National Institute for Health and Care Excellence

Final

Delirium: prevention, diagnosis and management

[A] Evidence review for diagnostic accuracy of tests to identify delirium

NICE guideline CG103

Evidence review underpinning recommendations 1.5.2 and 1.6.1 to 1.6.2 and research recommendations in the NICE guideline

January 2023

Final

National Institute for Health and Care Excellence

NICE accredited
www.nice.org.uk/accreditation

Disclaimer

The recommendations in this guideline represent the view of NICE, arrived at after careful consideration of the evidence available. When exercising their judgement, professionals are expected to take this guideline fully into account, alongside the individual needs, preferences and values of their patients or service users. The recommendations in this guideline are not mandatory and the guideline does not override the responsibility of healthcare professionals to make decisions appropriate to the circumstances of the individual patient, in consultation with the patient and/or their carer or guardian.

Local commissioners and/or providers have a responsibility to enable the guideline to be applied when individual health professionals and their patients or service users wish to use it. They should do so in the context of local and national priorities for funding and developing services, and in light of their duties to have due regard to the need to eliminate unlawful discrimination, to advance equality of opportunity and to reduce health inequalities. Nothing in this guideline should be interpreted in a way that would be inconsistent with compliance with those duties.

NICE guidelines cover health and care in England. Decisions on how they apply in other UK countries are made by ministers in the Welsh Government, Scottish Government, and Northern Ireland Executive. All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN: 978-1-4731-4955-7

Contents

1 Delirium: prevention, diagnosis and management

1.1 Review question

What is the diagnostic accuracy of diagnostic tests compared with the reference standard, to identify delirium in people in hospital and long-term residential care settings?

1.1.1 Introduction

Delirium (sometimes called 'acute confusional state') is a common clinical syndrome characterised by disturbed consciousness, cognitive function or perception, which has an acute onset and fluctuating course. It usually develops over 1–2 days. It is a serious condition that is associated with poor outcomes. However, it can be prevented and treated if dealt with urgently. Reporting of delirium is poor in the UK, indicating that awareness and reporting procedures need to be improved.

NICE is updating the guideline on delirium: prevention, diagnosis and management (CG103). The guideline was originally published in July 2010 and last updated in March 2019. It was developed as set out in the original scope (2008).

New evidence about diagnostic tests for delirium suggests that recommendations on diagnosis (specialist clinical assessment) may need updating. Full details are set out in the 2020 exceptional surveillance review decision.

1.1.2 Summary of the protocol

Table 1: Summary of protocol

Population	Adults (18 years and older) in: • hospital, including surgical, medical, ICU, and accident and emergency departments • long-term residential care settings Exclusion: • People receiving end-of-life care (within the last few days of life) • People with intoxication and/or withdrawing from drugs or alcohol, and people with delirium associated with these states.
Index Tests	Index tests, including the people operating them, subdivided by setting: a. 4AT b. Confusion Assessment Method Instrument (CAM) c. 3D CAM d. Delirium Observation Screening Scale (DOS) e. Single Question to Identify Delirium (SQID) f. Recognizing acute delirium as part of your routine (RADAR) g. Intensive Care Delirium Screening Checklist (ICD-SC) h. Confusion Assessment Method - Intensive Care Unit (CAM-ICU)

	i. Brief Confusion Assessment Method (B-CAM) j. Ultra-Brief Confusion Assessment Method (UB2-CAM) k. Nursing Delirium Screening Scale (NuDESC)
Comparator (reference standards)	DSM-IV/5 or ICD-10/11, applied by a trained specialist.
Outcome measures	Primary outcomes • Sensitivity and specificity • Likelihood ratios Secondary outcomes • Positive and negative predictive values if these are reported by SRs • ROC/AUC, c-statistic • Ease of use (for example, time taken, range of staff who can use)
Study types	A 2-stage approach to addressing this question will be taken: 1. A narrative review of systematic reviews for the index tests of interest 2. For tests where no SRs are identified, primary cross-sectional studies will be searched. Review of reviews • Systematic reviews of diagnostic studies. Review of primary studies • Cross-sectional studies

1.1.3 Methods and process

This evidence review was developed using the methods and process described in Developing NICE guidelines: the manual. Methods specific to this review question are described in the review protocol in appendix A and the methods appendix (appendix L)

Declarations of interest were recorded according to NICE's conflicts of interest policy.

1.1.3.1 Searches

The searches for the review-of-reviews of diagnostic test accuracy were run on the 19th July 2022. The following databases were searched: Epistemonikos; Cochrane Database of Systematic Reviews (Wiley); Medline ALL (Ovid); Embase (Ovid); PsycInfo (Ovid). Full search strategies for each database are provided in Appendix B.

The searches for cost-effectiveness evidence were run on the 20th July 2022. The following databases were searched: Medline ALL (Ovid); Embase (Ovid); PsycInfo (Ovid); Econlit (Ovid); INAHTA International HTA Database. Full search strategies for each database are provided in Appendix B.

A NICE information specialist conducted the searches. The MEDLINE strategy was quality assured by a trained NICE information specialist and all translated search strategies were peer reviewed to ensure their accuracy. Both procedures were adapted from the 2016 PRESS Checklist.

1.1.3.2 Protocol deviation

Some systematic reviews included primary studies that used other reference standards than those outlined in the protocol (for example CAM as a reference standard) in addition to primary studies that used the specified reference standards. Since it was not possible to disambiguate these, they were included.

1.1.3.3 Strategy for data synthesis

Whole systematic reviews were used. Data for individual studies were not extracted from the reviews.

Data from systematic reviews covered all of the index tests of interest and therefore no additional searches were undertaken to identify primary studies.

The results of all systematic reviews identified in the past 6 years were reported narratively by test and setting. The narrative reported the main outcomes of the systematic review alongside any assessment of confidence in the outcome (for example GRADE). Where GRADE was not reported, the RoB of included studies was reported instead. Matrices were constructed to show the spread of primary studies across systematic reviews for the same test to enable the committee to take into account the duplication of primary studies in several systematic reviews.

1.1.4 Diagnostic evidence

1.1.4.1 Included studies

In total 480 references were identified through systematic searches after duplicates were removed. Based on title and abstract, 20 of these references were considered relevant to the protocol and were ordered for full-text review.

Of the 20 references that progressed to full-text review, 17 references were included and 3 were excluded. All 17 included references were systematic reviews. See Appendix C for a flowchart for the search and study selection process (following the PRISMA guidelines). See Appendix D for the full evidence tables of included studies.

1.1.4.2 Excluded studies

See Appendix J for the full list of excluded studies along with reasons for their exclusion.

1.1.5 Summary of studies included in the diagnostic evidence

Table 2:Summary of included studies

Details of abbreviations can be found at the bottom of the table.

