as part of routine preoperative nursing care. *Age and ageing* 2005;34(4):387-8. doi: 10.1093/ageing/afi099 [published Online First: 2005/06/16]

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 and ageing* 2016 doi: 10.1093/ageing/afw198 [published Online First: 2016/11/07]

49. 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(12):1373-9. [published Online First: 1996/12/01]

50. 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*;0(0):0962280214558972. doi: doi:10.1177/0962280214558972

51. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ (Clinical research ed)* 2015;351:h3868. doi: 10.1136/bmj.h3868 [published Online First: 2015/08/13]

52. Grobman WA, Stamilio DM. Methods of clinical prediction. *American journal of obstetrics and gynecology* 2006;194(3):888-94. doi: 10.1016/j.ajog.2005.09.002 [published Online First: 2006/03/09]

53. Jones RN, Marcantonio ER, Saczynski JS, et al. Preoperative Cognitive Performance Dominates Risk for Delirium Among Older Adults. *J Geriatr Psychiatry Neurol* 2016 doi: 10.1177/0891988716666380 [published Online First: 2016/09/21]

54. Kazmierski J, Banys A, Latek J, et al. Mild Cognitive Impairment with Associated Inflammatory and Cortisol Alterations as Independent Risk Factor for Postoperative Delirium. *Dementia & Geriatric Cognitive Disorders* 2014;38(1/2):65-78 14p. doi: 10.1159/000357454

55. 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 Clinical And Experimental Research* 2007;19(3):172-77.

56. Fong TG, Hshieh TT, Wong B, et al. Neuropsychological profiles of an elderly cohort undergoing elective surgery and the relationship between cognitive performance and delirium. *Journal of the American Geriatrics Society* 2015;63(5):977-82. doi: 10.1111/jgs.13383

57. Rozzini R, Sabatini T, Cassinadri A, et al. Relationship between functional loss before hospital admission and mortality in elderly persons with medical illness. *J Gerontol A Biol Sci Med Sci* 2005;60(9):1180-3. [published Online First: 2005/09/27]

58. Fimognari FL, Pierantozzi A, De Alfieri W, et al. The Severity of Acute Illness and Functional Trajectories in Hospitalized Older Medical Patients. *J Gerontol A Biol Sci Med Sci* 2016 doi: 10.1093/gerona/glw096 [published Online First: 2016/06/04]

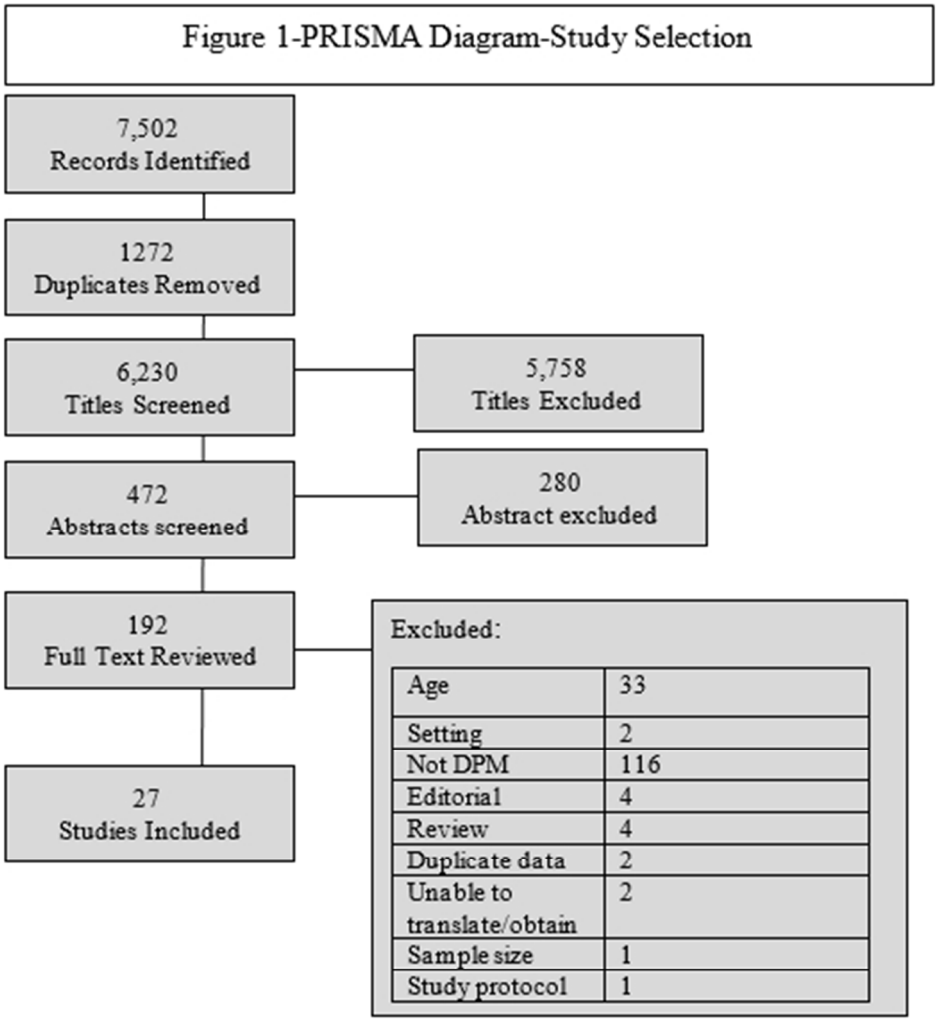
59. Li M, Harris I, Lu ZK. Differences in proxy-reported and patient-reported outcomes: assessing health and functional status among medicare beneficiaries. *BMC Medical Research Methodology* 2015;15(1):62. doi: 10.1186/s12874-015-0053-7

60. van Meenen LC, van Meenen DM, de Rooij SE, et al. Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *Journal of the American Geriatrics Society* 2014;62(12):2383-90. doi: 10.1111/jgs.13138

10.1111/jgs.13138.</p>

61. Newman MW, O'Dwyer LC, Rosenthal L. Predicting delirium: A review of risk-stratification models. *General Hospital Psychiatry* 2015;37(5):408-13. doi: 10.1016/j.genhosppsych.2015.05.003

62. Campaner CP. Valoración del riesgo de delirium en pacientes mayores hospitalizados. = Delirium risk assessment in elderly hospitalized patients. *Revista Española de Geriatría y Gerontología* 2010;45(5):285-90. doi: 10.1016/j.regg.2010.03.011



PRISMA Diagram

128x144mm (96 x 96 DPI)

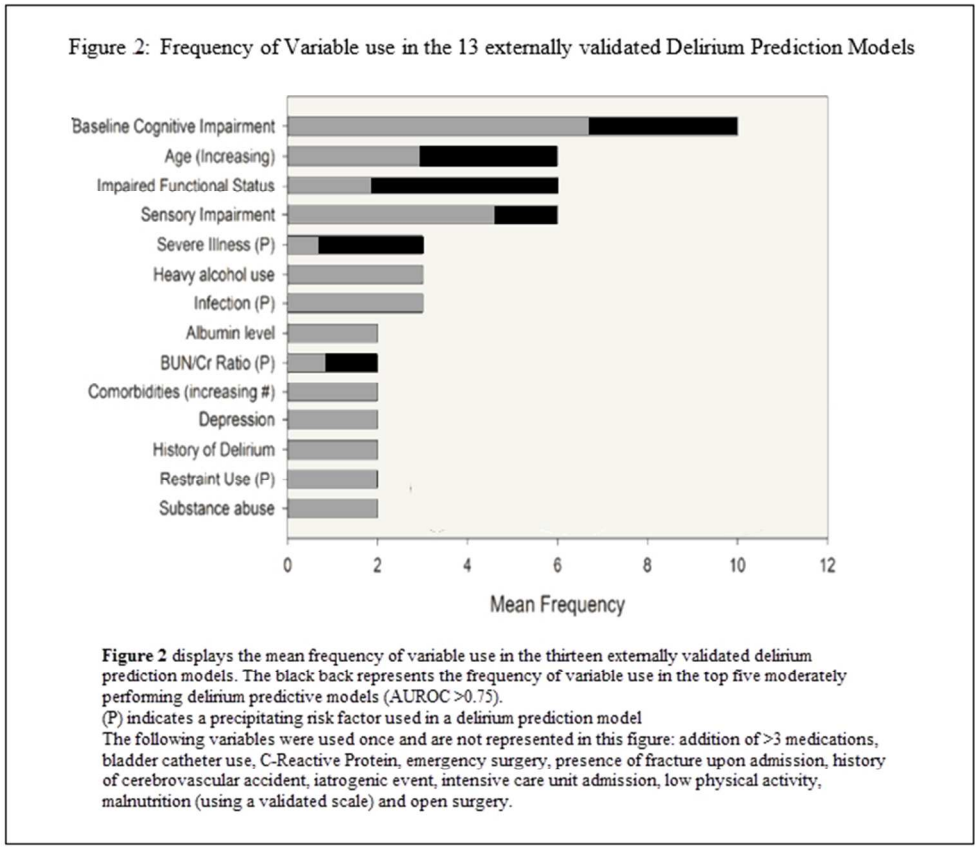


Figure 2 displays the mean frequency of variable use in the thirteen externally validated delirium prediction models. The black bar represents the frequency of variable use in the top five moderately performing models (AUROC>0.75). ‼ ┼ (P) indicates a precipitating risk factor used in DPMs.‼ ┼ The following variables were used once and are not represented in this figure: addition of >3 medications, bladder catheter use, C-Reactive Protein, emergency surgery, presence of fracture upon admission, history of cerebrovascular accident, iatrogenic event, intensive care unit admission, low physical activity, malnutrition(using a validated scale), and open surgery.

190x164mm (96 x 96 DPI)

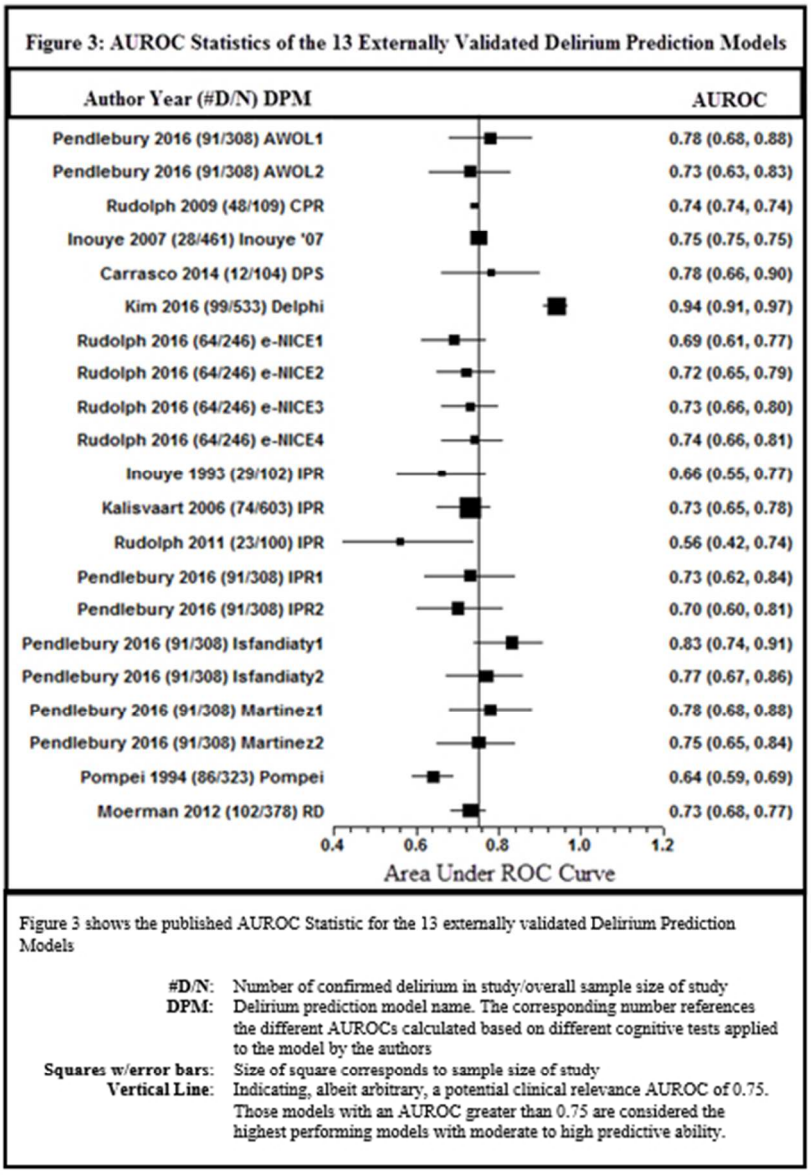


Figure 3 shows the published AUROC Statistic for the 13 externally validated Delirium Prediction Models

#D/N: Number of confirmed delirium in study/overall sample size of study
DPM: Delirium prediction model name. The corresponding number references the different AUROCs calculated based on different cognitive tests applied to the model by the authors
Squares w/error bars: Size of square corresponds to sample size of study
Vertical Line: Indicating, albeit arbitrary, a potential clinical relevance AUROC of 0.75. Those models with an AUROC greater than 0.75 are considered the highest performing models with moderate to high predictive ability.

122x175mm (96 x 96 DPI)

Appendix A – Review Protocol

Working Title of Review	Systematic Review of Delirium Prediction Models		Support	Modifications
Authors	1st & Corresponding	Heidi Lindroth	Literature search, data extraction, data synthesis and manuscript preparation.	
	Data Extraction	Heidi Lindroth Suzanne Purvis	Literature search, data extraction, data synthesis.	
	Content Experts	Lisa Bratzke	Assisted with content related to cognition. Results review.	
		Roger Brown	Statistical content expert	
		Mark Coburn	Results review, Manuscript preparation	
		Marko Mrkobrada	Results review, Manuscript preparation	
		Matthew TV Chan	Results review, Manuscript preparation	
		Daniel Davis	Geriatrician expertise, reviewed results, manuscript preparation.	
		Pratik Pandharipande	Results review, Manuscript preparation	
		Cynthia M. Carlsson	Geriatrician expertise, reviewed results, manuscript preparation.	
	Mentoring	Robert D. Sanders	Mentoring author, resolved content/data disagreements b/w authors, manuscript preparation.	
Aim	To identify existing prognostic delirium prediction models and evaluate their validity and statistical methodology in the older adult (≥60yo) acute hospital population.			
Search Terms	("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*)		UW-Madison Health Sciences librarian. Three meetings to refine search terms.	
Databases searched	PubMed, CINAHL, PsychINFO, Cochrane, SocINDEX and Medline		Health Sciences librarian.	Expanded to include SocINDEX
Timelines established	01/01/1990-12/31/2016			Originally was 12/31/15. Expanded to

			include all of 2016.
Inclusion criteria	• Age ≥ 60 • Inpatient population • Developing and/or validating a delirium prediction model		Age expanded from ≥ 70 years of age due to the literature
Exclusion criteria	• Emergency department • Hospice/palliative care • Pediatric population • Related to alcohol withdrawal • ≤50 sample size	Mentoring author	Sample size criteria added to build rigor in the studies that were included in the sys review
Selection process	Studies will be selected based on the inclusion/exclusion criteria. The data extraction authors (HL and SP) will conduct the literature search independently and meet monthly to discuss findings. Any disagreements will be resolved by the mentoring author (RDS)		
Data Management	A shared folder on the UW-Madison Box account will be created to share documents, data and meeting information.		
Data collection process	Data will be collected independently by HL and SP then data points will be shared at monthly meetings. Data collection tables will be created using Microsoft Excel then uploaded to the shared Box account. Any disagreement between authors will be resolved by the mentoring author (RDS).		
Data points collected	• Characteristics of studies (design, population, sample size) • Outcome measure including how it was identified, measured, defined. Prevalence. • Statistical methods applied • Statistical information about the delirium prediction models (sensitivity, specificity, positive predictive value, negative predictive value, AUROC) • Characteristics of DPMs (variables used, scoring, development) • Cognitive measures used in studies. • Criteria to fulfill the Newcastle Ottawa Scale.		
Outcomes	• AUROC will be the primary outcome measure • Characteristics of DPMs (variables, statistics) • Cognitive tests used		
Data synthesis	The first/corresponding author (HL) will synthesize the data into the		

	manuscript. The co-authors will verify this. RB will complete the meta-analysis.		
Manuscript preparation	HL will complete manuscript preparation. All co-authors are responsible for reviewing content and data to assure correctness and complete synthesis of data gathered.		