Author and year [countries of included studies]	Index test (number of studies)	Comparator	Setting	Patient population	HCP delivering index test	Outcomes	Risk of bias (ROBIS)
Aldwikat 2022 [Germany, USA]	4AT (1) CAM (1) 3D-CAM (1) CAM-ICU (1) NuDESC (3)	DSM-IV DSM-5	Hospitals (post-anaesthetic care units)	Adults 18 years and over who were admitted to the PACU following surgery, including those with any pre-existing conditions	Research assistants	Sensitivity Specificity	High
Brefka 2021 [NR]	4AT (1) CAM 3D-CAM (1) DOS (2) SQiD (2) RADAR (1) ICDSC (2) CAM-ICU (1) B-CAM (1) UB2-CAM (3) NuDESC (2)	DSM-IV 3D-CAM CAM Psychiatrist interview / diagnosis Geriatric psychiatrist rating after comprehensive assessment DSM-IV-TR	Hospitals (focus on acute care and emergency settings)	Older patients	Physicians Nurses Trained lay-raters Clinical staff / clinicians Untrained geriatricians	Sensitivity Specificity AUC	High
Calf 2021 [Germany, Brazil, Canada, USA,	4AT (2) CAM (2) SqiD (1)	DSM-5 DSM-IV DSM-IV-TR	Hospitals (emergency departments)	Patients with a mean or median age 65	NR	Sensitivity Specificity	Low

Author and year [countries of included studies]	Index test (number of studies)	Comparator	Setting	Patient population	HCP delivering index test	Outcomes	Risk of bias (ROBIS)
The Netherlands, UK]	CAM-ICU (2) B-CAM (2)	CAM		years or older, visiting an emergency department			
Chen 2021 [NR]	ICDSC (12) CAM-ICU (29)	DSM-IV DSM-IV-TR DSM-5	Hospitals (intensive care units)	Adult patients (aged≥18 years) who were admitted to an ICU	NR	Sensitivity Specificity	Low
Ho 2020 [NR]	ICDSC (8) CAM-ICU (23)	DSM-IV DSM-5 DSM-IV-TR Clinical diagnosis confirmed by a psychiatrist DSM-III-R	Hospitals (ICUs)	Patients 18 years and older	Nurses Doctors Independent investigators Intensivists Physician / nurse investigators Bachelor's-level psychologists Examiners	Sensitivity Specificity AUC	High
Ho 2022 [Portugal, Turkey, China, Sweden, Germany, USA]	NuDESC (11)	DSM-5 DSM-IV DSM-IV-TR ICDSC CAM-ICU	Hospitals	Adult (age ≥ 18 years) postoperative patients who received any type of surgery and any method of anaesthesia	Nurses Researchers / research assistants Physicians Psychiatrists	Sensitivity Specificity PLR NLR AUC	Low
Jeong 2020a [Iran, Norway, Australia, UK, Thailand, Italy,	4AT (11)	DSM-5 DSM-IV CAM	Hospitals Nursing homes and daily care centres	NR	NR	Sensitivity Specificity AUC	Low

Author and year [countries of included studies]	Index test (number of studies)	Comparator	Setting	Patient population	HCP delivering index test	Outcomes	Risk of bias (ROBIS)
Canada, Ireland, Germany]		CAM-ICU 3D-CAM					
Jeong 2020b [USA, Germany, Sweden, Hong Kong, Canada]	NuDESC (11)	DSM-5 DSM-IV CAM	Hospitals	NR	NR	Sensitivity Specificity AUC	High
Kim 2021 [Germany, The Netherlands, Sweden, USA, China, Thailand]	4AT (1) CAM (3) DOS (1) CAM-ICU (2) NuDESC (6)	DSM-IV	Hospitals (post-surgery)	Participants aged 20 years or older, who underwent general anaesthesia surgery	Nurses Psychiatrists Trainees (occupation not specified)	Sensitivity Specificity PLR NLR	High
Mansutti 2019 [Italy, Russia, UK, Czech Republic]	4AT (3) CAM-ICU (1)	DSM-IV CAM	Hospitals (stroke units / neurovascular departments)	Patients with acute stroke	Neurologists Trained medical students / junior physicians A panel of specialists, experts on delirium (two neurologists, two neuropsychologists, a psychiatrist and a speech therapist)	Sensitivity Specificity PPV NPV AUC PLR	Low
Park 2021 [The Netherlands, Switzerland, USA, Denmark, Belgium]	DOS (8)	DSM-IV DRS-R-98	Hospitals Home hospice	NR	NR	Sensitivity Specificity AUC	Low
Patel 2018 [NR]	4AT (1) ICDSC (1)	DSM-IV CAM	Hospitals (ICU)	Neuro-critically ill patients of any age	NR	Sensitivity Specificity	Low

Author and year [countries of included studies]	Index test (number of studies)	Comparator	Setting	Patient population	HCP delivering index test	Outcomes	Risk of bias (ROBIS)
	CAM-ICU (2)					PPV NPV	
Quispel-Aggenbach 2018 [NR]	RADAR (4)	DSM-IV-TR (with CAM) DSM-5 DSM-5 (with CAM) CAM	Acute care hospital and nursing homes	Patients aged 60 years or older	Nurses Nurse assistants Research assistants	Sensitivity Specificity	Low
Rosgen 2018 [Australia]	SqiD (1)	DSM-IV	Hospitals	Adult patients (≥ 18 years old) in any hospital setting	NR	Sensitivity Specificity PPV NPV	Low
Tieges 2021 [USA, Thailand, Russia, UK, Norway, Ireland, Germany, Iran, Italy, Canada, Australia]	4AT (17)	DSM-5 Chart review by 2-3 physicians CAM DSM-IV-TR	Hospitals Nursing homes and daily care centres	Participants aged ≥65	Nurse Psychiatrist Researcher Neurologist Medical student	Sensitivity Specificity	Low
Van Velthuijsen 2016 [The Netherlands, Italy, Sweden, Hong Kong, Germany, USA, Australia, Brazil, Spain, Finland, Portugal, Canada, Ireland, Thailand, Czech Republic]	4AT (1) CAM (11) 3D-CAM (1) DOS (2) SqiD (1) ICDSC (1) CAM-ICU (6) B-CAM (1) NuDESC (4)	DSM-III DSM-III-R DSM-IV DSM-IV-TR ICD-10	Hospitals	Older hospitalised patients (mean or median age 65+)	Nurses Informal carers Doctors Psychiatrists Psychologists Researchers / research assistants	Sensitivity Specificity	Low
Watt 2021 [USA, Belgium, The	CAM (1) DOS (3) SqiD (1)	DSM-IV DRS-R-98 CAM	Inpatient oncology Community hospice	Adults (18+ years) in palliative care	NR	Sensitivity Specificity	High

Author and year [countries of included studies]	Index test (number of studies)	Comparator	Setting	Patient population	HCP delivering index test	Outcomes	Risk of bias (ROBIS)
Netherlands, Ireland, Australia]	B-CAM (1) NuDESC (1)	DSM-5 MDAS	Palliative care units (including hospital and hospice units)				

Index tests – 3D-CAM: 3-Minute Diagnostic Interview for Confusion Assessment Method; 4AT: 4 'A's Test; B-CAM: Brief Confusion Assessment Method; CAM: Confusion Assessment Method; Brief Confusion Assessment Method for the Intensive Care Unit; DOS: Delirium Observation Screening Scale; ICDSC: Intensive Care Delirium Screening Checklist; NuDESC: Nursing Delirium Screening Scale; RADAR: Routine or Recognising Acute Delirium As part of your Routine; SqiD: Single Question to Identify Delirium; UB2-CAM: Ultra Brief Confusion Assessment Method

Reference standards – CAM: Confusion Assessment Method; DRS-R-98: Delirium Rating Scale Revised 98; DSM-III: Diagnostic and Statistical Manual of Mental Disorders, Third Edition; DSM-III-R: Diagnostic and Statistical Manual of Mental Disorders, Third Edition Revised; DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision; DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; ICD-10: International Classification of Diseases 10th Revision; MDAS: Memorial Delirium Assessment Scale

Setting – ICU: Intensive Care Unit

General: NR: Not Reported

See appendix D for full evidence tables.

1.1.6 Summary of the diagnostic evidence

Not all of the included systematic reviews undertook meta-analysis to generate pooled estimates of diagnostic accuracy.

For each index test, the median sensitivity and specificity values and interquartile ranges (IQR) were calculated from reviews that reported pooled estimates (columns 2-4 in table 3). Following this, the median sensitivity and specificity values and IQRs were also calculated for the remaining individual studies that were in reviews that did not pool results (columns 5-7 in table 3). These results are summarised in Table 3.