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	4
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4-5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5-6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	6
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	7-8Table1
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Table 1,2 Figure 3
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Table 1
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8-16
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	16-18
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	18-19
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16-21
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	6

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

PRISMA 2009 Checklist

BMJ Open

Systematic Review of Prediction Models for Delirium in the Older Adult Inpatient

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019223.R1
Article Type:	Research
Date Submitted by the Author:	22-Jan-2018
Complete List of Authors:	Lindroth, Heidi; University of Wisconsin Madison School of Medicine and Public Health, Department of Anesthesiology; University of Wisconsin Madison Graduate School, School of Nursing Bratzke, Lisa; University of Wisconsin-Madison, School of Nursing Purvis, Suzanne; University of Wisconsin Hospital and Clinics, Nursing Brown, Roger; University of Wisconsin Madison, School of Nursing Coburn, Mark; Uniklinik RWTH Aachen, Anaesthesiology Mrkobrada, Marko; Western University, Medicine Chan, MTV; The Chinese University of Hong Kong, Anesthesia and Intensive Care Davis, Daniel; University College London, MRC Unit for Lifelong Health and Ageing ; University College London Pandharipande, Pratik; Vanderbilt University School of Medicine Carlsson, Cynthia; William S. Middleton Veteran's Administration Geriatric Research Education and Clinical Center (GRECC), GRECC; University of Wisconsin Madison School of Medicine and Public Health Sanders, Robert; University of Wisconsin Madison School of Medicine and Public Health, Department of Anesthesiology
Primary Subject Heading:	Geriatric medicine
Secondary Subject Heading:	Neurology
Keywords:	Delirium, GERIATRIC MEDICINE, Statistic

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 TV Chan, Daniel Davis, Pratik Pandharipande, Cynthia M. Carlsson, Robert D. Sanders
University of Wisconsin School of Medicine and Public Health, Department of Anesthesiology, School of Nursing, 600 Highland Avenue, Madison, WI 53792 Heidi Lindroth BSN PhD Candidate[1]

University of Wisconsin School of Nursing, 701 Highland Avenue, Madison, WI 53705 Lisa Bratzke PhD Assistant Professor[2]

University Hospital, 600 Highland Avenue, Madison, WI 53792 Suzanne Purvis DNP, Clincial Nurse Specialist[3]

University of Wisconsin School of Nursing, 701 Highland Avenue, Madison, WI 53705 Roger L. Brown PhD, Professor[4]

University Hospital RWTH Aachen, Department of Anesthesiology, Pauwelsstrasse 30, 52074 Aachen, Mark Coburn MD, Professor[5]

Western University, Schulich School of Medicine and Dentistry, 339 Windermere Road, Room B9-100, London ON N5X 3W6, Marko Mrkobrada MD Associate Professor[6]

The Chinese University of Hong Kong, Department of Anaesthesia and Intensive Care, 4/F Main Clinical Bock and Trauma Center, Prince of Wales Hospital, Hong Kong Special Administrative Region, China, Matthew TV Chan, MB, BS, PhD, Professor[7]

MRC Unit for Lifelong Health and Ageing at UCL, 33 Bedford Place, London, WC1B 5JU, Daniel Davis PhD MRCP, Senior Clinical Researcher[8]

Vanderbilt University Medical Center, Department of Anesthesiology, 1211 Medical Center Drive, Nashville TN 37322, Pratik Pandharipande MD, Chief, Division of Anesthesiology Critical Care Medicine[9]

University of Wisconsin School of Medicine and Public Health, Madison VA Geriatric Research Education and Clinical Center (GRECC) (D4211), 2500 Overlook Terrace, Madison, WI, 53705, Cynthia M. Carlsson, MD, MS, Associate Professor of Medicine, Division of Geriatrics[10]

University of Wisconsin School of Medicine and Public Health, Department of Anesthesiology, 600 Highland Avenue, Madison, WI 53792, Robert D. Sanders MBBS PhD FRCA, Assistant Professor[11]

[+]**Corresponding author**: *Heidi Lindroth, Department of Anesthesiology, University of Wisconsin School of Medicine and Public Health, 600 Highland Avenue, B6/319 CSC Madison, WI 53792-3272 Telephone: 608-262-6469 Fax: 608-263-0575 Madison, USA. Email:* hlindroth@wisc.edu

Word count: 4043

Objective

To identify existing prognostic delirium prediction models and evaluate their validity and statistical methodology in the older adult (≥60yo) 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 PRISMA and CHARMS Statement guided protocol development. Inclusion criteria: Age ≥60, 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 (<60yrs) and 27 met the defined criteria. Twenty-three delirium prediction models were identified, fourteen were externally validated and three 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 AUROC range from 0.52-0.94. 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.

Funding

Department of Anesthesiology, University of Wisconsin-Madison.

1K23AG055700-01A1

R01 NHLBI (HL111111) and research grant from Hospira Inc., NIH collaboration

All authors completed ICMJE disclosure forms and no conflicts of interest declared.

Keywords

Delirium. Aging. Cognition. Prediction. Statistical Models.

Strengths and Limitations of this Study

- This study used the PRISMA 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, older adult inpatients, and could be expanded to include all ages and settings including palliative care, long term care and the emergency room.

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] Delirium has been independently associated with increased mortality, morbidity in terms of impaired cognition and functional disability along with an estimated annual U.S. 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 formalize 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 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 PRISMA Statement and the CHARMS checklist (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 predetermined modifiable and non-modifiable risk factors of delirium present. This review included studies focused on 1) older adult (≥ 60 years) population, (the U.S. Center 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 to 31 December 2016, and 4) developed and/or validated delirium prediction models. Studies were excluded if they 1) studied a different patient population (i.e. emergency department, skilled nursing facilities, palliative care, and hospice) as these are unique patient populations with characteristics requiring specific foci and are not readily generalizable to a medical or surgical inpatient hospital setting. Further, 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 ≤ 50 for methodological reasons (i.e. underpowered). All study designs were included. Studies were not limited by timeframe 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 CHARMS checklist.[15] Two authors (HL, 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, 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 4) characteristics of the models (variables used in model, 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, odds ratios, 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 Chi-square (X^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 pre-planned outcome measures included cognitive assessments, and predictive variable use per model.

Role of the Funding Source

The funding sources named has 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.

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).[25 40]

Twenty-three delirium prediction models were developed, fourteen were externally validated [23 27 29-31 33-35 41 43-46] and three were internally validated.[24 37 42] Prospective cohort design was used in 23 studies.[23 25-31 33-35 37-49] Retrospective design was 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 randomized control trial.[34 41] Eleven studies focused on the medical population [23 25 29-33 40 42 45 49], three included medical and surgical [24 43 44] and thirteen recruited a surgical population (seven-orthopaedic [26-28 34 38 41 48], one-cardiac [46], two-noncardiac [37 47], one general surgery[35], two-oncological[36 39]). None of the identified studies focused on critical care patients. Data collection occurred upon admission in seventeen studies [23 25 27 29-31 33-35 40-45 48 49]; participants were approached within forty-eight hours of admission. Seven studies collected data pre-operatively then followed participants post-operatively.[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; five 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.

Table 1

Author	Study Design Population Sample Size Sampling method Power Analysis	Study Grade (NOS)	Outcome Variable & Rate (%)	Delirium measurement & frequency	DPM Name & Regression Model used
Carrasco et al. (2014)[23]	P.Cohort Medical Dev: 374 Val: 104 Consecutive	S: **** C: - O: ** T: 6 stars	Delirium Dev: 25 (.06) Val: 12 (12)	CAM Every 48 h	Predictive Risk Score Forward stepwise
de Wit et al. (2016)[24]	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.** (2013)[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)	CAM-S & CAM Daily	Risk Stratification model (AWOL) Forward stepwise
Dworkin et al. (2016)[47]	P.Cohort Elective noncardiac surg Dev: 76 Consecutive	S: **** C: - O: ** T: 6 stars	Delirium Dev: 10(13)	CAM or FAM-CAM 1xafter surgery	Mini-Cog Stratified into a five-point score Stepwise
Fisher and Flowerdew (1995)[26]	P.Cohort Elective Orthopedic Dev: 80 Consecutive	S: ** C: - O: ** T: 4 stars	Delirium Dev: 14(17.5)	CAM 2xDaily	Prediction Model using two variables. Stewpsie
Freter et al. (2005)[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. (2005)[48]	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 al. (2015)[27]	P.Cohort Hip Fracture Val: 283 Consecutive	S: *** C: - O: ** T: 5 stars	Delirium Val: 119(42)	CAM POD1, 3 & 5	Risk stratification model (DEAR)
Inouye and Charpentier (1996)[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
Inouye et al. (2007)[31]	P.Cohort Medical Dev: 491 Val: 461 Consecutive	S: **** C: ** O: *** T: 9 stars	Delirium/ subsyndrome delirium at discharge Dev: 58(12) Val: 28(6)	CAM Every other day	Risk stratification model Log-binomial regression
Inouye et la. (1993)[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
Isfandiaty et al. (2012)[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 et al. (2006)[34]	P.Cohort Hip Surgery & Facture Val: 603 Consecutive	S: *** C: - O: *** T: 6 stars	Delirium Dev: 74(12)	CAM, DRS-98 Daily through POD5	Externally validated Inouye's '93 model.
Kim et al. (2016)[35]	P.Cohort Major General	S: *** C: **	Delirium Dev: 112(20)	Nu-Desc -every shift by RNs	Risk stratification model Backwards stepwise

	Surgery Dev: 561 Val: 533 Not stated Power analysis	O: *** T: 8 stars	Val: 99(18)	Confirmed with CAM.	
Korc-Grodzicki et al. (2014)[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 et al. (2013)[37]	P.Cohort Noncardiac surgery Dev: 581 Not stated	S: *** C: - O: ** T: 5 stars	Delirium Dev: 234(40)	CAM Daily	Risk stratification model Stepwise
Liang et al. (2015)[38]	P.Cohort Elective Orthopedic Surgery Dev: 461 Consecutive	S: *** C: ** O: ** T: 7 stars	Delirium Dev: 37(8)	CAM Daily Confirmed by psychologist DSM-IV	Built 2 DPMs CGA Risk stratification models Backward stepwise
Maekawa et al. (2015)[39]	P.Cohort Oncological; Gastrointestinal Surgery Dev: 517 Consecutive	S: ** C: * O: *** T: 6 stars	Delirium Dev: 124(24)	CAM Unknown frequency	Comprehensive Geriatric Assessment (CGA) as model. Proportional hazards
Martinez et al.(2012)[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 et al. (2012)[41]	P.Cohort Hip Fracture Val: 378 Consecutive Power analysis	S: *** C: - O: *** T: 6 stars	Delirium Val: 102(27)	Ward RN observation, 3xdaily Confirmed by chart review.	Risk stratification model (Risk Model for Delirium, RD) Built from literature
O'Keeffe and Lavan (1996)[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 et al. (2016)[49]	P. Cohort Medical Dev: 308 Consecutive	S: **** C: * O: *** T: 8 stars	Delirium Val: 95(31)	CAM Every 48-hours Confirmed by DSM-IV interview	Susceptibility Score Built from literature
Pendlebury et al. (2016)[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 4 DPMs
Pompei et al. (1994)[43]	P.Cohort Med/surg Dev: 432 V: 323 Not stated	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 64(14.8) Val: 86(26.3)	CAM 2xweekly. Confirmed with DSM III	Risk stratification model Stepwise
Rudolph et al. (2009)[46]	P.Cohort Cardiac Surgery Dev: 122 V: 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 et al. (2011)[45]	P.Cohort Medical V: 100 Consecutive	S: **** C: - O: *** T: 7 stars	Delirium Dev: 23(23)	DSM-IV Daily clinical interview	Externally validated Inouye's '93 model.
Rudolph et al. (2016)[44]	Dev: Retro Val: P.Cohort Med/surg Dev: 27625 Val: 246	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

	Consecutive				

Key:
**=Models developed in population ≤60 years of age, but validated in population ≥60 years of age.
Study Design: P.Cohort=Prospective Cohort, Retro=Retrospective design. Dev=Development, Val=Validation. Med=Medical, Surg=Surgical. Power analysis = reported in identified study.
Study Grade: NOS=Newcastle Ottawa Scale. S=Selection, C=Comparability, O=Ottawa. Max 9 stars.
Outcome Variable: Dev=Development, Val=Validation
Delirium Measurement: CAM=Confusion Assessment Method, DSM=Diagnostic Statistical Manual, POD=Postoperative Day, MDAS=Memorial Delirium Assessment Scale, Nu-Desc=Nursing Delirium Screening Scale, DRS-98=Delirium Rating Scale, EHR=Electronic Health Record
Type of Model: How authors designed their delirium prediction model (DPM), statistical method used
-Risk stratification model: Points (weighted or un-weighted) assigned per predictive risk factor present.
-CGA=Comprehensive Geriatric Assessment
-Built from Literature: Authors selected risk factors for DPM based on literature review.

Delirium assessment

The outcome variable was measured using the Confusion Assessment Method in twenty-one studies.[23 25-31 33-40 43 46-49] The frequency of delirium assessment varied from two or more assessments daily (three studies)[26 35 41], to once daily (twelve studies)[25 28 30 32 34 36-38 44-46 48], every-other day (eight studies)[23 27 29 31 33 42 43 49], once following surgery[47], and undefined (three 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 43-47 49] 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 recognize and chart delirium symptoms.[28 41 48] 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]

Model design and statistical methods

Various statistical techniques were employed by the twenty-three included studies. Twelve used univariate or bivariate analyses and selected variables with a pre-determined statistical value

(range for 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] Two used proportional hazards regression modeling paired with bivariate analyses and included variables with either a *p*-value <0.25[32] or a relative risk of ≥1.5.[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 categorized 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 eight of the fourteen externally validated models are categorized 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 fourteen 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 49] This cognitive test was administered upon study enrollment or extracted from past medical records.[49] 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 upon enrollment 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 mBDRS[31], and one study-dementia billing codes or prescription information.[44] One study defined baseline cognitive impairment as a pre-specified 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) needing assistance with any basic ADL,[27] (1) domestic help, help with meals or physical care[41] and (2) residence in nursing facility or at home with caregivers.[33] Two studies used validated functional assessment tools (iADL and Barthel Index) and evaluated functional status two weeks prior to hospitalization.[23 31]

Externally validated delirium prediction models are detailed in Table 2.