Table 4 summarises all diagnostic accuracy data for each test from systematic reviews that reported pooled estimates.

Table 5 summarises all unpooled diagnostic accuracy data from primary studies that were included in systematic reviews that did not undertake meta-analysis to produce a pooled estimate. This also includes primary studies that were already captured within systematic reviews that reported pooled results.

Table 3 Summary of median sensitivities and specificities from pooled results and results from single studies from within the included systematic reviews

This table summarises the median sensitivity and specificity values and interquartile ranges (IQR) for tests reported in systematic reviews that pooled their results and for single studies that were only captured by systematic reviews that did not report pooled results.

	From SRs that presented pooled results			From SRs that reported individual studies		
Index test	**Number of systematic reviews (primary studies) contributing to result**	**Median Sensitivity (IQR)**	**Median Specificity (IQR)**	**Number of primary studies contributing to result**	**Median Sensitivity (IQR)**	**Mean Specificity (SD)**
4AT	4 (19)	87% (3.25%)	88% (0.5%)	N/A	N/A	N/A
CAM	1 (4)	47% (NA)	99% (NA)	13	79% (14%)	98% (7%)
3D-CAM	N/A	N/A	N/A	2	98% (2.5%)	91% (3%)
DOS	1 (7)	90% (NA)	92% (NA)	1	100% (NA)	97% (NA)
SQiD	N/A	N/A	N/A	3	77% (9%)	71% (14%)
RADAR	N/A	N/A	N/A	4	100% (6.75%)	71% (4.75%)
ICDSC	3 (13)	87% (8%)	87% (13.5%)	2	67% (2.5%)	90% (10.5%)
CAM-ICU	3 (33)	85% (0.5%)	95% (1.5%)	3	45% (17%)	89% (12%)
B-CAM	N/A	N/A	N/A	3	80% (8%)	94% (4.75%)
UB2-CAM	1 (3)	93% (0%)	95% (0%)	N/A	N/A	N/A
NuDESC	5 (17)	71% (9.05%)	91% (4.2%)	1	63% (NA)	67% (NA)

Index tests – 3D-CAM: 3-Minute Diagnostic Interview for Confusion Assessment Method; 4AT: 4 ‘A’s Test; B-CAM: Brief Confusion Assessment Method; CAM: Confusion Assessment Method; Brief Confusion Assessment Method for the Intensive Care Unit; DOS: Delirium Observation Screening Scale; ICDSC: Intensive Care Delirium Screening Checklist; NuDESC: Nursing Delirium Screening Scale; RADAR: Routine or Recognising Acute Delirium As part of your Routine; SQiD: Single Question to Identify Delirium; UB2-CAM: Ultra Brief Confusion Assessment Method

Footnotes

CAM pooled results: Kim 2021 reported outcomes under CAM (and variants), which included data from studies on CAM-ICU.

SQiD results from single studies: Han 2018 (captured in Calf 2021) reported outcomes for SQiD to patient and SQiD to surrogate. The result for SQiD to patient was used in this analysis.

RADAR results from single studies: Voyer 2015 reported results for RADAR used 1–4 x daily or 3–4 x daily in Quispel-Aggenbach 2018, but only the 3–4 x daily result was reported from the same study in Brefka 2022. Therefore, the 3–4 x daily result was used in this analysis. Bilodeau 2016 (captured in Quispel-Aggenbach 2018) reported outcomes for dementia patients only.

B-CAM results from single studies: Calf 2021 reported a different result for Han 2013 compared to Brefka 2022 and van Velthuijsen 2016. The result from Brefka 2022 and van Velthuijsen 2016 was used in this analysis.

NuDESC pooled results: Kim 2021 reported separate outcomes for a cut-off score of ≥2 and ≥1. The result for a cut-off score of ≥2 was used to maintain consistency with the other reviews.

NuDESC results from single studies: De la Cruz reported results for NuDESC delivered by a nurse, caregiver in the evening and caregiver at night. The result for NuDESC delivered by a nurse was used as the test was designed for use by nurses.

Table 4: Summary of pooled diagnostic evidence

This table summarises the pooled diagnostic accuracy data from systematic reviews that undertook a meta-analysis of the diagnostic accuracy data of their included studies. For details of which primary studies were included in each review see the narrative section below (section 1.1.6.1).

Test	Review (number of studies included for test)	Pooled Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
4AT	Calf 2021 (2)	Sens: 87% (74-94) Spec: 87% (60-97)	NR	NR	NR	NR	NR
	Jeong 2020a (11)	Sens: 81.5% (70.7-89.0) Spec: 87.5% (79.5-92.7)	NR	NR	0.911 (NR)	>3	NR
	Tieges 2021 (17)	Sens: 88% (80-93) Spec:88% (82-92)	NR	NR	NR	≥4	NR
CAM	Kim 2021* (3)	Sens: 47% (37-56) Spec: 99% (98-99)	NR	LR+: 32.10 (7.01-146.93) LR-: 0.55 (0.34-0.87)	NR	NR	5.27-14 min
DOS	Park 2021 (8)	Sens: 90% (76-97) Spec: 92% (88-94)	NR	NR	0.94 (NR)	≥3	NR
ICDSC	Brefka 2022 (2)	Sens: 99% (NR) Spec: 64% (NR)	NR	NR	NR	≥ 4 = suspected delirium	<5 min
	Chen 2021 (12)	Sens: 83% (74-90) Spec: 87% (78-93)	NR	NR	NR	4 (9 studies) 3 (1 study) NR (2 studies)	2 min

Test	Review (number of studies included for test)	Pooled Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	Ho 2020 (8)	Sens: 87% (70–95) Spec: 91% (85–95)	NR	NR	0.95 (NR)	NR	NR
CAM-ICU	Calf 2021 (2)	Sens: 85% (39-98) Spec: 98% (94-99)	NR	NR	NR	NR	NR
	Chen 2021 (29)	Sens: 84% (77-88) Spec: 95% (91-97)	NR	NR	NR	NR	2-3 min but may be up to 10 mins when users are unfamiliar with the content
	Ho 2020 (23)	Sens: 85% (77-91) Spec: 95% (90-97)	NR	NR	0.96 (NR)	NR	NR
UB2-CAM	Brefka 2022 (3)	Sens: 93% (NR) Spec: 95% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	2 min
NuDESC	Brefka 2022 (3)	Sens: 86% (NR) Spec: 87% (NR)	NR	NR	NR	≥2 = suspected delirium	<2 min
	Ho 2022 (11)	Sens: 73% (44-90) Spec: 93% (87-96)	PPV: 10.2 (6.8–15.2) NPV: 0.29 (0.12–0.69)	NR	0.94 (0.91–0.96)	≥2	2.13 (SD: 0.05)
	Jeong 2020b (11)	Sens: 68.6% (55.3-79.5)	NR	NR	0.882 (NR)	>1	NR

Test	Review (number of studies included for test)	Pooled Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
		Spec: 89.4% (83.3-93.5)					
	Kim 2021 (5)	Sens: 63% (56-69) Spec: 93% (91-94)	PPV: 7.97 (4.38–14.49) NPV: 0.33 (0.16–0.67)	NR	NR	≥2	1.27-13 min
	Kim 2021 (1)	Sens: 69% (60-76) Spec: 94% (92-96)	PPV: 7.76 (2,058–23.32) NPV: 0.38 (0.29–0.48)	NR	NR	≥1	1.27-13 min

Headings – CI: Confidence Interval; LR: Likelihood Ratio; NPV: Negative Predictive Value; PPV: Positive Predictive Value; Sens: Sensitivity; Spec: Specificity

Index tests – 3D-CAM: 3-Minute Diagnostic Interview for Confusion Assessment Method; 4AT: 4 ‘A’s Test; B-CAM: Brief Confusion Assessment Method; CAM: Confusion Assessment Method; Brief Confusion Assessment Method for the Intensive Care Unit; DOS: Delirium Observation Screening Scale; ICDSC: Intensive Care Delirium Screening Checklist; NuDESC: Nursing Delirium Screening Scale; RADAR: Routine or Recognising Acute Delirium As part of your Routine; SQiD: Single Question to Identify Delirium; UB2-CAM: Ultra Brief Confusion Assessment Method

Footnotes

*Kim 2021 reported outcomes under CAM (and variants), which included data from studies on CAM-ICU.