Table 2

External Validated DPM Name	Citation Type of Validation	Delirium #(%)	Sens Spec PPV NPV (external)	AUROC (95%CI)	Model Components	Cog. Assess Tool & Cutoff
AWOL Tool	Pendlebury et al. (2016) [33] Broad val.	1st Val: 14(9) 2nd Val: 95(31) (any delirium) 67-prevalent 28-incident	Mod. AWOL Cutoff - 3 Any Delirium Sens ·7 Spec .66 PPV .55 NPV .79 Incident Del Sens .76 Spec .66 PPV .27 NPV .94	1st Val: 0.69 (0·54-0·83) Incident delirium 2nd Val: Cohort 1 (MMSE) 0·78 (0·68-0·88) Cohort 2 (AMTS) 0·73 (0·63-0·83)	Original AWOL Tool Age >80 — 1 pt Failure to spell WORLD backwards — 1 pt Disorientation — 1 pt Illness Severity — 1 pt Modified AWOL Tool Age >80 — 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 et al. (2009) [46] 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 pt 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	MMSE -Stratified score

					2 pts ≥ 3 pts – High risk group RR in High risk group: 4.9 (3.8-6.2)	
DEAR	Freter et al. (2015)[27] Narrow val.	Dev: (2005) 18(14) Val: (2015) Pre-Op= 163(58) Post-op= 118(42)	Sens .68 Spec .73 PPV .65 NPV .76 Optimal cut-off score: 3pts (Incident post-op delirium)	Dev: (2005) 0·77 (0·64-0·87) Val: (2015) AUROC Not published	MMSE ≤ 23 — 1 pt Functional dependence — 1 pt Sensory impairment — 1 pt Substance use — 1 pt Age >80 — 1 pt Not weighted. 0-5 Score, cut-off of 3 indicating high risk.	MMSE Cut-off ≤ 23
Delirium at Discharge Prediction Model	Inouye et al. (2007)[31] Narrow val.	Dev: 58(12) Val: 28(6) (incident delirium)	Not reported	Dev: 0.80 Val: 0.75 Did not report CI Calibration: X^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 pt = 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 et al. (2014)[23] Narrow val.	Dev: 25(.06) Val: 12(12) (incident delirium)	Sens .88 Spec .74 PPV .22 NPV .99	Dev: 0.86 (0.82-0.91) Val: 0.78 (0.66-0.90)	DPS=[5xBUN/Cr ratio]-(3xBarthel 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
Delphi Score	Kim et al. (2016)[35] Narrow val.	Dev: 112(20) Val: 99(18) (incident delirium)	Sens .81 Spec .93 PPV .70 NPV .96 Optimal cut-off score: 6.5pts	Dev: 0.911 (0.88-0.94) Val: 0.938 (0.91-0.97)	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 Pre-Op CRP (mg/dL)	No measure of cognition. Excluded participants if MMSE <24

					<10 0 ≥10 1 Max points: 15 Optimal cut-off: 6.5 High Risk: ≥7 points	
e-NICE Rule	**Rudolph et al. (2016)[44]** Broad val. Dev: 2343(8) Val:64(26) (incident delirium)	Cohort / AUROC / CI / TPR / FPR Dev 0.81 (0.80-0.82) Validation AUROCs* Original 0.69 (0.61-0.77) 64% 33% mRASS 0.72 (0.65-0.79) 69% 35% TMYB 0.73 (0.66-0.80) 78% 43% MoCA 0.74 (0.66-0.81) 75% 43% *Any delirium Original model-AUROC of 0.68 (95%CI0.59-0.77) in incident delirium. Did not report sens, spec, PPV, NPV			Weighted Points/OR Cog impair -Medications, diagnosis or both: 4 pt Age ≥ 65 y: 2 pt Age ≥ 80 y: 3 pt Infection: 2 pt Fracture: 4 pt Vision: 1 pt Severe Illness: 2 pt 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 et al. (1993)[30] Narrow val.	Dev: 27(25) Val: 29(17) (incident delirium)	Did not report	Dev: 0.74 (0.63-0.85) Val: 0.66 (0.55-0.77) Calibration: Dev: X^2 Trend $p<0{\cdot}00001$ Val: X^2 Trend $p<0{\cdot}002$	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 pts = 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/care giver bDRS Excluded those w/history of severe dementia
IPR	Kalisvaart et al. (2006)[34] Broad val.	Val: 74(12)	Did not report	Val: 0.73 (0.65-0.78) Calibration: X^2 $p<0{\cdot}05$ X^2 Trend $p<0{\cdot}002$	Externally validated IPR in surgical hip fracture population. -Addition of age & type of admission improved model performance, R^2=0.20 RR of High risk group: 9.8	MMSE Cut-off < 24
IPR	Rudolph et al. (2011)[45] Broad val.	Val: 23(23) Any delirium 10-Prevalent 13-Incident	Did not report	Val: 0.56 (0.42-0.74) Incident del. 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
IPR	Pendlebury et al. (2016)[33] Broad val.	Val: 95(31) Any delirium 67-prevalent 28-incident	Cutoff 2pts All Delirium Sens .57 Spec .80 PPV .64 NPV .76	Val: Incident delirium Cohort 1 (MMSE) 0.73	Baseline cognitive impairment: 1 pt High BUN/Cr ratio: 1 pt Severe illness (SIRS ≥ 2): 1 pt Vision impairment: 1 pt	Original Model: MMSE <24 Modified Model:

			Incident D Sens .52 Spec .80 PPV .31 NPV .91	(0.62-0.84) Cohort 2-(AMTS) 0.70 (0.60-0.81)	4pts=Incident delirium	MMSE < 24 AMTS < 9
Isfandiaty model	Pendlebury et al. (2016) [33] Broad val.	Dev: 87(19) Val: 95 (31) Any delirium 67-prevalent 28-incident	Cutoff 4pts Any Delirium Sens .74 Spec .71 PPV .60 NPV .82 Incident Del Sens .81 Spec .71 PPV .31 NPV .96	Dev: 0.82 (0.77-0.88) Val: Incident delirium Cohort 1 (MMSE) 0.83 (0.74-0.91) Cohort 2 (AMTS) 0.77 (0.67-0.86)	Baseline cognitive impairment: 3 pt Functional dependency: 2 pt Infection w/sepsis: 2 pt Infection w/out sepsis: 1 pt Weighted Score Score = 7 for incident delirium Cohort 1: MMSE Cohort 2: AMTS	<u>Original Model:</u> Chart review <u>Modified Model:</u> MMSE < 24 AMTS < 9
Martinez et al. 2012 model	Pendlebury et al. (2016) [33] Broad val.	1st Val: 76(25) 2nd Val: 95(31) Any delirium 67-prevalent 28-incident	Modified Model Cutoff 2pts Any Delirium Sens .62 Spec .68 PPV .54 NPV .75 Incident Del Sens .81 Spec .68 PPV .29 NPV .96	1st Val: 0.85 (0.80-0.88) Incident delirium 2nd Val: Cohort 1 (MMSE) 0.78 (0.68-0.88) Cohort 2 (AMTS) 0.75 (0.65-0.84)	Martinez et al. 2012 Original Model Age >85: 1 pt Dependent in ≥5 ADLs: 1 pt Drugs on admit: -Antidepressants -Antidementia -anticonvulsants -antipsychotics: 1pt/drug 2pt/ antipsych Score 0-3 Score >1 = high risk for delirium Modified Model Age >85: 1 pt Dependency in ≥ 5 ADLs: 1 pt Diag of Dementia MMSE<24 AMTS<9: 1 pt	<u>Original Model:</u> -No cognitive measure <u>Modified Model:</u> MMSE < 24 AMTS < 9
Pompei et al. 1994 model	Pompei et al. (1994)[43] Broad val.	Dev: 64(15) Val: 86(26) (21=prevalent delirium)	Sens .83 Spec .50 PPV .38 NPV .89 *Pts stratified as low or moderate to high-risk	Dev: 0.74 +/- 0.05 Val: 0.64 +/- 0.05 Calibration: X^2 Trend $p<0{\cdot}0001$	Weighted Points Baseline cognitive impairment: 2 pt Depression: 2 pt Alcoholism: 3 pt ≥ 4 comorbidities: 3 pt 0-3 pts = Low risk 4-7 pts = Moderate risk 8-10 pts = High risk	MMSE Less than HS <21 High school <23 College edu < 24
Precipitating Risk Factors	Inouye and Charpentier (1996)[29] Narrow val.	Dev: 35(18) Val: 47(15) (incident delirium)	Not reported	No AUROC reported Calibration: X^2 Trend $p<0{\cdot}001$	Physical restraint use: 1 pt Malnutrition: 1 pt ≥3 medications added: 1 pt Bladder catherization: 1 pt Any iatrogenic event: 1 pt Not weighted. 0 pt = Low Risk 1-2 pt = Intermediate ≥ 3 pt = High Risk RR of High Risk: 17.5 (8.1-37.4)	None used in model

Risk Model for Delirium (RD)	Moerman et al. (2012) [41] Narrow val.	Val: 102(27) (incident delirium)	Sens .81 Spec .56 PPV .41 NPV .89 Optimal cut-off score: 4 pts	Val: 0.73 (0.68-0.77)	Weighted Points Delirium-previous hospitalization: 5 pt Dementia: 5 pt Clock Drawing: -Sm mistake: 1 pt -big mistake: 2 pt Age: -70 to 85 years old: 1 pt - >85 years: 2 pt Impaired hearing: 1 pt Impaired vision: 1 pt Problems w/ADL: -Help w/meal prep: .5p -help w/physical: .5p Use of heroin, methadone, morphine: 2 pt Daily >4 alcohol: 2 pt ≥ 5 pts = High risk	CDT -11:10 -Two Categories 1 Small mistakes 2 Big mistakes
Susceptibility Score	Pendlebury et al. (2016) [49] 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 (0-70-0.92) Improved w/age eliminated to 0.84 (0.77-0.92)	Weighted Points Dementia/cog impair: 2 Age >80 years: 2 Severe illness (SIRS+): 1 Infection-working diagnosis: 1 Vision impairment: 1 >5 pts=High Risk ORs for >5 risk score: 25.0 (3.0-208.9) RR for >5 risk score: 5.4	Known diagnosis of dementia or MMSE < 24 AMTS < 9

Key:
Dev=Development, Val=Validation
Sens=Sensitivity, Spec=Specificity, PPV=Positive Predictive Value, NPV=Negative Predictive Value
Area Under the Receiver Operating Curve Statistic, Dev=Development, Val=Validation, mRASS=Modified Richmond Agitation-Sedation Scale, TMTYB=The Months of the Year Backwards
ADL=Activities of Daily Living
MMSE=Mini Mental Status Exam, AMTS=Abbreviated Mental Test Score, CDT=Clock Drawing Test, mBDR=Modified Blessed Dementia Rating, MoCA=Montreal Cognitive Assessment.

Predictive ability

Reported AUROC in externally validated delirium prediction models ranged from 0.52-0.94 (Figure 4). Of these models, the highest performing model (AUROC 0.94, CI 0.91-0.97) was developed and validated in a surgical population.[35] Two models reported an external validation AUROC above 0.80, indicating moderate predictive ability.[33] [49] 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 fourteen externally validated delirium prediction models reported calibration metrics.[29-31 34 43 45] The reported chi-square 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 reported calibration statistics.[32 37 40 42] None of the included studies reported calibration plots or slopes.

Risk of overfitting

Events per variable (EPV) were examined in each of the fourteen 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 48] 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 penalized 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 odds ratios, relative risk, sensitivity and specificity, receiver operator curves, R squared 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 odds ratios or relative risk statistics evaluating the highest risk stratification cut-off point.[29-31 34 46 49] Seven studies reported sensitivity and specificity[23 27 33 35 41 43 49] and one study reported the rate of true positives and false positives.[44] None of the identified studies reported decision curse 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. (2016) 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 49] Moerman et al. reported feasibility and reliability statistics following the incorporation of the risk prediction tool into practice.[41]

DISCUSSION

This review identified moderate predictive ability (AUROC 0.52-0.94) in fourteen externally validated delirium prediction models with eight out of fourteen 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. (2016) 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.[49] 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 interrelationship between premorbid and precipitating factors,[29] the time course of data collection is important. Nine of the fourteen 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 (1993).[30] However, if variables are collected after the onset of delirium this would exaggerate model performance (e.g. ICU admission). As an example, one delirium prediction model has a robust AUROC of 0.94 (CI 0.91-0.97).[35] This study excluded those with a 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 upon admission and either excluded those with prevalent delirium or calculated separate AUROCs for prevalent delirium versus incident delirium.[23 30 33 44 49]

Model underperformance may be explained by low powered studies, insufficient events per

variable (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. (2016) 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] Further, 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, 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[67] 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. (2016), an improved AUROC resulted when age was removed from the prediction model (0.81 to 0.84).[49] 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] Further, 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. (2016) 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. (2015) 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. (2015), 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 twenty-seven studies with fourteen externally validated delirium prediction models. Our author team is interprofessional, providing the opportunity for different perspectives on model evaluation. Further, this review synthesizes 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. Further, 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. Further 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 generalizability and bolstering broad external validation.[81] Variable definition could be harmonized 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 re-fitting. 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 mild cognitive impairment and sensitive to the spectrum of cognition. Further, 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 Steyerberg (2009)[22] and Steyerberg and Vergouwe (2014)[56] including strategies to

counter low sample size and overly optimistic model performance, the use of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to assess model fit, and consider broad validations to expand case-mix and generalizability; 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 0.52-0.94. Assessment of the outcome variable, delirium, is often non-systematic and future studies would be improved with more standardized 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 generalizable to several populations within the remit of understanding the relationship with the precipitating event.

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 synthesized 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. DD 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.

Declaration of Interests

All authors have completed ICMJE disclosure forms and no conflicts of interest are declared.

Data Sharing Statement

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

Acknowledgements

HL and RDS acknowledge funding support from the Department of Anesthesiology at University of Wisconsin-Madison. PP acknowledges funding support from R01 NHLBI (HL111111) and research grant from Hospira Inc. in collaboration with NIH. 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 and Lily Turner for their assistance with proofreading the manuscript.

Figure Legends

Figure 1: No legend

Figure 2: Figure 2 displays the CHARMS Risk of Bias assessment on all included studies.
Study Participants: design of included study, sampling method, 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,

missing data

Statistical Analysis: Selection of predictors, internal validation, type of external validation

Figure 3: Figure 3 displays the mean frequency of variable use in the fourteen 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, 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.

Figure 4: Figure 4 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 w/error bars: Size of square corresponds to sample size of study

AUROC: Reported Area Under the Receiver Curve Statistic, 95% Confidence Intervals

References

1. Sanders RD, Pandharipande PP, Davidson AJ, et al. Anticipating and managing postoperative delirium and cognitive decline in adults. *BMJ: British Medical Journal (Overseas & Retired Doctors Edition)* 2011;343(7816):197-200 4p. doi: 10.1136/bmj.d4331
2. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet* 2014;383(9920):911-22. doi: 10.1016/s0140-6736(13)60688-110.1016/S0140-6736(13)60688-1. Epub 2013 Aug 28.</p>
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(1) doi: 10.1136/bmjopen-2012-001772
4. Pandharipande PP, Girard TD, Jackson JC, et al. Long-Term Cognitive Impairment after Critical Illness. *The New England Journal Of Medicine* 2013;369(14):1306-16. doi: 10.1056/NEJMoa1301372
5. Leslie DL, Inouye SK. The importance of delirium: economic and societal costs. *Journal of the American Geriatrics Society* 2011;59 Suppl 2:S241-3. doi: 10.1111/j.1532-5415.2011.03671.x [published Online First: 2011/12/07]
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. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* 2016 doi: 10.1016/j.jagp.2016.05.016 [published Online First: 2016/09/14]
7. Pollard C, Fitzgerald M, Ford K. Delirium: The lived experience of older people who are delirious post-orthopaedic surgery. *International Journal of Mental Health Nursing* 2015;24(3):213-21. doi: 10.1111/inm.12132
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 Surgery* 2015;150(12):1134-40. doi: 10.1001/jamasurg.2015.2606
9. Robinson TN, Raeburn CD, Tran ZV, et al. Postoperative delirium in the elderly: risk factors and outcomes. *Annals of surgery* 2009;249(1):173-8. doi: 10.1097/SLA.0b013e31818e4776 [published Online First: 2008/12/25]
10. Adams ST, Leveson SH. Clinical prediction rules. *BMJ (Clinical research ed)* 2012;344
11. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of internal medicine* 2006;144(3):201-09.
12. Moons KGM, 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(9):683-90. doi: 10.1136/heartjnl-2011-301246

13. Debray TPA, Moons KGM, Ahmed I, et al. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in medicine* 2013;32(18):3158-80. doi: 10.1002/sim.5732
14. Hayhurst CJ, Pandharipande PP, Hughes CG. Intensive Care Unit DeliriumA Review of Diagnosis, Prevention, and Treatment. *Anesthesiology* 2016;125(6):1229-41. doi: 10.1097/ALN.0000000000001378
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 medicine* 2014;11(10):e1001744. doi: 10.1371/journal.pmed.1001744 [published Online First: 2014/10/15]
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(5):336-41. doi: 10.1016/j.ijsu.2010.02.007 [published Online First: 2010/02/23]
17. Nations U. Ageing 2016 [Available from: http://www.un.org/en/sections/issues-depth/ageing/index.html accessed March 7th 2017.
18. Centers for Disease C. Elder Abuse: Definitions 2016 [Available from: https://www.cdc.gov/violenceprevention/elderabuse/definitions.html accessed March 7th 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(1):263-306. doi: 10.1097/CCM.0b013e3182783b72 [published Online First: 2012/12/28]
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(11):1047-56. doi: 10.1001/jama.2017.11468 [published Online First: 2017/10/05]
21. Well GA, Shea, B., O'Connell, J.P., Welch, V., Losos, M., Tugwell, P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses Ottawa: Ottawa Hospital; 2014 [Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp accessed 05-31 2016.
22. Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating: New York : Springer, [2009] ©2009 2009.
23. Carrasco MP, Villarroel L, Andrade M, et al. Development and validation of a delirium predictive score in older people. *Age and ageing* 2014;43(3):346-51. doi: 10.1093/ageing/aft141 [published Online First: 2013/09/26]
24. de Wit HAJM, Winkens B, Mestres Gonzalvo C, et al. The development of an automated ward independent delirium risk prediction model. *Int J Clin Pharm* 2016 doi: 10.1007/s11096-016-0312-7
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(9):493-9. doi: 10.1002/jhm.2062 [published Online First: 2013/08/08]
26. Fisher BW, Flowerdew G. A simple model for predicting postoperative delirium in older patients undergoing elective orthopedic surgery. *Journal of the American Geriatrics Society* 1995;43(2):175-8. [published Online First: 1995/02/01]