Table 5: Summary of unpooled diagnostic evidence

This table summarises the unpooled diagnostic accuracy data from primary studies that were included in systematic reviews that did not undertake a meta-analysis. For details of the reviews and all of the included primary studies in them, see the narrative section below (section 1.1.6.1).

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
4AT	**Reported in Mansutti 2019**						
	Infante 2017 (at admission)	Sens: 90.2% (NR) Spec: 64.5% (NR)	NR	NR	0.82 (NR)	NR	NR
	Infante 2017 (after 7 days)	Sens: 96.4% (NR) Spec: 76.7% (NR)	NR	NR	0.88 (NR)	NR	NR
	Kutlubaev 2016	Sens: 93% (NR) Spec: 86% (NR)	PPV: 86% (NR) NPV: 85.6% (NR)	NR	NR	NR	NR
	Lees 2013	Sens: 100% (74-100) Spec: 82% (72-89)	NR	NR	NR	NR	NR
	Reported in Aldwikat 2022						
	Saller 2019	Sens: 96% (NR) Spec: 99% (NR)	NR	NR	NR	NR	2 min
	Reported in Brefka 2022						
	Bellelli 2014	Sens: 90% (NR) Spec: 84% (NR)	NR	NR	0.89-0.93 (NR)	≥4 = possible delirium	<5 min
	Reported in Kim 2021						
	Saller 2019	Sens: 95% (77-99)	NR	LR+: 975.91 (60.94-15,614.60)	NR	≥3	NR

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
		Spec: 100% (99-100)		LR-: 0.06 (0.01-0.30)			
	Reported in Patel 2018						
	Lees 2013	Sens: 100% (74-100) Spec: 82% (72-89)	PPV: 43% (NR) NPV: 100% (NR)	NR	NR	NR	NR
	Reported in van Velthuijsen 2016						
	Bellelli 2014	Sens: 90% (NR) Spec: 84% (NR)	NR	NR	NR	4	<2 min
CAM	**Reported in Aldwikat 2022**						
	Radtke 2008	Sens: 43% (NR) Spec: 98% (NR)	NR	NR	NR	NR	20 min
	Reported in Brefka 2022						
	Inouye 1990	Sens: 94-100% (NR) Spec: 90-95% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	5-10 min
	Reported in Calf 2021						
	Fabbri 2001	Sens: 94% (71-100) Spec: 96% (90-99)	NR	NR	NR	NR	NR
	Shenkin 2019	Sens: 40% (26-57) Spec: 100% (98-100)	NR	NR	NR	NR	NR
	Reported in van Velthuijsen 2016						
	Fabbri 2001	Sens: 94% (NR) Spec: 96% (NR)	NR	NR	NR	NR	5min - <15 min

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	González 2004	Sens: 90% (NR) Spec: 100% (NR)	NR	NR	NR	NR	5min - <15 min
	Hestermann 2009	Sens: 77% (NR) Spec: 96-100% (NR)	NR	NR	NR	NR	5min - <15 min
	Laurila 2002	Sens: 80-85% (NR) Spec: 63-84% (NR)	NR	NR	NR	NR	5min - <15 min
	Leung 2008	Sens:76% (NR) Spec: 100% (NR)	NR	NR	NR	NR	5min - <15 min
	Martins 2015	Sens: 79% (NR) Spec: 99% (NR)	NR	NR	NR	NR	5min - <15 min
	Monette 2001	Sens: 64% (NR) Spec: 93% (NR)	NR	NR	NR	NR	5min - <15 min
	Pompei 1995	Sens: 46% (NR) Spec: 92% (NR)	NR	NR	NR	NR	5min - <15 min
	Ryan 2009	Sens: 88% (NR) Spec: 100% (NR)	NR	NR	NR	NR	5min - <15 min
	Thomas 2012	Sens: 74-82% (NR) Spec: 91-100% (NR)	NR	NR	NR	NR	5min - <15 min
	Wongpakaran 2011	Sens: 92% (NR) Spec: 100% (NR)	NR	NR	NR	NR	5min - <15 min
	Reported in Watt 2021						

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	Ryan 2009	Sens: 88% (62-98) Spec: 100% (88-100)	NR	NR	NR	Binary	NR
3D CAM	**Reported in Aldwikat 2022**						
	Olbert 2019	Sens: 100% (NR) Spec: 88% (NR)	NR	NR	NR	NR	3 min
	Reported in Brefka 2022						
	Marcantonio 2014-Total sample	Sens: 95% (NR) Spec: 94% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	3 min
	Marcantonio 2014-Patients with dementia	Sens: 96% (NR) Spec: 86% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	3 min
	Marcantonio 2014-Patients without dementia	Sens: 93% (NR) Spec: 96% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	3 min
	Reported in van Velthuijsen 2016						
	Marcantonio 2014	Sens: 95% (NR) Spec: 94% (NR)	NR	NR	NR	NR	3 min
DOS	**Reported in Brefka 2022**						
	Van Gemert 2007	Sens: 89% (NR) Spec: 88% (NR)	NR	NR	NR	≥3 = suspected delirium	5 min

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	Koster 2009	Sens: 100% (NR) Spec: 97% (NR)	NR	NR	NR	≥3 = suspected delirium	5 min
	Reported in Kim 2021						
	Koster 2009	Sens: 100% (86–100) Spec: 97% (90–99)	NR	LR+: 24.92 (8.91–69.69) LR-:0.02 (0.00–0.32)	NR	≥3	NR
	Reported in van Velthuijsen 2016						
	Koster 2009	Sens: 100% (NR) Spec: 97% (NR)	NR	NR	NR	≥2	5 min
	Van Gemert 2007	Sens: 89% (NR) Spec: 87% (NR)	NR	NR	NR	3	5 min
	Reported in Watt 2021						
	Detroyer 2014	Sens: 81.8% (52−95) Spec: 96.1% (90−98)	NR	NR	NR	optimal cut-off score ⩾3; diagnostic score binary	NR
	Jorgensen 2017	Sens: 97% (81−100) Spec: 89% (75−96)	NR	NR	NR	optimal cut-off score ⩾3; diagnostic score ⩾18	NR
	Neefjes 2019	Sens: >99.9% (95.8–100) Spec: 99.6.% (95.5–100)	NR	NR	NR	optimal cut-off score ⩾3; diagnostic score ⩾17.5	NR
SQiD	**Reported in Brefka 2022**						
	Sands 2010 – vs. psychiatrist interview	Sens: 80% (NR) Spec: 71% (NR)	NR	NR	NR	"yes" = suspected delirium	<1 min