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(4):212-16. doi: 10.5770/cgj.18.185
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 and ageing* 2005;34(2):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(11):852-7. [published Online First: 1996/03/20]
30. Inouye SK, Viscoli CM, Horwitz RI, et al. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. *Annals of internal medicine* 1993;119(6):474-81. [published Online First: 1993/09/15]
31. Inouye SK, Zhang Y, Jones RN, et al. Risk factors for delirium at discharge: development and validation of a predictive model. *Archives of internal medicine* 2007;167(13):1406-13. doi: 10.1001/archinte.167.13.1406 [published Online First: 2007/07/11]
32. Isfandiaty R, Harimurti K, Setiati S, et al. Incidence and predictors for delirium in hospitalized elderly patients: a retrospective cohort study. *Acta medica Indonesiana* 2012;44(4):290-7. [published Online First: 2013/01/15]
33. Pendlebury ST, Lovett N, Smith SC, et al. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of externally derived risk scores. *Age and ageing* 2016;45(1):60-65. doi: 10.1093/ageing/afv177
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. *Journal of the American Geriatrics Society* 2006;54(5):817-22. doi: 10.1111/j.1532-5415.2006.00704.x [published Online First: 2006/05/16]
35. Kim MY, Park UJ, Kim HT, et al. DELirium Prediction Based on Hospital Information (Delphi) in General Surgery Patients. *Medicine (Baltimore)* 2016;95(12):e3072. doi: 10.1097/MD.0000000000003072
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. *Annals of surgery* 2014 doi: 10.1097/sla.0000000000000742 [published Online First: 2014/06/03]
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? *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* 2013;21(10):946-56. doi: 10.1016/j.jagp.2013.01.069 [published Online First: 2013/05/11]
38. Liang C-K, Chu C-L, Chou M-Y, 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(4):347-55. doi: 10.1089/rej.2014.1645
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* 2015 doi: 10.1111/ggi.12587
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(5) doi: 10.1136/bmjopen-2012-001599 [published Online First: 2012/09/18]

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(2):153-9. doi: 10.1016/j.genhosppsych.2011.11.011 [published Online First: 2012/01/10]
42. O'Keeffe ST, Lavan JN. Predicting delirium in elderly patients: development and validation of a risk-stratification model. *Age and ageing* 1996;25(4):317-21. [published Online First: 1996/07/01]
43. Pompei P, Foreman M, Rudberg MA, et al. Delirium in hospitalized older persons: outcomes and predictors. *Journal of the American Geriatrics Society* 1994;42(8):809-15. [published Online First: 1994/08/01]
44. Rudolph JL, Doherty K, Kelly B, et al. Validation of a Delirium Risk Assessment Using Electronic Medical Record Information. *Journal of the American Medical Directors Association* 2016;17(3):244-48 5p. doi: 10.1016/j.jamda.2015.10.020
45. Rudolph JL, Harrington MB, Lucatorto MA, et al. Validation of a medical record-based delirium risk assessment. *Journal of the American Geriatrics Society* 2011;59 Suppl 2:S289-94. doi: 10.1111/j.1532-5415.2011.03677.x
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(2):229-36 8p. doi: 10.1161/CIRCULATIONAHA.108.795260
47. Dworkin A, Lee DS, An AR, et al. A Simple Tool to Predict Development of Delirium After Elective Surgery. *Journal of the American Geriatrics Society* 2016;64(11):e149-e53. doi: 10.1111/jgs.14428 [published Online First: 2016/09/22]
48. Freter SH, George J, Dunbar MJ, et al. Prediction of delirium in fractured neck of femur as part of routine preoperative nursing care. *Age and ageing* 2005;34(4):387-8. doi: 10.1093/ageing/afi099 [published Online First: 2005/06/16]
49. 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 and ageing* 2016 doi: 10.1093/ageing/afw198 [published Online First: 2016/11/07]
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 (Clinical research ed)* 2015;350:g7594. doi: 10.1136/bmj.g7594 [published Online First: 2015/01/09]
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. *Advances in anatomic pathology* 2015;22(5):303-5. doi: 10.1097/pap.0000000000000072 [published Online First: 2015/08/12]
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. *Annals of internal medicine* 2015;162(1):W1-73. doi: 10.7326/m14-0698 [published Online First: 2015/01/07]
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(8):971-80. doi: 10.1093/aje/kwq223 [published Online First: 2010/09/03]

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*;0(0):0962280214558972. doi: doi:10.1177/0962280214558972
55. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005 [published Online First: 2015/05/20]
56. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal* 2014;35(29):1925-31. doi: 10.1093/eurheartj/ehu207 [published Online First: 2014/06/06]
57. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass)* 2010;21(1):128-38. doi: 10.1097/EDE.0b013e3181c30fb2 [published Online First: 2009/12/17]
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(12):1373-9. [published Online First: 1996/12/01]
59. Campbell DJ. The clinical utility curve: a proposal to improve the translation of information provided by prediction models to clinicians. *BMC research notes* 2016;9:219. doi: 10.1186/s13104-016-2028-0 [published Online First: 2016/04/16]
60. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ (Clinical research ed)* 2015;351:h3868. doi: 10.1136/bmj.h3868 [published Online First: 2015/08/13]
61. Grobman WA, Stamilio DM. Methods of clinical prediction. *American journal of obstetrics and gynecology* 2006;194(3):888-94. doi: 10.1016/j.ajog.2005.09.002 [published Online First: 2006/03/09]
62. Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? *Contemporary clinical trials* 2013;36(2):636-41. doi: 10.1016/j.cct.2013.06.011 [published Online First: 2013/07/03]
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. doi: 10.1016/j.jclinepi.2016.02.031 [published Online First: 2016/03/12]
64. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., et al. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Medical decision making : an international journal of the Society for Medical Decision Making* 2001;21(1):45-56. doi: 10.1177/0272989x0102100106 [published Online First: 2001/02/24]
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. *Journal of biomedical informatics* 2015;56:87-93. doi: 10.1016/j.jbi.2015.05.001 [published Online First: 2015/05/20]
66. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *Plos One* 2017;12(4):e0174944. doi: 10.1371/journal.pone.0174944 [published Online First: 2017/04/05]
67. Jones RN, Marcantonio ER, Saczynski JS, et al. Preoperative Cognitive Performance Dominates Risk for Delirium Among Older Adults. *J Geriatr Psychiatry Neurol* 2016 doi: 10.1177/0891988716666380 [published Online First: 2016/09/21]

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(10):1495-504. doi: 10.1016/j.psyneuen.2011.04.003 [published Online First: 2011/05/24]
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. doi: 10.1186/1471-2288-11-103 [published Online First: 2011/07/14]
70. Moore RC, Eyler LT, Mausbach BT, et al. Complex interplay between health and successful aging: role of perceived stress, resilience, and social support. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* 2015;23(6):622-32. doi: 10.1016/j.jagp.2014.08.004 [published Online First: 2014/09/14]
71. Arnberg FK, Lekander M, Morey JN, et al. Self-rated health and interleukin-6: Longitudinal relationships in older adults. *Brain, behavior, and immunity* 2016;54:226-32. doi: 10.1016/j.bbi.2016.02.008 [published Online First: 2016/02/14]
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 and ageing* 2017;46(6):932-39. doi: 10.1093/ageing/afx126 [published Online First: 2017/10/07]
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? *Research on aging* 2017;39(7):879-905. doi: 10.1177/0164027516637410 [published Online First: 2016/03/20]
74. Yates JA, Clare L, Woods RT. What is the Relationship between Health, Mood, and Mild Cognitive Impairment? *J Alzheimers Dis* 2017;55(3):1183-93. doi: 10.3233/jad-160611 [published Online First: 2016/10/30]
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. *Dementia & Geriatric Cognitive Disorders* 2014;38(1/2):65-78 14p. doi: 10.1159/000357454
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 Clinical And Experimental Research* 2007;19(3):172-77.
77. Fong TG, Hshieh TT, Wong B, et al. Neuropsychological profiles of an elderly cohort undergoing elective surgery and the relationship between cognitive performance and delirium. *Journal of the American Geriatrics Society* 2015;63(5):977-82. doi: 10.1111/jgs.13383
78. van Meenen LC, van Meenen DM, de Rooij SE, et al. Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *Journal of the American Geriatrics Society* 2014;62(12):2383-90. doi: 10.1111/jgs.1313810.1111/jgs.13138.</p>
79. Newman MW, O'Dwyer LC, Rosenthal L. Predicting delirium: A review of risk-stratification models. *General Hospital Psychiatry* 2015;37(5):408-13. doi: 10.1016/j.genhosppsych.2015.05.003

80. Campaner CP. Valoración del riesgo de delirium en pacientes mayores hospitalizados. = Delirium risk assessment in elderly hospitalized patients. *Revista Española de Geriatría y Gerontología* 2010;45(5):285-90. doi: 10.1016/j.regg.2010.03.011
81. Debray TP, Koffijberg H, Nieboer D, et al. Meta-analysis and aggregation of multiple published prediction models. *Stat Med* 2014;33(14):2341-62. doi: 10.1002/sim.6080 [published Online First: 2014/04/23]
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 journal of radiology* 2011;12(5):588-94. doi: 10.3348/kjr.2011.12.5.588 [published Online First: 2011/09/20]

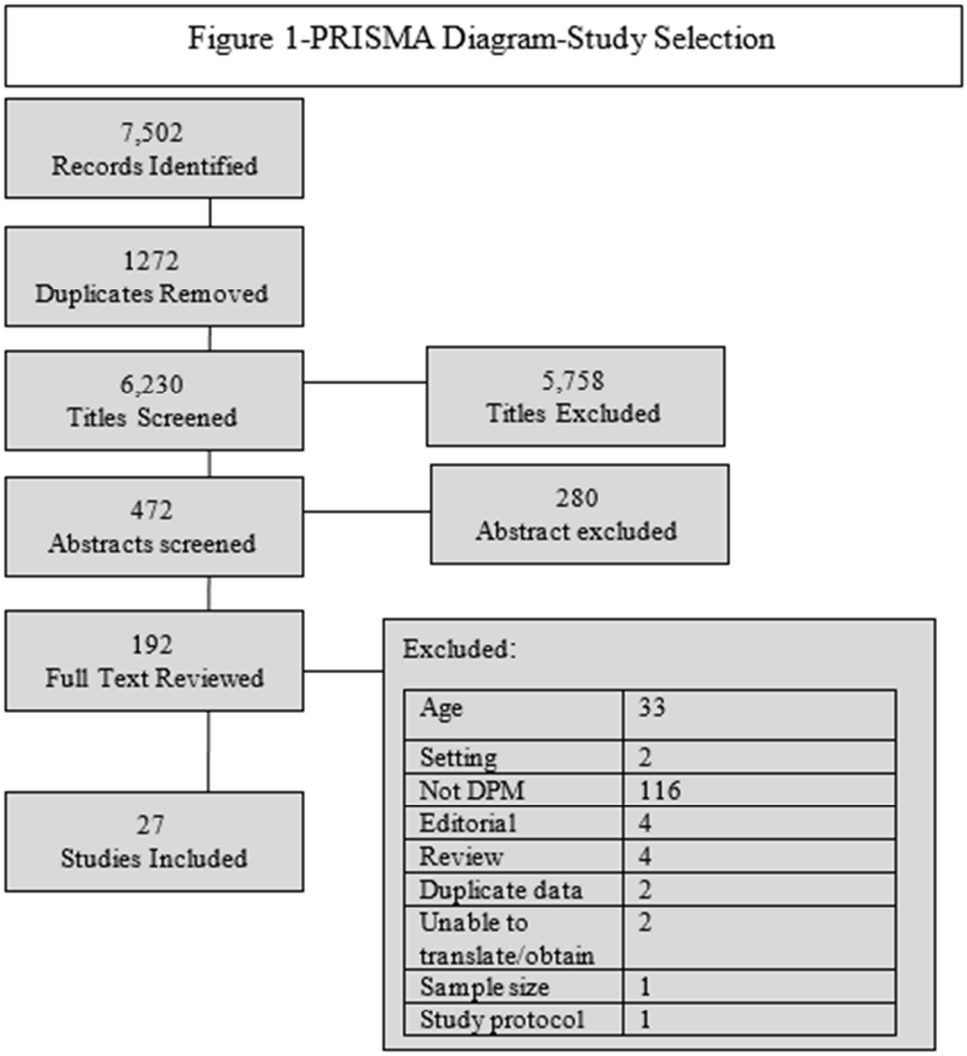


Figure 1: PRISMA Diagram - Study Selection

128x145mm (300 x 300 DPI)

Figure 2: CHARMS Risk of Bias Assessment

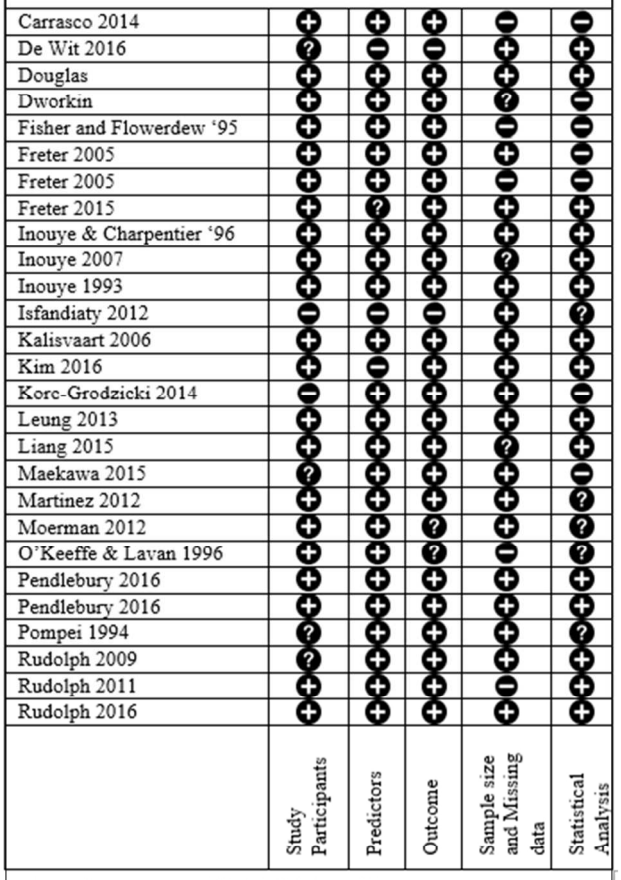


Figure 2 displays the CHARMS Risk of Bias assessment on all included studies.
Study Participants: design of included study, sampling method, inclusion/exclusion criteria
Predictors: definition, timing, and measurement
Outcome: definition, timing, and measurement
Sample Size and Missing Data: sample size, events per variable, missing data
Statistical Analysis: selection of predictors, internal validation, type of external validation done

Figure 2: CHARMS Risk of Bias Assessment

110x202mm (300 x 300 DPI)

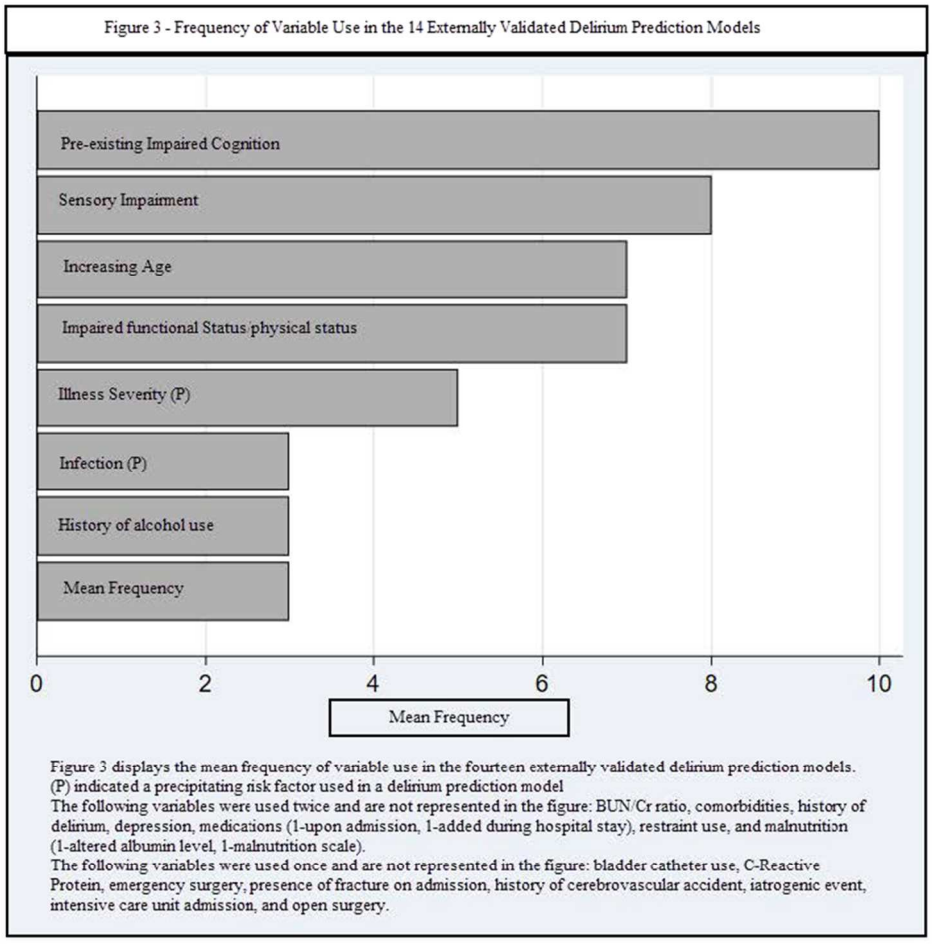


Figure 3: Frequency of Variable Use in the 14 Externally Validated Delirium Prediction Models

187x183mm (300 x 300 DPI)

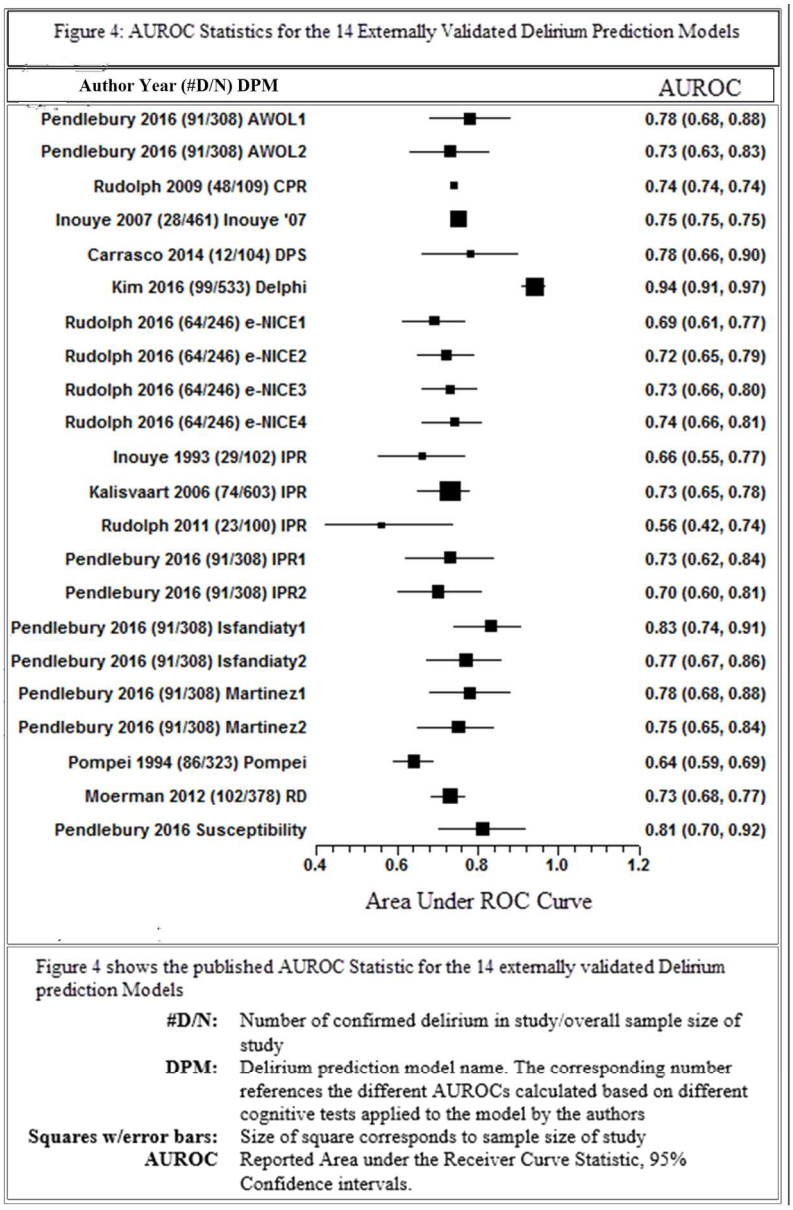


Figure 4: AUROC Statistics for the 14 Externally Validated Delirium Prediction Models

137x210mm (300 x 300 DPI)

Appendix A – Review Protocol

Working Title of Review	Systematic Review of Delirium Prediction Models			Support	Modifications
Authors	1st & Corresponding	Heidi Lindroth	Literature search, data extraction, data synthesis and manuscript preparation.		
	Data Extraction	Heidi Lindroth Suzanne Purvis	Literature search, data extraction, data synthesis.		
	Content Experts	Lisa Bratzke	Assisted with content related to cognition. Results review.		
		Roger Brown	Statistical content expert		
		Mark Coburn	Results review, Manuscript preparation		
		Marko Mrkobrada	Results review, Manuscript preparation		
		Matthew TV Chan	Results review, Manuscript preparation		
		Daniel Davis	Geriatrician expertise, reviewed results, manuscript preparation.		
		Pratik Pandharipande	Results review, Manuscript preparation		
		Cynthia M. Carlsson	Geriatrician expertise, reviewed results, manuscript preparation.		
	Mentoring	Robert D. Sanders	Mentoring author, resolved content/data disagreements b/w authors, manuscript preparation.		
Aim	To identify existing prognostic delirium prediction models and evaluate their validity and statistical methodology in the older adult (≥60yo) acute hospital population.				
Search Terms	(“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*)			UW-Madison Health Sciences librarian. Three meetings to refine search terms.	
Databases searched	PubMed, CINAHL, PsychINFO, Cochrane, SocINDEX and Medline			Health Sciences librarian.	Expanded to include SocINDEX
Timelines established	01/01/1990-12/31/2016				Originally was 12/31/15.

			Expanded to include all of 2016.
Inclusion criteria	• Age ≥ 60 • Inpatient population • Developing and/or validating a delirium prediction model		Age expanded from ≥ 70 years of age due to the literature
Exclusion criteria	• Emergency department • Hospice/palliative care • Pediatric population • Related to alcohol withdrawal • ≤50 sample size	Mentoring author	Sample size criteria added to build rigor in the studies that were included in the sys review
Selection process	Studies will be selected based on the inclusion/exclusion criteria. The data extraction authors (HL and SP) will conduct the literature search independently and meet monthly to discuss findings. Any disagreements will be resolved by the mentoring author (RDS)		
Data Management	A shared folder on the UW-Madison Box account will be created to share documents, data and meeting information.		
Data collection process	Data will be collected independently by HL and SP then data points will be shared at monthly meetings. Data collection tables will be created using Microsoft Excel then uploaded to the shared Box account. Any disagreement between authors will be resolved by the mentoring author (RDS).		
Data points collected	• Characteristics of studies (design, population, sample size) • Outcome measure including how it was identified, measured, defined. Prevalence. • Statistical methods applied • Statistical information about the delirium prediction models (sensitivity, specificity, positive predictive value, negative predictive value, AUROC) • Characteristics of DPMs (variables used, scoring, development) • Cognitive measures used in studies. • Criteria to fulfill the Newcastle Ottawa Scale.		
Outcomes	• AUROC will be the primary outcome measure • Characteristics of DPMs (variables, statistics) • Cognitive tests used		

Data synthesis	The first/corresponding author (HL) will synthesize the data into the manuscript. The co-authors will verify this. RB will complete the meta-analysis.		
Manuscript preparation	HL will complete manuscript preparation. All co-authors are responsible for reviewing content and data to assure correctness and complete synthesis of data gathered.		

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	4
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4-5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5-6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6-7
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	7-8Table1, Figure 2
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Table 1,2 Figure 4
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Table 1, Figure 2
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8-19
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	19-20
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	19-24
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	19-26
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	6

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

PRISMA 2009 Checklist

CHARMS 2014 Relevant items to extract from individual studies in a systematic review of prediction models

Domain	Key items	Reported on page #
SOURCE OF DATA	Source of data (e.g., cohort, case-control, randomized trial participants, or registry data)	7
PARTICIPANTS	Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centers, setting, inclusion and exclusion criteria)	7
	Participant description	7
	Details of treatments received, if relevant	7
	Study dates	7
OUTCOME(S) TO BE PREDICTED	Definition and method for measurement of outcome	11
	Was the same outcome definition (and method for measurement) used in all patients?	11
	Type of outcome (e.g., single or combined endpoints)	11
	Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)?	11
	Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)?	11
	Time of outcome occurrence or summary of duration of follow-up	11
CANDIDATE PREDICTORS (OR INDEX TESTS)	Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics)	12-17
	Definition and method for measurement of candidate predictors	12-17
	Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation)	7
	Were predictors assessed blinded for outcome, and for each other (if relevant)?	N/A
	Handling of predictors in the modelling (e.g., continuous, linear, non-linear transformations or categorised)	11-12
SAMPLE SIZE	Number of participants and number of outcomes/events	Table 1
	Number of outcomes/events in relation to the number of candidate predictors (Events Per Variable)	18
MISSING DATA	Number of participants with any missing value (include predictors and outcomes)	Appendix B
	Number of participants with missing data for each predictor	
	Handling of missing data (e.g., complete-case analysis, imputation, or other methods)	Appendix B
MODEL DEVELOPMENT	Modelling method (e.g., logistic, survival, neural network, or machine learning techniques)	Table 1, pg 11
	Modelling assumptions satisfied	Not reported
	Method for selection of predictors **for inclusion** in multivariable modelling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)	12
	Method for selection of predictors **during multivariable modelling** (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion)	12
	Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)	12,18
MODEL PERFORMANCE	Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (C-statistic, D-statistic, log-rank) measures with confidence intervals	Table 2, 17-18
	Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a-priori cut points were used	Table 2, 18-19
MODEL EVALUATION	Method used for testing model performance: development dataset only (random split of data, resampling methods e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators)	12
	In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)	Table 2
RESULTS	Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals)	Table 2, 17-18
	Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance	Table 2
	Comparison of the distribution of predictors (including missing data) for development and validation datasets	Table 2
INTERPRETATION AND DISCUSSION	Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)	19-26
	Comparison with other studies, discussion of generalizability, strengths and limitations.	19-26

BMJ Open

Systematic Review of Prediction Models for Delirium in the Older Adult Inpatient

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019223.R2
Article Type:	Research
Date Submitted by the Author:	13-Mar-2018
Complete List of Authors:	Lindroth, Heidi; University of Wisconsin Madison School of Medicine and Public Health, Department of Anesthesiology; University of Wisconsin Madison Graduate School, School of Nursing Bratzke, Lisa; University of Wisconsin-Madison, School of Nursing Purvis, Suzanne; University of Wisconsin Hospital and Clinics, Nursing Brown, Roger; University of Wisconsin Madison, School of Nursing Coburn, Mark; Uniklinik RWTH Aachen, Anaesthesiology Mrkobrada, Marko; Western University, Medicine Chan, MTV; The Chinese University of Hong Kong, Anesthesia and Intensive Care Davis, Daniel; University College London, MRC Unit for Lifelong Health and Ageing ; University College London Pandharipande, Pratik; Vanderbilt University School of Medicine Carlsson, Cynthia; William S. Middleton Veteran's Administration Geriatric Research Education and Clinical Center (GRECC), GRECC; University of Wisconsin Madison School of Medicine and Public Health Sanders, Robert; University of Wisconsin Madison School of Medicine and Public Health, Department of Anesthesiology
Primary Subject Heading:	Geriatric medicine
Secondary Subject Heading:	Neurology
Keywords:	Delirium, GERIATRIC MEDICINE, Statistic

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 TV Chan, Daniel Davis, Pratik Pandharipande, Cynthia M. Carlsson, Robert D. Sanders
University of Wisconsin School of Medicine and Public Health, Department of Anesthesiology, School of Nursing, 600 Highland Avenue, Madison, WI 53792 Heidi Lindroth BSN PhD Candidate[1]

University of Wisconsin School of Nursing, 701 Highland Avenue, Madison, WI 53705 Lisa Bratzke PhD Assistant Professor[2]

University Hospital, 600 Highland Avenue, Madison, WI 53792 Suzanne Purvis DNP, Clincial Nurse Specialist[3]

University of Wisconsin School of Nursing, 701 Highland Avenue, Madison, WI 53705 Roger L. Brown PhD, Professor[4]

University Hospital RWTH Aachen, Department of Anesthesiology, Pauwelsstrasse 30, 52074 Aachen, Mark Coburn MD, Professor[5]

Western University, Schulich School of Medicine and Dentistry, 339 Windermere Road, Room B9-100, London ON N5X 3W6, Marko Mrkobrada MD Associate Professor[6]

The Chinese University of Hong Kong, Department of Anaesthesia and Intensive Care, 4/F Main Clinical Bock and Trauma Center, Prince of Wales Hospital, Hong Kong Special Administrative Region, China, Matthew TV Chan, MB, BS, PhD, Professor[7]

MRC Unit for Lifelong Health and Ageing at UCL, 33 Bedford Place, London, WC1B 5JU, Daniel Davis PhD MRCP, Senior Clinical Researcher[8]

Vanderbilt University Medical Center, Department of Anesthesiology, 1211 Medical Center Drive, Nashville TN 37322, Pratik Pandharipande MD, Chief, Division of Anesthesiology Critical Care Medicine[9]

University of Wisconsin School of Medicine and Public Health, Madison VA Geriatric Research Education and Clinical Center (GRECC) (D4211), 2500 Overlook Terrace, Madison, WI, 53705, Cynthia M. Carlsson, MD, MS, Associate Professor of Medicine, Division of Geriatrics[10]

University of Wisconsin School of Medicine and Public Health, Department of Anesthesiology, 600 Highland Avenue, Madison, WI 53792, Robert D. Sanders MBBS PhD FRCA, Assistant Professor[11]

[+]**Corresponding author**: *Heidi Lindroth, Department of Anesthesiology, University of Wisconsin School of Medicine and Public Health, 600 Highland Avenue, B6/319 CSC Madison, WI 53792-3272 Telephone: 608-262-6469 Fax: 608-263-0575 Madison, USA. Email:* hlindroth@wisc.edu

Word count: 4043

Objective

To identify existing prognostic delirium prediction models and evaluate their validity and statistical methodology in the older adult (≥60yo) 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 PRISMA and CHARMS Statement guided protocol development. Inclusion criteria: Age ≥60, 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 (<60yrs) and 27 met the defined criteria. Twenty-three delirium prediction models were identified, fourteen were externally validated and three 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 AUROC range from 0.52-0.94. 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.

Funding

Department of Anesthesiology, University of Wisconsin-Madison.

1K23AG055700-01A1

R01 NHLBI (HL111111) and research grant from Hospira Inc., NIH collaboration

All authors completed ICMJE disclosure forms and no conflicts of interest declared.

Keywords

Delirium. Aging. Cognition. Prediction. Statistical Models.

Strengths and Limitations of this Study

- This study used the PRISMA 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, older adult inpatients, and could be expanded to include all ages and settings including palliative care, long term care and the emergency room.

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] Delirium has been independently associated with increased mortality, morbidity in terms of impaired cognition and functional disability along with an estimated annual U.S. 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 formalize 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 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 PRISMA Statement and the CHARMS checklist (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 predetermined modifiable and non-modifiable risk factors of delirium present. This review included studies focused on 1) older adult (≥ 60 years) population, (the U.S. Center 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 to 31 December 2016, and 4) developed and/or validated delirium prediction models. Studies were excluded if they 1) studied a different patient population (i.e. emergency department, skilled nursing facilities, palliative care, and hospice) as these are unique patient populations with characteristics requiring specific foci and are not readily generalizable to a medical or surgical inpatient hospital setting. Further, 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 ≤ 50 for methodological reasons (i.e. underpowered). All study designs were included. Studies were not limited by timeframe 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 CHARMS checklist.[15] Two authors (HL, 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, 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 4) characteristics of the models (variables used in model, 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, odds ratios, 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 Chi-square (X^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 pre-planned outcome measures included cognitive assessments, and predictive variable use per model.