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	Lin 2015 – vs. DSM-IV	Sens: 77% (NR) Spec: 51% (NR)	NR	NR	NR	"yes" = suspected delirium	<1 min
	Reported in Calf 2021						
	Han 2018 – answered by patient	Sens: 62% (47-75) Spec: 79% (74-83)	NR	NR	NR	NR	NR
	Han 2018 – answered by surrogate	Sens: 91% (76-98) Spec: 77% (71-82)	NR	NR	NR	NR	NR
	Reported in Rosgen 2018						
	Sands 2010	Sens: 80% (28.4-99.5) Spec: 71% (41.9-91.6)	PPV: 50 (15.7-84.3) NPV: 91 (58.7-99.8)	NR	NR	"yes" = suspected delirium	NR
	Reported in van Velthuijsen 2016						
	Lin 2015	Sens: 77% (NR) Spec: 51% (NR)	NR	NR	NR	NR	NR
	Reported in Watt 2021						
	Sands 2010	Sens: 80% (28.4–99.5) Spec: 71% (41.9–91.6)	NR	NR	NR	Binary	NR
RADAR	**Reported in Quispel-Aggenbach 2018**						
	Voyer 2015 (RADAR 1–4 × daily)	Sens: 65% (43–84) Spec: 71% (64–78)	NR	NR	NR	>0 item present	7 seconds - <1 min

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	Voyer 2015 (RADAR 3–4 × daily)	Sens: 73% (39–94) Spec: 67% (57–76)	NR	NR	NR	>0 item present	7 seconds - <1 min
	Bilodeau 2016 (dementia patients only)	Sens: 100% (3–100) Spec: 77% (58–90)	NR	NR	NR	>0 item present	7 seconds - <1 min
	Koop 2016	Sens: 100% (3–100) Spec: 69% (39–91)	NR	NR	NR	>0 item present	7 seconds - <1 min
	Pelletier 2017	Sens: 100% (16–100) Spec: 72% (59–86)	NR	NR	NR	>0 item present	7 seconds - <1 min
	Reported in Brefka 2022						
	Voyer 2015 (RADAR 3–4 × daily)	Sens: 73% (NR) Spec: 67% (NR)	NR	NR	NR	≥1 “yes” = suspected delirium	<1 min
ICDSC	**Reported in Patel 2018**						
	Frenette 2016	Sens: 64% (49-77) Spec: 79% (63-89)	PPV: 74% (55-87) NPV: 69% (54-81)	NR	NR	NR	NR
	Reported in van Velthuijsen 2016						
	Giusti and Piergentili 2012	Sens: 69% (NR) Spec: 100% (NR)	NR	NR	NR	≥4	NR
CAM ICU	**Reported in Aldwikat 2022**						
	Neufeld 2013	Sens: 28% (NR)	NR	NR	NR	NR	2 min

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
		Spec: 98% (NR)					
	Reported in Brefka 2022						
	Ely 2001a - Total sample	Sens: 95-100% (NR) Spec: 89 -93% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	<5 min
	Ely 2001a - Patients ≥65 years	Sens: 90-100% (NR) Spec: 83-100% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	<5 min
	Ely 2001a - Patients with dementia	Sens: 100% (NR) Spec: 100% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	<5 min
	Reported in Mansutti 2019						
	Mitasova 2012	Sens: 76% (55-91) Spec: 98% (93-100)	PPV: 91% (70–99) NPV: 94% (88–98)	NR	NR	NR	NR
	Reported in Patel 2018						
	Frenette 2016	Sens: 62% (44-76) Spec: 74% (59-85)	PPV: 63% (45-78) NPV: 70% (55-82)	NR	NR	NR	NR
	Mitasova 2012	Sens: 76% (55-91) Spec: 98% (93-100)	PPV: 91% (70-99) NPV: 94% (88-98)	NR	NR	NR	NR
	Reported in van Velthuijsen 2016						

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	Han 2014	Sens: 69-72% (NR) Spec: 99% (NR)	NR	NR	NR	NR	1-2 min
	Luetz 2010	Sens: 79% (NR) Spec: 97% (NR)	NR	NR	NR	NR	1-2 min
	Mitasova 2012	Sens: 76% (NR) Spec: 98% (NR)	NR	NR	NR	NR	1-2 min
	Neufeld 2013	Sens: 28% (NR) Spec: 98% (NR)	NR	NR	NR	NR	1-2 min
	Pipanmekaporn 2014	Sens: 92% (NR) Spec: 95% (NR)	NR	NR	NR	NR	1-2 min
	Powers 2013	Sens: 45% (NR) Spec: 89% (NR)	NR	NR	NR	NR	1-2 min
B-CAM	**Reported in Brefka 2022**						
	Han 2013	Sens: 78-84% (NR) Spec: 96-97% (NR)	NR	NR	NR	CAM algorithm:1+2+(3 or 4) positive = suspected delirium	<5 min
	Reported in Calf 2021						
	Baten 2018	Sens: 65% (50-79) Spec: 94% (90-96)	NR	NR	NR	NR	NR
	Han 2013	Sens: 84% (71-93) Spec: 96% (93-98)	NR	NR	NR	NR	NR
	Reported in van Velthuijsen 2016						
	Han 2013	Sens: 78-84%	NR	NR	NR	>1	<1 min

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
		Spec: 96-97%					
	Reported in Watt 2021						
	Wilson 2019	Sens: 80% (40−96) Spec: 87% (67−96)	NR	NR	NR	Binary	NR
NuDESC	**Reported in Aldwikat 2022**						
	Neufeld 2013	Sens: 32% (NR) Spec: 92% (NR)	NR	NR	NR	≥2	2-3 min
	Neufeld 2013	Sens: 80% (NR) Spec: 69% (NR)	NR	NR	NR	≥1	2-3 min
	Radke 2008	Sens: 95% (NR) Spec: 87% (NR)	NR	NR	NR	NR	2-3 min
	Saller 2019	Sens: 27% (NR) Spec: 99% (NR)	NR	NR	NR	NR	2-3 min
	Reported in van Velthuijsen 2016						
	Lingehall 2013	Sens: 72% (NR) Spec: 81% (NR)	NR	NR	NR	≥2	<2 min
	Leung 2008	Sens: 96% (NR) Spec: 79% (NR)	NR	NR	NR	>0	<2 min
	Luetz 2010	Sens: 82% (NR) Spec: 83% (NR)	NR	NR	NR	2 and 1	<2 min
	Neufeld 2013	Sens: 32-80% (NR) Spec: 69-92% (NR)	NR	NR	NR	≥2 and ≥1	<2 min
	Reported in van Watt 2021						
	de la Cruz 2015 – nurse	Sens: 63% (NR) Spec: 67% (NR)	NR	NR	NR	diagnostic score ⩾7	NR

Test	Primary study	Sens/Spec (95% CI)	PPV/NPV (95% CI)	LR+/LR- (95% CI)	Area under curve (95% CI)	Cut-off score	Time to administer test
	de la Cruz 2015 – caregiver evening	Sens: 35% (NR) Spec: 80% (NR)	NR	NR	NR	diagnostic score ⩾7	NR
	de la Cruz 2015 – caregiver night	Sens: 21% (NR) Spec: 85% (NR)	NR	NR	NR	diagnostic score ⩾7	NR

Headings – CI: Confidence Interval; LR: Likelihood Ratio; NPV: Negative Predictive Value; PPV: Positive Predictive Value; Sens: Sensitivity; Spec: Specificity

Index tests – 3D-CAM: 3-Minute Diagnostic Interview for Confusion Assessment Method; 4AT: 4 ‘A’s Test; B-CAM: Brief Confusion Assessment Method; CAM: Confusion Assessment Method; Brief Confusion Assessment Method for the Intensive Care Unit; DOS: Delirium Observation Screening Scale; ICDSC: Intensive Care Delirium Screening Checklist; NuDESC: Nursing Delirium Screening Scale; RADAR: Routine or Recognising Acute Delirium As part of your Routine; SQiD: Single Question to Identify Delirium; UB2-CAM: Ultra Brief Confusion Assessment Method

See appendix D for full evidence tables.

1.1.6.1 Narrative summary of the diagnostic evidence

Matrices were constructed for each test to show the spread of primary studies across systematic reviews and to identify any duplication of primary studies across several systematic reviews. The matrices also helped prevent double counting of primary studies when calculating median sensitivity and specificity values for each test.