Role of the Funding Source

The funding sources named has 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).[25 40]

Twenty-three delirium prediction models were developed, fourteen 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 23 studies.[23 25-31 33-35 37-49] Retrospective design was 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 randomized control trial.[34 41] Eleven studies focused on the medical population [23 25 29-33 40 42 45 48], three included medical and surgical [24 43 44] and thirteen recruited a surgical population (seven-orthopaedic [26-28 34 38 41 49], one-cardiac [46], two-noncardiac [37 47], one general surgery[35], two-oncological[36 39]). None of the identified studies focused on critical care patients. Data collection occurred upon admission in seventeen studies [23 25 27 29-31 33-35 40-45 48 49]; participants were approached within forty-eight hours of admission. Seven studies collected data pre-operatively

then followed participants post-operatively.[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.

Table 1

Author	Study Design Population Sample Size Sampling method Power Analysis	Study Grade (NOS)	Outcome Variable & Rate (%)	Delirium measurement & frequency	DPM Name & Regression Model used
Carrasco et al. (2014)[23]	P.Cohort Medical Dev: 374 Val: 104 Consecutive	S: **** C: - O: ** T: 6 stars	Delirium Dev: 25 (.06) Val: 12 (12)	CAM Every 48 h	Predictive Risk Score Forward stepwise
de Wit et al. (2016)[24]	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.** (2013)[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)	CAM-S & CAM Daily	Risk Stratification model (AWOL) Forward stepwise
Dworkin et al. (2016)[47]	P.Cohort Elective noncardiac surg Dev: 76 Consecutive	S: **** C: - O: ** T: 6 stars	Delirium Dev: 10(13)	CAM or FAM-CAM 1xafter surgery	Mini-Cog Stratified into a five-point score Stepwise
Fisher and Flowerdew (1995)[26]	P.Cohort Elective Orthopedic Dev: 80 Consecutive	S: ** C: - O: ** T: 4 stars	Delirium Dev: 14(17.5)	CAM 2xDaily	Prediction Model using two variables. Stewpsie
Freter et al. (2005)[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. (2005)[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 al. (2015)[27]	P.Cohort Hip Fracture Val: 283 Consecutive	S: *** C: - O: ** T: 5 stars	Delirium Val: 119(42)	CAM POD1, 3 & 5	Risk stratification model (DEAR)
Inouye and Charpentier (1996)[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
Inouye et al. (2007)[31]	P.Cohort Medical Dev: 491 Val: 461 Consecutive	S: **** C: ** O: *** T: 9 stars	Delirium/ subsyndrome delirium at discharge Dev: 58(12) Val: 28(6)	CAM Every other day	Risk stratification model Log-binomial regression
Inouye et la. (1993)[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
Isfandiaty et al. (2012)[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 et al. (2006)[34]	P.Cohort Hip Surgery & Facture Val: 603 Consecutive	S: *** C: - O: *** T: 6 stars	Delirium Dev: 74(12)	CAM, DRS-98 Daily through POD5	Externally validated Inouye's '93 model.
Kim et al. (2016)[35]	P.Cohort Major General	S: *** C: **	Delirium Dev: 112(20)	Nu-Desc -every shift by RNs	Risk stratification model Backwards stepwise

	Surgery Dev: 561 Val: 533 Not stated Power analysis	O: *** T: 8 stars	Val: 99(18)	Confirmed with CAM.	
Korc-Grodzicki et al. (2014)[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 et al. (2013)[37]	P.Cohort Noncardiac surgery Dev: 581 Not stated	S: *** C: - O: ** T: 5 stars	Delirium Dev: 234(40)	CAM Daily	Risk stratification model Stepwise
Liang et al. (2015)[38]	P.Cohort Elective Orthopedic Surgery Dev: 461 Consecutive	S: *** C: ** O: ** T: 7 stars	Delirium Dev: 37(8)	CAM Daily Confirmed by psychologist DSM-IV	Built 2 DPMs CGA Risk stratification models Backward stepwise
Maekawa et al. (2015)[39]	P.Cohort Oncological; Gastrointestinal Surgery Dev: 517 Consecutive	S: ** C: * O: *** T: 6 stars	Delirium Dev: 124(24)	CAM Unknown frequency	Comprehensive Geriatric Assessment (CGA) as model. Proportional hazards
Martinez et al.(2012)[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 et al. (2012)[41]	P.Cohort Hip Fracture Val: 378 Consecutive Power analysis	S: *** C: - O: *** T: 6 stars	Delirium Val: 102(27)	Ward RN observation, 3xdaily Confirmed by chart review.	Risk stratification model (Risk Model for Delirium, RD) Built from literature
O'Keeffe and Lavan (1996)[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 et al. (2016)[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 et al. (2016)[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 4 DPMs
Pompei et al. (1994)[43]	P.Cohort Med/surg Dev: 432 V: 323 Not stated	S: **** C: ** O: *** T: 9 stars	Delirium Dev: 64(14.8) Val: 86(26.3)	CAM 2xweekly. Confirmed with DSM III	Risk stratification model Stepwise
Rudolph et al. (2009)[46]	P.Cohort Cardiac Surgery Dev: 122 V: 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 et al. (2011)[45]	P.Cohort Medical V: 100 Consecutive	S: **** C: - O: *** T: 7 stars	Delirium Dev: 23(23)	DSM-IV Daily clinical interview	Externally validated Inouye's '93 model.
Rudolph et al. (2016)[44]	Dev: Retro Val: P.Cohort Med/surg Dev: 27625 Val: 246	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

	Consecutive				

Key:
**=Models developed in population ≤60 years of age, but validated in population ≥60 years of age.
Study Design: P.Cohort=Prospective Cohort, Retro=Retrospective design. Dev=Development, Val=Validation. Med=Medical, Surg=Surgical. Power analysis = reported in identified study.
Study Grade: NOS=Newcastle Ottawa Scale. S=Selection, C=Comparability, O=Ottawa. Max 9 stars.
Outcome Variable: Dev=Development, Val=Validation
Delirium Measurement: CAM=Confusion Assessment Method, DSM=Diagnostic Statistical Manual, POD=Postoperative Day, MDAS=Memorial Delirium Assessment Scale, Nu-Desc=Nursing Delirium Screening Scale, DRS-98=Delirium Rating Scale, EHR=Electronic Health Record
Type of Model: How authors designed their delirium prediction model (DPM), statistical method used
-Risk stratification model: Points (weighted or un-weighted) assigned per predictive risk factor present.
-CGA=Comprehensive Geriatric Assessment
-Built from Literature: Authors selected risk factors for DPM based on literature review.

Delirium assessment

The outcome variable was measured using the Confusion Assessment Method in twenty-one studies.[23 25-31 33-40 43 46-49] The frequency of delirium assessment varied from two or more assessments daily (three studies)[26 35 41], to once daily (twelve studies)[25 28 30 32 34 36-38 44-46 49], every-other day (eight studies)[23 27 29 31 33 42 43 48], once following surgery[47], and undefined (three 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 recognize 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]

Model design and statistical methods

Various statistical techniques were employed by the twenty-three included studies. Twelve used univariate or bivariate analyses and selected variables with a pre-determined statistical value

(range for 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 modeling paired with bivariate analyses and included variables with either a *p*-value <0.25[32] or a relative risk of ≥1.5.[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 categorized 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 eight of the fourteen externally validated models are categorized 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 fourteen 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 upon study enrollment 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 upon enrollment 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 mBDRS[31], and one study-dementia billing codes or prescription information.[44] One study defined baseline cognitive impairment as a pre-specified 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) needing assistance with any basic ADL,[27] (1) domestic help, help with meals or physical care[41] and (2) residence in nursing facility or at home with caregivers[33], (2) requiring a home care package with professional caregivers or residence in a care home.[33 48] The latter being obtained upon admission from medical records.[33 48] Two studies used validated functional assessment tools (iADL and Barthel Index) and evaluated functional status two weeks prior to hospitalization.[23 31]

Externally validated delirium prediction models are detailed in Table 2.

Table 2

External Validated DPM Name	Citation Type of Validation	Delirium #(%)	Sens Spec PPV NPV (external)	AUROC (95%CI)	Model Components	Cog. Assess Tool & Cutoff
AWOL Tool	Pendlebury et al. (2016)[33] Broad_val.	1st Val: 14(9) 2nd Val: 95(31) (any delirium) 67-prevalent 28-incident	Mod. AWOL Cutoff - 3 Any Delirium Sens ·7 Spec .66 PPV .55 NPV .79 Incident Del Sens .76 Spec .66 PPV .27 NPV .94	1st Val: 0.69 (0·54-0·83) Incident delirium 2nd Val: Cohort 1 (MMSE) 0·78 (0·68-0·88) Cohort 2 (AMTS) 0·73 (0·63-0·83)	Original AWOL Tool Age >80 — 1 pt Failure to spell WORLD backwards — 1 pt Disorientation — 1 pt Illness Severity — 1 pt Modified AWOL Tool Age >80 — 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 et al. (2009)[46]	Dev: 63(52) Val: 48(44)	Not reported	Dev: 0·74 Val: 0·75 Did not report	Weighted Points-Regression MMSE ≤ 23 — 2 pt MMSE 24-27 — 1 pt	MMSE -Stratified score

	Narrow val.	(incident delirium)		CI	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)	
DEAR	Freter et al. (2015)[27] Narrow val.	Dev: (2005) 18(14) Val: (2015) Pre-Op= 163(58) Post-op= 118(42)	Sens .68 Spec .73 PPV .65 NPV .76 Optimal cut-off score: 3pts (Incident post-op delirium)	Dev: (2005) 0·77 (0·64-0·87) Val: (2015) AUROC Not published	MMSE ≤ 23: 1 pt Functional dependence: 1 pt Sensory impairment: 1 pt Substance use: 1 pt Age >80: 1 pt Not weighted. 0-5 Score, cut-off of 3 indicating high risk.	MMSE Cut-off ≤ 23
Delirium at Discharge Prediction Model	Inouye et al. (2007)[31] Narrow val.	Dev: 58(12) Val: 28(6) (incident delirium)	Not reported	Dev: 0.80 Val: 0.75 Did not report CI Calibration: X^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 pt = 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 et al. (2014)[23] Narrow val.	Dev: 25(.06) Val: 12(12) (incident delirium)	Sens .88 Spec .74 PPV .22 NPV .99	Dev: 0.86 (0.82-0.91) Val: 0.78 (0.66-0.90)	DPS=[5xBUN/Cr ratio]-(3xBarthel 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
Delphi Score	Kim et al. (2016)[35] Narrow val.	Dev: 112(20) Val: 99(18) (incident delirium)	Sens .81 Spec .93 PPV .70 NPV .96 Optimal cut-off score: 6.5pts	Dev: 0.911 (0.88-0.94) Val: 0.938 (0.91-0.97)	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	No measure of cognition. Excluded participants if MMSE <24

					Yes 2 ICU Admission No 0 Yes 3 Pre-Op CRP (mg/dL) <10 0 ≥10 1 Max points: 15 Optimal cut-off: 6.5 High Risk: ≥7 points	
e-NICE Rule	**Rudolph et al. (2016)[44]** Broad val. Dev: 2343(8) Val:64(26) (incident delirium)	Cohort / AUROC / CI / TPR / FPR Dev 0.81 (0.80-0.82) Validation AUROCs* Original 0.69 (0.61-0.77) 64% 33% mRASS 0.72 (0.65-0.79) 69% 35% TMYB 0.73 (0.66-0.80) 78% 43% MoCA 0.74 (0.66-0.81) 75% 43% *Any delirium Original model-AUROC of 0.68 (95%CI0.59-0.77) in incident delirium. Did not report sens, spec, PPV, NPV			Weighted Points/OR Cog impair -Medications, diagnosis or both 4 pt Age ≥ 65 y 2 pt Age ≥ 80 y 3 pt Infection 2 pt Fracture 4 pt Vision 1 pt Severe Illness 2 pt 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 et al. (1993)[30] Narrow val.	Dev: 27(25) Val: 29(17) (incident delirium)	Did not report	Dev: 0.74 (0.63-0.85) Val: 0.66 (0.55-0.77) Calibration: Dev: X^2 Trend $p<0{\cdot}00001$ Val: X^2 Trend $p<0{\cdot}002$	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 pts = 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/care giver bDRS Excluded those w/history of severe dementia
IPR	Kalisvaart et al. (2006)[34] Broad val.	Val: 74(12)	Did not report	Val: 0.73 (0.65-0.78) Calibration: X^2 $p<0{\cdot}05$ X^2 Trend $p<0{\cdot}002$	Externally validated IPR in surgical hip fracture population. -Addition of age & type of admission improved model performance, R^2=0.20 RR of High risk group: 9.8	MMSE Cut-off < 24
IPR	Rudolph et al. (2011)[45] Broad val.	Val: 23(23) Any delirium 10-Prevalent 13-Incident	Did not report	Val: 0.56 (0.42-0.74) Incident del. 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
IPR	Pendlebury et al. (2016)[33] Broad val.	Val: 95(31) Any delirium 67-prevalent 28-incident	Cutoff 2pts All Delirium Sens .57 Spec .80 PPV .64 NPV .76	Val: Incident delirium Cohort 1 (MMSE) 0.73	Baseline cognitive impairment 1 pt High BUN/Cr ratio 1 pt Severe illness (SIRS ≥ 2) 1 pt Vision impairment 1 pt	Original Model: MMSE <24 Modified Model:

			Incident D Sens .52 Spec .80 PPV .31 NPV .91	(0.62-0.84) Cohort 2-(AMTS) 0.70 (0.60-0.81)	4pts=Incident delirium	MMSE < 24 AMTS < 9
Isfandiaty model	Pendlebury et al. (2016)[33] Broad val.	Dev: 87(19) Val: 95 (31) Any delirium 67-prevalent 28-incident	Cutoff 4pts Any Delirium Sens .74 Spec .71 PPV .60 NPV .82 Incident Del Sens .81 Spec .71 PPV .31 NPV .96	Dev: 0.82 (0.77-0.88) Val: Incident delirium Cohort 1 (MMSE) 0.83 (0.74-0.91) Cohort 2 (AMTS) 0.77 (0.67-0.86)	Baseline cognitive impairment: 3 pt Functional dependency: 2 pt Infection w/sepsis: 2 pt Infection w/out sepsis: 1 pt Weighted Score Score = 7 for incident delirium Cohort 1: MMSE Cohort 2: AMTS	<u>Original Model:</u> Chart review <u>Modified Model:</u> MMSE < 24 AMTS < 9
Martinez et al. 2012 model	Pendlebury et al. (2016)[33] Broad val.	1st Val: 76(25) 2nd Val: 95(31) Any delirium 67-prevalent 28-incident	Modified Model Cutoff 2pts Any Delirium Sens .62 Spec .68 PPV .54 NPV .75 Incident Del Sens .81 Spec .68 PPV .29 NPV .96	1st Val: 0.85 (0.80-0.88) Incident delirium 2nd Val: Cohort 1 (MMSE) 0.78 (0.68-0.88) Cohort 2 (AMTS) 0.75 (0.65-0.84)	Martinez et al. 2012 Original Model Age >85: 1 pt Dependent in ≥5 ADLs: 1 pt Drugs on admit: -Antidepressants -Antidementia -anticonvulsants -antipsychotics: 1pt/drug 2pt/ antipsych Score 0-3 Score >1 = high risk for delirium Modified Model Age >85: 1 pt Dependency in ≥ 5 ADLs: 1 pt Diag of Dementia MMSE<24 AMTS<9: 1 pt	<u>Original Model:</u> -No cognitive measure <u>Modified Model:</u> MMSE < 24 AMTS < 9
Pompei et al. 1994 model	Pompei et al. (1994)[43] Broad val.	Dev: 64(15) Val: 86(26) (21=prevalent delirium)	Sens .83 Spec .50 PPV .38 NPV .89 *Pts stratified as low or moderate to high-risk	Dev: 0.74 +/- 0.05 Val: 0.64 +/- 0.05 Calibration: X^2 Trend $p<0{\cdot}0001$	Weighted Points Baseline cognitive impairment: 2 pt Depression: 2 pt Alcoholism: 3 pt ≥ 4 comorbidities: 3 pt 0-3 pts = Low risk 4-7 pts = Moderate risk 8-10 pts = High risk	MMSE Less than HS <21 High school <23 College edu < 24
Precipitating Risk Factors	Inouye and Charpentier (1996)[29] Narrow val.	Dev: 35(18) Val: 47(15) (incident delirium)	Not reported	No AUROC reported Calibration: X^2 Trend $p<0{\cdot}001$	Physical restraint use: 1 pt Malnutrition: 1 pt ≥3 medications added: 1 pt Bladder catherization: 1 pt Any iatrogenic event: 1 pt Not weighted. 0 pt = Low Risk 1-2 pt = Intermediate ≥ 3 pt = High Risk RR of High Risk: 17.5 (8.1-37.4)	None used in model

<table>
<tr>
<td>Risk Model for Delirium (RD)</td>
<td>Moerman et al. (2012)⁴¹

Narrow val.</td>
<td>Val: 102(27)

(incident delirium)</td>
<td>Sens .81
Spec .56
PPV .41
NPV .89

Optimal cut-off score:
4 pts</td>
<td>Val: 0.73 (0.68-0.77)</td>
<td>Weighted Points
<table>
<tr><td>Delirium-previous hospitalization</td><td>5 pt</td></tr>
<tr><td>Dementia</td><td>5 pt</td></tr>
<tr><td>Clock Drawing</td><td></td></tr>
<tr><td>-Sm mistake</td><td>1 pt</td></tr>
<tr><td>-big mistake</td><td>2 pt</td></tr>
<tr><td>Age</td><td></td></tr>
<tr><td>-70 to 85 years old</td><td>1 pt</td></tr>
<tr><td>- >85 years</td><td>2 pt</td></tr>
<tr><td>Impaired hearing</td><td>1 pt</td></tr>
<tr><td>Impaired vision</td><td>1 pt</td></tr>
<tr><td>Problems w/ADL</td><td></td></tr>
<tr><td>-Help w/meal prep</td><td>.5p</td></tr>
<tr><td>-help w/physical</td><td>.5p</td></tr>
<tr><td>Use of heroin, methadone, morphine</td><td>2 pt</td></tr>
<tr><td>Daily >4 alcohol</td><td>2 pt</td></tr>
</table>
≥ 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>Pendlebury et al. (2016)⁴⁸

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 (0-70-0.92)

Improved w/age eliminated to 0.84 (0.77-0.92)</td>
<td>Weighted Points
<table>
<tr><td>Dementia/cog impair</td><td>2</td></tr>
<tr><td>Age >80 years</td><td>2</td></tr>
<tr><td>Severe illness (SIRS+)</td><td>1</td></tr>
<tr><td>Infection-working diagnosis</td><td>1</td></tr>
<tr><td>Vision impairment</td><td>1</td></tr>
</table>
>5 pts=High Risk

ORs for >5 risk score: 25.0 (3.0-208.9)
RR for >5 risk score: 5.4</td>
<td>Known diagnosis of dementia or MMSE < 24 AMTS < 9</td>
</tr>
<tr>
<td colspan="6">Key:
Dev=Development, Val=Validation
Sens=Sensitivity, Spec=Specificity, PPV=Positive Predictive Value, NPV=Negative Predictive Value
Area Under the Receiver Operating Curve Statistic, Dev=Development, Val=Validation, mRASS=Modified Richmond Agitation-Sedation Scale, TMTYB=The Months of the Year Backwards
ADL=Activities of Daily Living
MMSE=Mini Mental Status Exam, AMTS=Abbreviated Mental Test Score, CDT=Clock Drawing Test, mBDR=Modified Blessed Dementia Rating, MoCA=Montreal Cognitive Assessment.</td>
</tr>
</table>

Predictive ability

Reported AUROC in externally validated delirium prediction models ranged from 0.52-0.94 (Figure 4). Of these models, the highest performing model (AUROC 0.94, CI 0.91-0.97) 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 fourteen externally validated delirium prediction models reported calibration metrics.[29-31 34 43 45] The reported chi-square 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 reported calibration statistics.[32 37 40 42] None of the included studies reported calibration plots or slopes.

Risk of overfitting

Events per variable (EPV) were examined in each of the fourteen 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 penalized 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 odds ratios, relative risk, sensitivity and specificity, receiver operator curves, R squared 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 odds ratios 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 curse 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. (2016) 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]

DISCUSSION

This review identified moderate predictive ability (AUROC 0.52-0.94) in fourteen externally validated delirium prediction models with eight out of fourteen 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. (2016) 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 interrelationship between premorbid and precipitating factors,[29] the time course of data collection is important. Nine of the fourteen 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 (1993).[30] However, if variables are collected after the onset of delirium this would exaggerate model performance (e.g. ICU admission). As an example, one delirium prediction model has a robust AUROC of 0.94 (CI 0.91-0.97).[35] This study excluded those with a 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 upon 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 events per

variable (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. (2016) 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] Further, 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, 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[67] 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. (2016), 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] Further, 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. (2016) 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. (2015) 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. (2015), 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 twenty-seven studies with fourteen externally validated delirium prediction models. Our author team is interprofessional, providing the opportunity for different perspectives on model evaluation. Further, this review synthesizes 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. Further, 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. Further 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 generalizability and bolstering broad external validation.[81] Variable definition could be harmonized 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 re-fitting. 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 mild cognitive impairment and sensitive to the spectrum of cognition. Further, 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 Steyerberg (2009)[22] and Steyerberg and Vergouwe (2014)[56] including strategies to

counter low sample size and overly optimistic model performance, the use of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to assess model fit, and consider broad validations to expand case-mix and generalizability; 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 0.52-0.94. Assessment of the outcome variable, delirium, is often non-systematic and future studies would be improved with more standardized 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 generalizable to several populations within the remit of understanding the relationship with the precipitating event.

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 synthesized 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. DD 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.

Declaration of Interests

All authors have completed ICMJE disclosure forms and no conflicts of interest are declared.

Data Sharing Statement

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

Acknowledgements

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 NIH. 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 and Lily Turner for their assistance with proofreading the manuscript.

Figure Legends

Figure 1: No legend

Figure 2: Figure 2 displays the CHARMS Risk of Bias assessment on all included studies.
Study Participants: design of included study, sampling method, 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,

missing data

Statistical Analysis: Selection of predictors, internal validation, type of external validation

Figure 3: Figure 3 displays the mean frequency of variable use in the fourteen 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, 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.

Figure 4: Figure 4 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 w/error bars: Size of square corresponds to sample size of study

AUROC: Reported Area Under the Receiver Curve Statistic, 95% Confidence Intervals

References

1. Sanders RD, Pandharipande PP, Davidson AJ, et al. Anticipating and managing postoperative delirium and cognitive decline in adults. *BMJ: British Medical Journal (Overseas & Retired Doctors Edition)* 2011;343(7816):197-200 4p. doi: 10.1136/bmj.d4331
2. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet* 2014;383(9920):911-22. doi: 10.1016/s0140-6736(13)60688-1

10.1016/S0140-6736(13)60688-1. Epub 2013 Aug 28.</p>
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(1) doi: 10.1136/bmjopen-2012-001772
4. Pandharipande PP, Girard TD, Jackson JC, et al. Long-Term Cognitive Impairment after Critical Illness. *The New England journal of medicine* 2013;369(14):1306-16. doi: 10.1056/NEJMoa1301372
5. Leslie DL, Inouye SK. The importance of delirium: economic and societal costs. *J Am Geriatr Soc* 2011;59 Suppl 2:S241-3. doi: 10.1111/j.1532-5415.2011.03671.x [published Online First: 2011/12/07]
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* 2016 doi: 10.1016/j.jagp.2016.05.016 [published Online First: 2016/09/14]
7. Pollard C, Fitzgerald M, Ford K. Delirium: The lived experience of older people who are delirious post-orthopaedic surgery. *International Journal of Mental Health Nursing* 2015;24(3):213-21. doi: 10.1111/inm.12132
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 surgery* 2015;150(12):1134-40. doi: 10.1001/jamasurg.2015.2606
9. Robinson TN, Raeburn CD, Tran ZV, et al. Postoperative delirium in the elderly: risk factors and outcomes. *Ann Surg* 2009;249(1):173-8. doi: 10.1097/SLA.0b013e31818e4776 [published Online First: 2008/12/25]
10. Adams ST, Leveson SH. Clinical prediction rules. *BMJ (Clinical research ed)* 2012;344
11. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of internal medicine* 2006;144(3):201-09.
12. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart (British Cardiac Society)* 2012;98(9):683-90. doi: 10.1136/heartjnl-2011-301246
13. Debray TPA, Moons KGM, Ahmed I, et al. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in medicine* 2013;32(18):3158-80. doi: 10.1002/sim.5732

14. Hayhurst CJ, Pandharipande PP, Hughes CG. Intensive Care Unit DeliriumA Review of Diagnosis, Prevention, and Treatment. *Anesthesiology* 2016;125(6):1229-41. doi: 10.1097/ALN.0000000000001378
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(10):e1001744. doi: 10.1371/journal.pmed.1001744 [published Online First: 2014/10/15]
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(5):336-41. doi: 10.1016/j.ijsu.2010.02.007 [published Online First: 2010/02/23]
17. Nations U. Ageing 2016 [Available from: http://www.un.org/en/sections/issues-depth/ageing/index.html accessed March 7th 2017.
18. Centers for Disease C. Elder Abuse: Definitions 2016 [Available from: https://www.cdc.gov/violenceprevention/elderabuse/definitions.html accessed March 7th 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(1):263-306. doi: 10.1097/CCM.0b013e3182783b72 [published Online First: 2012/12/28]
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(11):1047-56. doi: 10.1001/jama.2017.11468 [published Online First: 2017/10/05]
21. Well GA, Shea, B., O'Connell, J.P., Welch, V., Losos, M., Tugwell, P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses Ottawa: Ottawa Hospital; 2014 [Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp accessed 05-31 2016.
22. Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating: New York : Springer, [2009] ©2009 2009.
23. Carrasco MP, Villarroel L, Andrade M, et al. Development and validation of a delirium predictive score in older people. *Age Ageing* 2014;43(3):346-51. doi: 10.1093/ageing/aft141 [published Online First: 2013/09/26]
24. de Wit HAJM, Winkens B, Mestres Gonzalvo C, et al. The development of an automated ward independent delirium risk prediction model. *International journal of clinical pharmacy* 2016 doi: 10.1007/s11096-016-0312-7
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(9):493-9. doi: 10.1002/jhm.2062 [published Online First: 2013/08/08]
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(2):175-8. [published Online First: 1995/02/01]
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. *Canadian geriatrics journal : CGJ* 2015;18(4):212-16. doi: 10.5770/cgj.18.185
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 And Ageing* 2005;34(2):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(11):852-7. [published Online First: 1996/03/20]

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(6):474-81. [published Online First: 1993/09/15]
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(13):1406-13. doi: 10.1001/archinte.167.13.1406 [published Online First: 2007/07/11]
32. Isfandiaty 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(4):290-7. [published Online First: 2013/01/15]
33. Pendlebury ST, Lovett N, Smith SC, et al. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of externally derived risk scores. *Age and ageing* 2016;45(1):60-65. doi: 10.1093/ageing/afv177
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(5):817-22. doi: 10.1111/j.1532-5415.2006.00704.x [published Online First: 2006/05/16]
35. Kim MY, Park UJ, Kim HT, et al. DELirium Prediction Based on Hospital Information (Delphi) in General Surgery Patients. *Medicine* 2016;95(12):e3072. doi: 10.1097/MD.0000000000003072
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* 2014 doi: 10.1097/sla.0000000000000742 [published Online First: 2014/06/03]
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(10):946-56. doi: 10.1016/j.jagp.2013.01.069 [published Online First: 2013/05/11]
38. Liang C-K, Chu C-L, Chou M-Y, 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 research* 2015;18(4):347-55. doi: 10.1089/rej.2014.1645
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. *Geriatrics & gerontology international* 2015 doi: 10.1111/ggi.12587
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(5) doi: 10.1136/bmjopen-2012-001599 [published Online First: 2012/09/18]
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(2):153-9. doi: 10.1016/j.genhosppsych.2011.11.011 [published Online First: 2012/01/10]
42. O'Keeffe ST, Lavan JN. Predicting delirium in elderly patients: development and validation of a risk-stratification model. *Age Ageing* 1996;25(4):317-21. [published Online First: 1996/07/01]
43. Pompei P, Foreman M, Rudberg MA, et al. Delirium in hospitalized older persons: outcomes and predictors. *J Am Geriatr Soc* 1994;42(8):809-15. [published Online First: 1994/08/01]
44. Rudolph JL, Doherty K, Kelly B, et al. Validation of a Delirium Risk Assessment Using Electronic Medical Record Information. *Journal of the American Medical Directors Association* 2016;17(3):244-48 5p. doi: 10.1016/j.jamda.2015.10.020
45. Rudolph JL, Harrington MB, Lucatorto MA, et al. Validation of a medical record-based delirium risk assessment. *Journal of the American Geriatrics Society* 2011;59 Suppl 2:S289-94. doi: 10.1111/j.1532-5415.2011.03677.x

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(2):229-36 8p. doi: 10.1161/CIRCULATIONAHA.108.795260
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(11):e149-e53. doi: 10.1111/jgs.14428 [published Online First: 2016/09/22]
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* 2016 doi: 10.1093/ageing/afw198 [published Online First: 2016/11/07]
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(4):387-8. doi: 10.1093/ageing/afi099 [published Online First: 2005/06/16]
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. doi: 10.1136/bmj.g7594 [published Online First: 2015/01/09]
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(5):303-5. doi: 10.1097/pap.0000000000000072 [published Online First: 2015/08/12]
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(1):W1-73. doi: 10.7326/m14-0698 [published Online First: 2015/01/07]
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(8):971-80. doi: 10.1093/aje/kwq223 [published Online First: 2010/09/03]
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. *Statistical Methods in Medical Research*;0(0):0962280214558972. doi: doi:10.1177/0962280214558972
55. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005 [published Online First: 2015/05/20]
56. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35(29):1925-31. doi: 10.1093/eurheartj/ehu207 [published Online First: 2014/06/06]
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(1):128-38. doi: 10.1097/EDE.0b013e3181c30fb2 [published Online First: 2009/12/17]
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(12):1373-9. [published Online First: 1996/12/01]
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. doi: 10.1186/s13104-016-2028-0 [published Online First: 2016/04/16]
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. doi: 10.1136/bmj.h3868 [published Online First: 2015/08/13]

61. Grobman WA, Stamilio DM. Methods of clinical prediction. *Am J Obstet Gynecol* 2006;194(3):888-94. doi: 10.1016/j.ajog.2005.09.002 [published Online First: 2006/03/09]
62. Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? *Contemp Clin Trials* 2013;36(2):636-41. doi: 10.1016/j.cct.2013.06.011 [published Online First: 2013/07/03]
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. doi: 10.1016/j.jclinepi.2016.02.031 [published Online First: 2016/03/12]
64. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., et al. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001;21(1):45-56. doi: 10.1177/0272989x0102100106 [published Online First: 2001/02/24]
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. doi: 10.1016/j.jbi.2015.05.001 [published Online First: 2015/05/20]
66. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;12(4):e0174944. doi: 10.1371/journal.pone.0174944 [published Online First: 2017/04/05]
67. Jones RN, Marcantonio ER, Saczynski JS, et al. Preoperative Cognitive Performance Dominates Risk for Delirium Among Older Adults. *J Geriatr Psychiatry Neurol* 2016 doi: 10.1177/0891988716666380 [published Online First: 2016/09/21]
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(10):1495-504. doi: 10.1016/j.psyneuen.2011.04.003 [published Online First: 2011/05/24]
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. doi: 10.1186/1471-2288-11-103 [published Online First: 2011/07/14]
70. Moore RC, Eyler 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(6):622-32. doi: 10.1016/j.jagp.2014.08.004 [published Online First: 2014/09/14]
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. doi: 10.1016/j.bbi.2016.02.008 [published Online First: 2016/02/14]
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(6):932-39. doi: 10.1093/ageing/afx126 [published Online First: 2017/10/07]
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(7):879-905. doi: 10.1177/0164027516637410 [published Online First: 2016/03/20]
74. Yates JA, Clare L, Woods RT. What is the Relationship between Health, Mood, and Mild Cognitive Impairment? *J Alzheimers Dis* 2017;55(3):1183-93. doi: 10.3233/jad-160611 [published Online First: 2016/10/30]
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. *Dementia & Geriatric Cognitive Disorders* 2014;38(1/2):65-78 14p. doi: 10.1159/000357454
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 Clinical And Experimental Research* 2007;19(3):172-77.