Evidence for 4AT

Table 6 Matrix of primary studies focussing on 4AT captured within systematic reviews

		Systematic reviews								
		Aldwikat 2022	Brefka 2022	Calf 2021	Jeong 2020a	Kim 2021	Mansutti 2019	Patel 2018	Tieges 2021	van Velthuijsen 2016
Primary studies	Bellelli 2014	No	Yes	No	Yes	No	No	No	Yes	Yes
	Gagné 2018	No	No	Yes	Yes	No	No	No	Yes	No
	Shenkin 2019	No	No	Yes	No	No	No	No	Yes	No
	O'Sullivan 2018	No	No	Yes	Yes	No	No	No	Yes	No
	Asadollahi 2016	No	No	No	Yes	No	No	No	Yes	No
	Myrstad 2019	No	No	No	Yes	No	No	No	Yes	No
	Casey 2019	No	No	No	Yes	No	No	No	No	No
	MacLullich 2019	No	No	No	Yes	No	No	No	No	No

		Systematic reviews								
		Aldwikat 2022	Brefka 2022	Calf 2021	Jeong 2020a	Kim 2021	Mansutti 2019	Patel 2018	Tieges 2021	van Velthuijsen 2016
	Kuladee 2016	No	No	No	Yes	No	No	No	Yes	No
	Hendry 2016	No	No	No	Yes	No	No	No	Yes	No
	De 2017	No	No	No	Yes	No	No	No	Yes	No
	Saller 2019	Yes	No	No	Yes	Yes	No	No	Yes	No
	Infante 2017	No	No	No	Yes	No	Yes	No	Yes	No
	Lees 2013	No	No	No	Yes	No	Yes	Yes	Yes	No
	Kutlubaev 2016	No	No	No	No	No	Yes	No	Yes	No
	Al-Jumayli 2018	No	No	No	No	No	No	No	Yes	No
	Chang 2019	No	No	No	No	No	No	No	Yes	No
	Kazim 2016 (2 studies)	No	No	No	No	No	No	No	Yes	No

Aldwikat 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of 4AT delivered by research assistants in post-anaesthetic care units compared to DSM-5. A sensitivity of 96% and sepcificity of 99% were reported. Study designs considered for this review included prospective and retrospective cohort studies, randomised and non-randomised controlled trials.

Brefka 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of 4AT delivered by untrained getriatricians in acute care and emergency settings compared to DSM-IV. A sensitivity of 90% and sepcificity of 84% were reported. An AUC range of 0.89-0.93 was also found. The study design was not reported.

Calf 2021 (RoB [assessed with ROBIS]: Low) included 3 studies that looked at the diagnostic accuracy of 4AT in emergency departments. The person administering the test was not reported. One study used a reference standard of DSM-IV, one used DSM-5 and one used CAM. A pooled sensitivity of 87% (95% CI 74-94) and pooled specificity of 87% (95% CI 60-97) were reported. Cohort and case-control studies were considered for this review.

Jeong 2020a (RoB [assessed with ROBIS]: Low) included 11 studies that looked at the diagnostic accuracy of 4AT. One study was set in nursing homes and daily caring centres, with the remaining 10 studies being set in various hospital units. Of the included studies, 8 used either DSM-IV or DSM-V as a reference standard, with 2 of these studies using a combination DSM-IV or DSM-5 and another reference. The remaining 3 studies used CAM or 3D-CAM as a reference. The person administering the test was not reported. A pooled sensitivity of 81.5% (95% CI 70.7–89.0) and pooled specificity of 87.5% (95% CI 79.5–92.7) were reported. An AUC of 0.911 was also found. Positive and negative likelihood ratios were reported for individual studies but not pooled. The study design was not reported.

Kim 2021 (RoB [assessed with ROBIS]: High) included 1 prospective cohort study that looked at the diagnostic accuracy of 4AT delivered by trainees in a post-surgery setting compared to DSM-IV. A sensitivity of 95% (95% CI 77–99) and specificity of 100% (95% CI 99–100). A positive likelihood ratio of 975.91 (95% CI 60.94–15,614.60) and negative likelihood ratio of 0.06 (95% CI 0.01–0.30) were also found.

Mansutti 2019 (RoB [assessed with ROBIS]: Low) included 3 studies that looked at the diagnostic accuracy of 4AT in hospital stroke units or neurovascular departments. In 2 studies, tests were delivered by neurologists and compared to DSM-IV. In the remaining study, the test was delivered by trained medical students and CAM was used as a reference standard. Study designs included a quasi-experimental and observational study. Outcomes were reported on an individual study basis. Sensitivity ranged from 90.2–100% and specificity ranged from 64.5–86%. Positive and negative predictive values ranged from 43–86% and 85.6–100% respectively. Finally, AUC ranged from 0.82 to 0.89. See Table 5 for the full set of results.

Patel 2018 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of 4AT in ICUs compared to CAM. The person administering the test was not reported. A sensitivity of 100% (95%CI: 74–100) and specificity of 82% (95%CI: 72–89). Positive and negative predictive values of 43% and 100% were also found. The study design was not reported.

Tieges 2021 (RoB [assessed with ROBIS]: Low) included 17 studies (11 prospective, 3 retrospective and 2 cross-sectional) from 16 papers that looked at the diagnostic accuracy of 4AT. One study was set in nursing homes and daily caring centres, with the remaining studies being set in various hospital units. Of the included studies, 12 used either DSM-IV, DSM-IV-TR or DSM-5 as a reference standard. A pooled sensitivity of 88% (95% CI 80–93) and pooled specificity of 88% (95% CI 82–92) were reported.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of 4AT delivered by doctors in an acute geriatric setting compared to

DSM-IV-TR. A sensitivity of 90% and specificity of 84% were found. The study design was not reported.

Evidence for CAM

Table 7 Matrix of primary studies focussing on CAM captured within systematic reviews

		Systematic reviews					
		Aldwikat 2022	**Brefka 2022**	**Calf 2021**	**Kim 2021**	**van Velthuijsen 2016**	**Watt 2021**
Primary studies	Inouye 1990	No	Yes	No	No	No	No
	Fabri 2001	No	No	Yes	No	Yes	No
	Shenkin 2019	No	No	Yes	No	No	No
	Radtke 2008	Yes	No	No	Yes	No	No
	Radtke 2010	No	No	No	Yes	No	No
	Shi 2014	No	No	No	Yes	No	No
	Ryan 2009	No	No	No	No	Yes	Yes
	González 2004	No	No	No	No	Yes	No
	Hestermann 2009	No	No	No	No	Yes	No
	Laurila 2002	No	No	No	No	Yes	No
	Leung 2008	No	No	No	No	Yes	No
	Lin 2015	No	No	No	No	Yes	No
	Martins 2015	No	No	No	No	Yes	No
	Monette 2001	No	No	No	No	Yes	No
	Pompei 1995	No	No	No	No	Yes	No
	Thomas 2012	No	No	No	No	Yes	No
	Wongpakaran 2011	No	No	No	No	Yes	No

Aldwikat 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of CAM delivered by research assistants in post-anaesthetic care units compared to DSM-IV. A sensitivity of 43% and sepcificity of 98% were reported. Study designs considered for this review included prospective and retrospective cohort studies, randomised and non-randomised controlled trials.

Brefka 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of CAM delivered by clinicians or lay raters in acute care and emergency settings compared to geriatric psychiatrist rating after comprehensive assessment. A sensitivity range of 94-100% and specificity range of 90-95% were found. The study design was not reported.