77. Fong TG, Hshieh TT, Wong B, et al. Neuropsychological profiles of an elderly cohort undergoing elective surgery and the relationship between cognitive performance and delirium. *Journal of the American Geriatrics Society* 2015;63(5):977-82. doi: 10.1111/jgs.13383
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(12):2383-90. doi: 10.1111/jgs.13138

10.1111/jgs.13138.</p>
79. Newman MW, O'Dwyer LC, Rosenthal L. Predicting delirium: A review of risk-stratification models. *General Hospital Psychiatry* 2015;37(5):408-13. doi: 10.1016/j.genhosppsych.2015.05.003
80. Campaner CP. Valoración del riesgo de delirium en pacientes mayores hospitalizados. = Delirium risk assessment in elderly hospitalized patients. *Revista Española de Geriatría y Gerontología* 2010;45(5):285-90. doi: 10.1016/j.regg.2010.03.011
81. Debray TP, Koffijberg H, Nieboer D, et al. Meta-analysis and aggregation of multiple published prediction models. *Stat Med* 2014;33(14):2341-62. doi: 10.1002/sim.6080 [published Online First: 2014/04/23]
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(5):588-94. doi: 10.3348/kjr.2011.12.5.588 [published Online First: 2011/09/20]

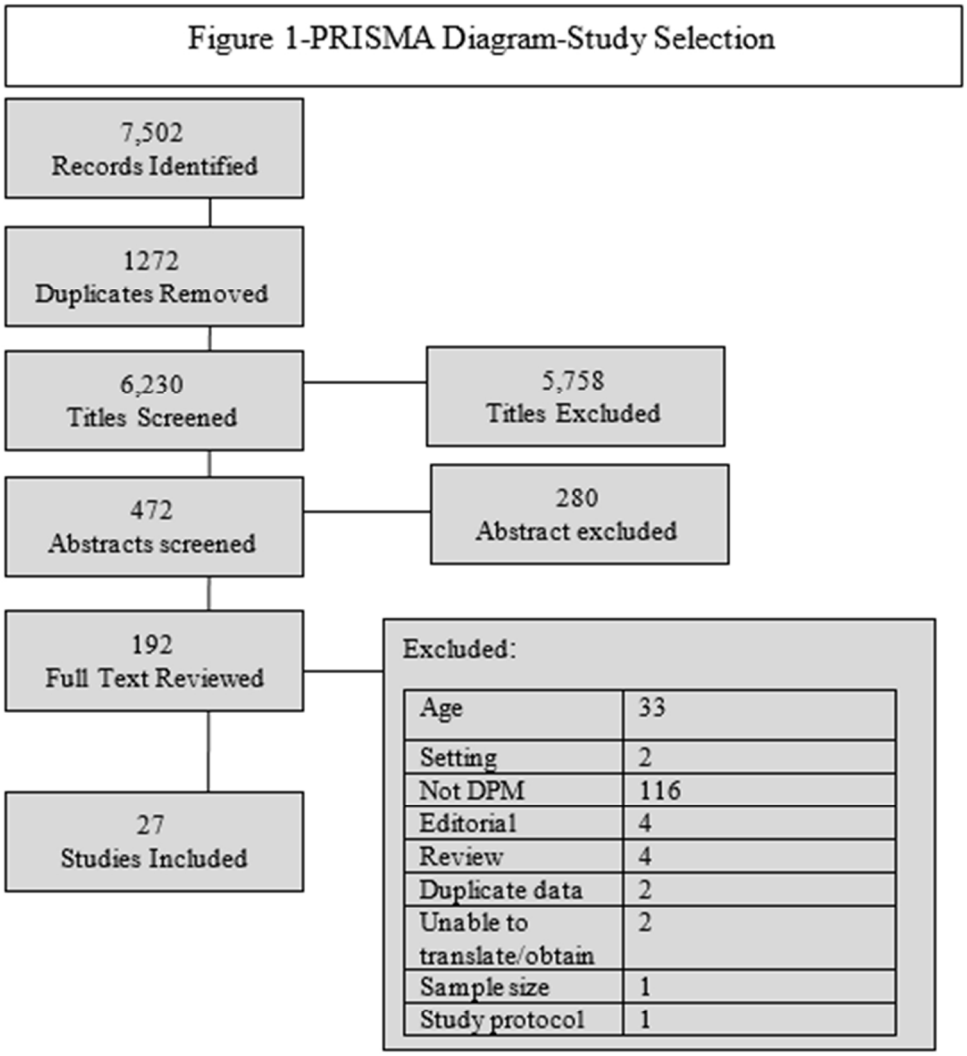


Figure 1: PRISMA Diagram - Study Selection

128x145mm (300 x 300 DPI)

Figure 2: CHARMS Risk of Bias Assessment

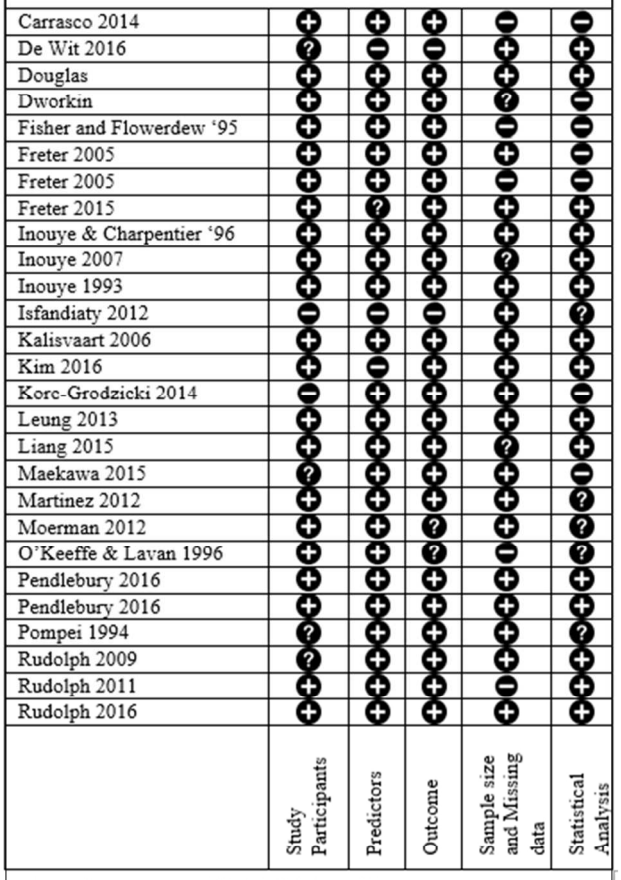


Figure 2 displays the CHARMS Risk of Bias assessment on all included studies.
Study Participants: design of included study, sampling method, inclusion/exclusion criteria
Predictors: definition, timing, and measurement
Outcome: definition, timing, and measurement
Sample Size and Missing Data: sample size, events per variable, missing data
Statistical Analysis: selection of predictors, internal validation, type of external validation done

Figure 2: CHARMS Risk of Bias Assessment

110x202mm (300 x 300 DPI)

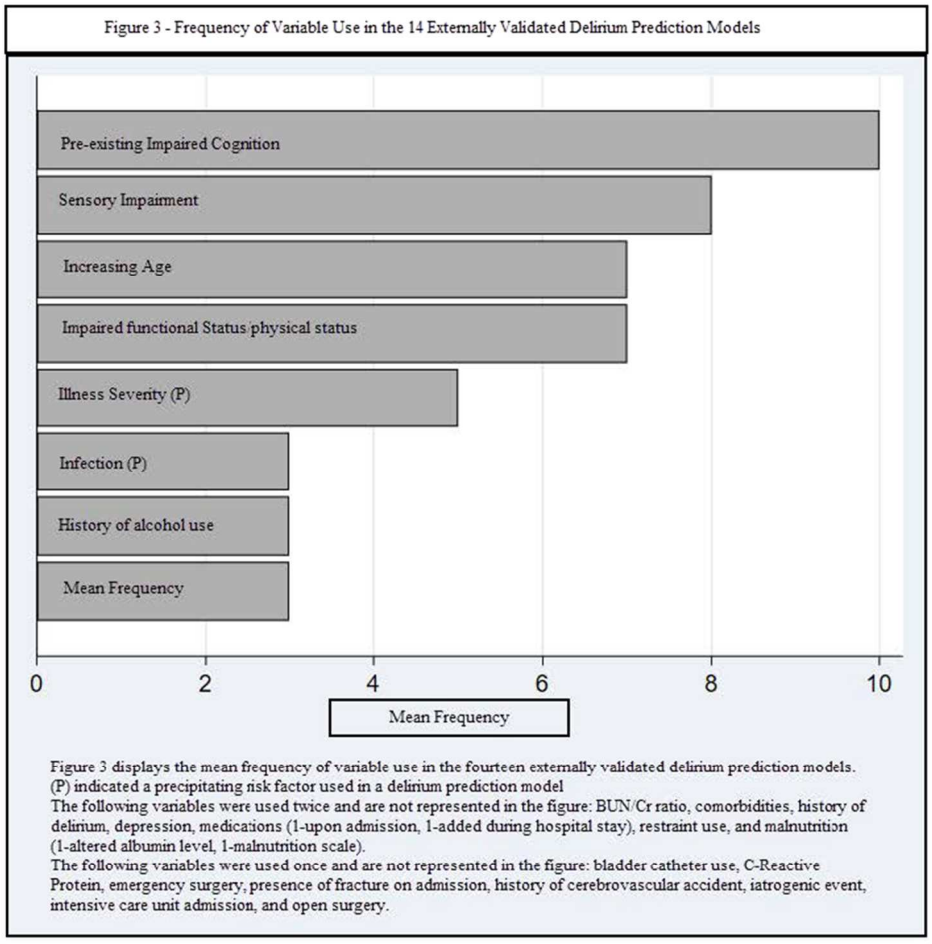


Figure 3: Frequency of Variable Use in the 14 Externally Validated Delirium Prediction Models

187x183mm (300 x 300 DPI)

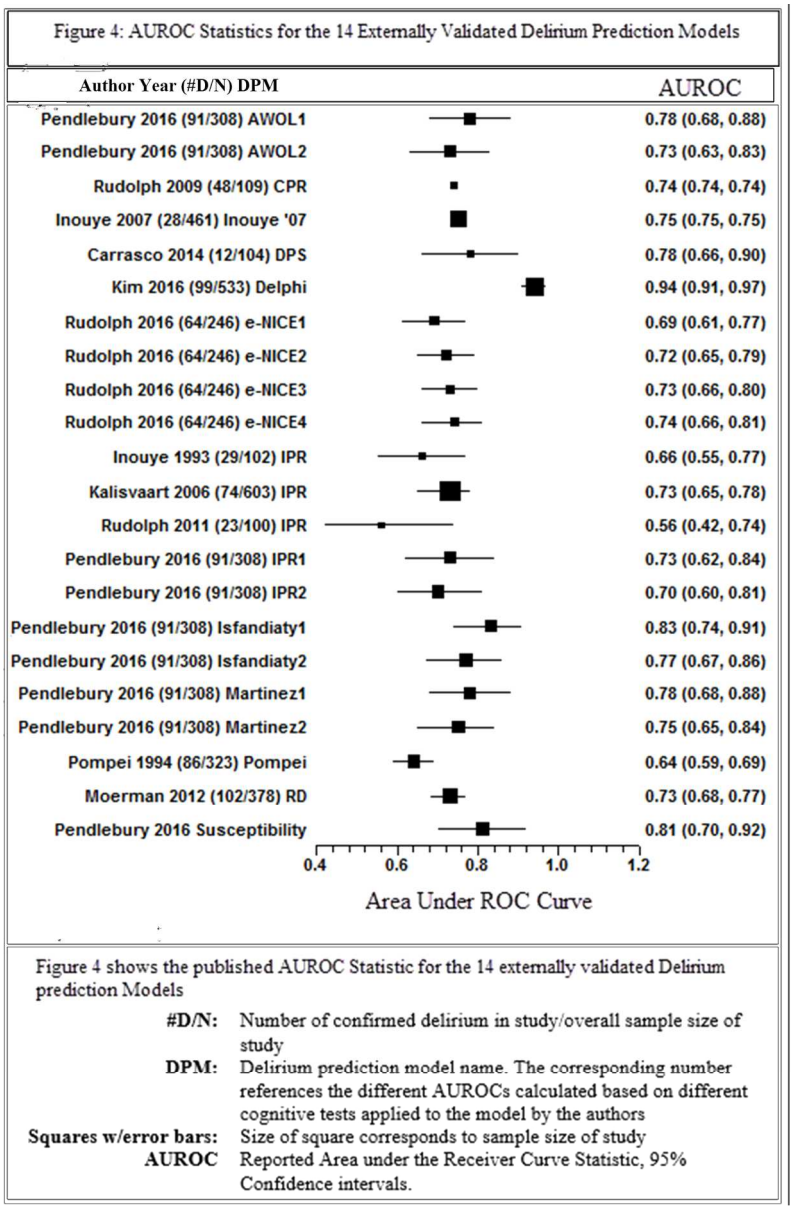


Figure 4: AUROC Statistics for the 14 Externally Validated Delirium Prediction Models

137x210mm (300 x 300 DPI)

Appendix A – Review Protocol

Working Title of Review	Systematic Review of Delirium Prediction Models			Support	Modifications
Authors	1st & Corresponding	Heidi Lindroth	Literature search, data extraction, data synthesis and manuscript preparation.		
	Data Extraction	Heidi Lindroth Suzanne Purvis	Literature search, data extraction, data synthesis.		
	Content Experts	Lisa Bratzke	Assisted with content related to cognition. Results review.		
		Roger Brown	Statistical content expert		
		Mark Coburn	Results review, Manuscript preparation		
		Marko Mrkobrada	Results review, Manuscript preparation		
		Matthew TV Chan	Results review, Manuscript preparation		
		Daniel Davis	Geriatrician expertise, reviewed results, manuscript preparation.		
		Pratik Pandharipande	Results review, Manuscript preparation		
		Cynthia M. Carlsson	Geriatrician expertise, reviewed results, manuscript preparation.		
	Mentoring	Robert D. Sanders	Mentoring author, resolved content/data disagreements b/w authors, manuscript preparation.		
Aim	To identify existing prognostic delirium prediction models and evaluate their validity and statistical methodology in the older adult (≥60yo) acute hospital population.				
Search Terms	(“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*)			UW-Madison Health Sciences librarian. Three meetings to refine search terms.	
Databases searched	PubMed, CINAHL, PsychINFO, Cochrane, SocINDEX and Medline			Health Sciences librarian.	Expanded to include SocINDEX
Timelines established	01/01/1990-12/31/2016				Originally was 12/31/15.

			Expanded to include all of 2016.
Inclusion criteria	• Age ≥ 60 • Inpatient population • Developing and/or validating a delirium prediction model		Age expanded from ≥ 70 years of age due to the literature
Exclusion criteria	• Emergency department • Hospice/palliative care • Pediatric population • Related to alcohol withdrawal • ≤50 sample size	Mentoring author	Sample size criteria added to build rigor in the studies that were included in the sys review
Selection process	Studies will be selected based on the inclusion/exclusion criteria. The data extraction authors (HL and SP) will conduct the literature search independently and meet monthly to discuss findings. Any disagreements will be resolved by the mentoring author (RDS)		
Data Management	A shared folder on the UW-Madison Box account will be created to share documents, data and meeting information.		
Data collection process	Data will be collected independently by HL and SP then data points will be shared at monthly meetings. Data collection tables will be created using Microsoft Excel then uploaded to the shared Box account. Any disagreement between authors will be resolved by the mentoring author (RDS).		
Data points collected	• Characteristics of studies (design, population, sample size) • Outcome measure including how it was identified, measured, defined. Prevalence. • Statistical methods applied • Statistical information about the delirium prediction models (sensitivity, specificity, positive predictive value, negative predictive value, AUROC) • Characteristics of DPMs (variables used, scoring, development) • Cognitive measures used in studies. • Criteria to fulfill the Newcastle Ottawa Scale.		
Outcomes	• AUROC will be the primary outcome measure • Characteristics of DPMs (variables, statistics) • Cognitive tests used		

Data synthesis	The first/corresponding author (HL) will synthesize the data into the manuscript. The co-authors will verify this. RB will complete the meta-analysis.		
Manuscript preparation	HL will complete manuscript preparation. All co-authors are responsible for reviewing content and data to assure correctness and complete synthesis of data gathered.		

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	4
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4-5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5-6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6-7
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	7-8Table1, Figure 2
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Table 1,2 Figure 4
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Table 1, Figure 2
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8-19
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	19-20
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	19-24
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	19-26
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	6

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

PRISMA 2009 Checklist