Calf 2021 (RoB [assessed with ROBIS]: Low) included 2 studies that looked at the diagnostic accuracy of CAM in emergency departments compared to DSM-IV. The person administering the test was not reported. Outcomes were reported on an individual study basis. Sensitivity ranged from 40-94% and specificity ranged from.96-100%. See Table 5 for the full set of results. Cohort and case-control studies were considered for this review.

Kim 2021 (RoB [assessed with ROBIS]: High) included 3 studies that looked at the diagnostic accuracy of CAM delivered by registered nurses in a post-surgery setting compared to DSM-IV. The studies had a prospective cohort design. Outcomes were pooled from 4 studies and reported under CAM (and variants), which included results from studies that focussed on CAM-ICU as well as CAM. A pooled sensitivity of 47% (95% CI 37–56%) and pooled specificity of 99% (98–99). A pooled positive likelihood of 32.10 (95% CI 7.01–146.93) and pooled negative likelihood ratio of 0.55 (95% CI 0.34–0.87). It was noted that, despite there being a total of 5 included studies that reported on the diagnostic accuracy of CAM and CAM-ICU, only 4 were considered for the pooled outcomes.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 11 studies that looked at the diagnostic accuracy of CAM delivered by a medical doctors, psychiatrists, psychologists, nurses and researchers. Study settings icluded general hospitals, geriatric units, intermediate care units, emergency departments, post-operative units and palliative care. In 9 studies, DSM-IV or DSM-IV-TR was used as a reference standard. The remaining two studies used DMS-III-R. Outcomes were reported on an individual study basis. Sensitivity ranged from 46-94% and specificity ranged from.63-100%. See Table 5 for the full set of results. Study designs were not reported.

Watt 2021 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of CAM in hospice PCUs compared to DSM-IV. The person administering the test was not reported. A sensitivity of 88% (95% CI 62–98) and sepcificity of 100% (95% CI 88–100) were found. Primary quantitative research studies were considered for this review.

Evidence for 3D-CAM

Table 8 Matrix of primary studies focussing on 3D-CAM captured within systematic reviews

		Systematic reviews		
		Aldwikat 2022	**Brefka 2022**	**van Velthuijsen 2016**
Primary studies	Marcantonio 2014	No	Yes	Yes
	Olbert 2019	Yes	No	No

Aldwikat 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of 3D-CAM delivered by research assistants in post-anaesthetic care units compared to DSM-5. A sensitivity of 100% and sepcificity of 88% were reported. Study designs considered for this review included prospective and retrospective cohort studies, randomised and non-randomised controlled trials.

Brefka 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of 3D-CAM delivered by trained physicians or nurses in acute care and emergency settings compared to DSM-IV. A sensitivity of 95% and specificity 94% were reported. Outcomes sub-groups of patients with and without dementia were also reported (see Table 5 for details). The study design was not reported.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of 3D-CAM delivered by reserarch assistants in general or geriatirc units compared to DSM-IV. A sensitivity of 95% and specificity of 94% were found. The study design was not reported.

Evidence for DOS

Table 9 Matrix of primary studies focussing on DOS captured within systematic reviews

		Systematic reviews				
		Brefka 2022	**Kim 2021**	**Park 2021**	**van Velthuijsen 2016**	**Watt 2021**
Primary studies	van Gemert 2007	Yes	No	Yes	Yes	No
	Koster 2009	Yes	Yes	No	Yes	No
	Neefjes 2019	No	No	Yes	No	Yes
	Hasemann 2018	No	No	Yes	No	No
	Jorgensen 2017	No	No	Yes	No	Yes
	Gavinski 2016	No	No	Yes	No	No
	Schrøder Pedersen 2014	No	No	Yes	No	No
	Detroyer 2014	No	No	Yes	No	Yes

Brefka 2022 (RoB [assessed with ROBIS]: High) included 2 studies that looked at the diagnostic accuracy of 3D-CAM delivered by nurses in acute care and emergency settings compared to DSM-IV. Outcomes were reported on an individual study basis. A sensitivity range of 89-100% and specificity range of 97-100% were found. The study designs were not reported.

Kim 2021 (RoB [assessed with ROBIS]: High) included 1 prospective cohort study that looked at the diagnostic accuracy of DOS delivered by registered nurses in a post-surgery setting compared to DSM-IV. A sensitivity of 100% (95% CI 86–100) and specificity of 97% (95% CI 90–99). A positive likelihood ratio of 24.92 (95% Ci 8.91–69.69) and negative likelihood ratio of 0.02 (95% Ci 0.00–0.32) were also found.

Park 2021 (RoB [assessed with ROBIS]: Low) included 8 studies (7 prospective cohort studies and that looked at the diagnostic accuracy of DOS. One study was set in a home hospice, with the remaining 7 set in various hospital wards (general, cardiac surgical and palliative. In 5 studies, DSM-IV was used as the reference standard, whilst DRS-R-98 was used in 3 studies. A pooled sensitivity of 90% (95% CI 76-97) and specificity of 92% (95% CI 88-94) were reported. An AUC of 0.94 was also found.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 2 studies that looked at the diagnostic accuracy of DOS delivered by nurses in general hospital and cardiac surgery settings compared to DSM-IV. The person administering the test was not reported. Outcomes were reported on an individual study basis. Sensitivity ranged from 89-100%% and specificity ranged from.87-97%%. See Table 5 for the full set of results. Study designs were not reported.

Watt 2021 (RoB [assessed with ROBIS]: High) included 3 studies that looked at the diagnostic accuracy of DOS in palliative care units, community hospices and inpatient oncology units. The person administering the test was not reported. Two studies used DRS-R-98 as a reference standard and 1 study used CAM. Outcomes were reported on an individual study basis. A sensitivity range of 81.8->99.9% and a specificity range 96.1-99.6%. The study designs were not reported.

Evidence for SQiD

Table 10 Matrix of primary studies focussing on SQiD captured within systematic reviews

		Systematic reviews				
		Brefka 2022	Calf 2021	Rosgen 2018	van Velthuijsen 2016	Watt 2021
Primary studies	Sands 2010	Yes	No	Yes	No	Yes
	Han 2018	No	Yes	No	No	No
	Lin 2015	No	No	No	Yes	No

Brefka 2022 (RoB [assessed with ROBIS]: High) included 2 studies that looked at the diagnostic accuracy of SQiD in acute care and emergency settings. In 1 study, tests were administered by trained clinicians or lay raters and a psychiatrist interview was used as a reference standard. The other did not the person administering the test and used DSM-IV as a reference standard. A sensitivity range of 77-80% and specificity range of 51-71% were found. The study designs were not reported.

Calf 2021 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of SQiD in emergency departments compared to DSM-IV-TR. The person administering the test was not reported. A sensitivity of 62% (95% CI 47-75) and a specificity of 79% (95% CI 74-83) were reported when the patient completed the index test. When a surrogate completed the test, a separate sensitivity and specificity of 91% (95% CI 76-98) and 77% (95% CI 71-82) respectively were found. Cohort and case-control studies were considered for this review.

Rosgen 2018 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of SQiD in an inpatient oncology unit compare to a psychiatrist interview using DSM-IV criteria. The person administering the test was not reported. A sensitivity of 80% (95% CI, 28.3-99.5%) and specificity of 71% (95% CI, 41.9-91.6%) were reported. Positive and negative predicitve values of 50% (95% CI, 15.7-84.3%) and 91% (95% CI, 58.7-99.8%) respectively were also found. Observational study designs such as cohort and cross-sectional studies were considered for this review.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of SQiD delivered by informal carers in general medicine wards compared to DSM-IV. A sensitivity of 77% and a specificity of 51% were found. The study design was not reported.

Watt 2021 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of SQiD in inpatient oncology settings compared to DSM-IV. The person administering the test was not reported. A sensitivity of 80% (95% CI 28.4-99.5) an a specificity of 71% (95% CI 41.9-91.6) were found. Primary quantitative research studies were considered for this review.

Evidence for RADAR

Table 11 Matrix of primary studies focussing on RADAR captured within systematic reviews

		Systematic reviews	
		Brefka 2022	Quispel-Aggenbach 2018
Primary studies	Voyer 2015	Yes	Yes
	Bilodeau 2016	No	Yes
	Koop 2016	No	Yes
	Pelletier 2017	No	Yes

Brefka 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of RADAR delivered by trained clinical staff or lay raters in acute care and emergency settings compared to DSM-IV. A sensitvity of 73% and a specificity of 67% were found. The study design was not reported.

Quispel-Aggenbach (RoB [assessed with ROBIS]: Low) 2018 included 4 studies that looked at the diagnostic accuracy of RADAR delivered by nurses, nurse assistants and research assistants. Settings included acute care hospitals,nursing homes and the rehabilitation ward of a nursing home. Outcomes were reported on an individual study basis. A sensitvity range of 65-100% and a specificity range of 67-77% were found. Study designs were not reported.

Evidence for ICDSC

Table 12 Matrix of primary studies focussing on ICDSC captured within systematic reviews

		Systematic reviews				
		Brefka 2022	Chen 2021	Ho 2020	Patel 2018	van Velthuijsen 2016
Primary studies	Bergeron 2001	Yes	Yes	Yes	No	No
	Yu 2013	Yes	No	No	No	No
	Barman 2018	No	Yes	Yes	No	No
	Boettger 2017	No	Yes	Yes	No	No
	Chanques 2018	No	Yes	Yes	No	No
	Domenico and Federica 2012	No	Yes	No	No	No
	George 2011	No	Yes	No	No	No

Systematic reviews						
		Brefka 2022	Chen 2021	Ho 2020	Patel 2018	van Velthuijsen 2016
	Gusmao-Flores 2011	No	Yes	Yes	No	No
	Kose 2016	No	Yes	Yes	No	No
	Larsen 2019	No	Yes	No	No	No
	Nishimura 2016	No	Yes	Yes	No	No
	Radtke 2009	No	Yes	No	No	No
	Van Eijk 2009	No	Yes	Yes	No	No
	Frenette 2016	No	No	No	Yes	No
	Giusti and Piergentili 2012	No	No	No	No	Yes

Brefka 2022 (RoB [assessed with ROBIS]: High) included 2 studies that looked at the diagnostic accuracy of ICDSC delivered by a clinician or nurse in acute care and emergency settings compared to delirium diagnosis by psychiatrist. A pooled sensitivity of 99% and specificity of 64% were found. The study designs were not reported.

Chen 2021 (RoB [assessed with ROBIS]: Low) included 12 studies (8 prospective and 4 cross-sectional) that looked at the diagnostic accuracy of ICDSC in intensive care units. The person administering the test was not reported. Reference standards were not reported for each study, but it was noted that most studies (n=9) used either DSM-IV, DSM-IV-TR or DSM-5. A pooled sensitivity of 83% (95% CI 74-90) a specificity of 87% (95%CI 78-93).

Ho 2020 (RoB [assessed with ROBIS]: High) included 8 studies (6 prospective and 2 cross-sectional) that looked at the diagnostic accuracy of ICDSC in intensive care units. Those delivering the test included nurses, doctors, intensivists and investigators. DSM-IV, DSM-IV-TR and DSM-5 were used as reference standards in 7 studies, with 1 study using clinical diagnosis confirmed by a psychiatrist. A pooled sensitivity of 87% (95% CI 70-95) and specificity of 91% (95% CI 85-95). An AUC of 0.95 was also found.

Patel 2018 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of ICDSC in ICUs compared to DSM-IV. The person administering the test was not reported. A sensitivity of 64% (95%CI: 49-77) and specificity of 79% (95%CI: 63–89) were reported. Positive and negative predictive values of 74% (95%CI: 55–87) and 69% (95%CI: 54-81) respectively were also found. The study design was not reported.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of ICDSC delivered by nurses in general hospital settings compared to DSM-IV-TR. A sensitivity of 69% and a specificity of 100% were found. The study design was not reported.

Evidence for CAM-ICU

Table 13 Matrix of primary studies focussing on CAM-ICU captured within systematic reviews

		Systematic reviews								
		Aldwikat 2022	**Brefka 2022**	**Calf 2021**	**Chen 2021**	**Ho 2020**	**Kim 2021**	**Mansutti 2019**	**Patel 2018**	**van Velthuijsen 2016**
Primary studies	Ely 2001a	No	Yes	No	Yes	Yes	No	No	No	No
	Ely 2001b	No	No	No	Yes	Yes	No	No	No	No
	Han 2014	No	No	Yes	No	No	No	No	No	Yes
	Meeberg 2016	No	No	Yes	No	No	No	No	No	No
	Adamis 2012	No	No	No	Yes	Yes	No	No	No	No
	Akinci 2005	No	No	No	Yes	No	No	No	No	No
	Aljuaid 2018	No	No	No	Yes	Yes	No	No	No	No
	Barman 2018	No	No	No	Yes	Yes	No	No	No	No
	Boettger 2017	No	No	No	Yes	No	No	No	No	No
	Bui 2017	No	No	No	Yes	No	No	No	No	No
	Chanques 2018	No	No	No	Yes	Yes	No	No	No	No
	Chuang 2007	No	No	No	Yes	Yes	No	No	No	No
	Danzeng 2019	No	No	No	Yes	No	No	No	No	No
	Gusmao-Flores 2011	No	No	No	Yes	Yes	No	No	No	No

		Systematic reviews								
		Aldwikat 2022	Brefka 2022	Calf 2021	Chen 2021	Ho 2020	Kim 2021	Mansutti 2019	Patel 2018	van Velthuijsen 2016
	Heo 2011	No	No	No	Yes	Yes	No	No	No	No
	Karlicic 2016	No	No	No	Yes	Yes	No	No	No	No
	Koga 2015	No	No	No	Yes	Yes	No	No	No	No
	Larsen 2019	No	No	No	Yes	No	No	No	No	No
	Lin 2004	No	No	No	Yes	Yes	No	No	No	No
	Luetz 2010	No	No	No	Yes	Yes	No	No	No	Yes
	Mitasova 2010	No	No	No	Yes	No	No	No	No	No
	Mitasova 2012	No	No	No	Yes	Yes	No	Yes	Yes	Yes
	Nishimura 2016	No	No	No	Yes	Yes	No	No	No	No
	Pipanmekaporn 2014	No	No	No	Yes	Yes	No	No	No	Yes
	Selim 2018	No	No	No	Yes	Yes	No	No	No	No
	Tobar 2010	No	No	No	Yes	No	No	No	No	No
	Toro 2009	No	No	No	Yes	No	No	No	No	No

		Systematic reviews								
		Aldwikat 2022	Brefka 2022	Calf 2021	Chen 2021	Ho 2020	Kim 2021	Mansutti 2019	Patel 2018	van Velthuijsen 2016
	Van Eijk 2011	No	No	No	Yes	Yes	No	No	No	No
	Van Eijk 2009	No	No	No	Yes	Yes	No	No	No	No
	Vreeswijk 2009	No	No	No	Yes	Yes	No	No	No	No
	Wang 2013	No	No	No	Yes	Yes	No	No	No	No
	Guenther 2010	No	No	No	No	Yes	No	No	No	No
	Boettger 2018	No	No	No	No	Yes	No	No	No	No
	Neufeld 2013	Yes	No	No	No	No	Yes	No	No	Yes
	Wongviriyawong 2019	No	No	No	No	No	Yes	No	No	No
	Frenette 2016	No	No	No	No	No	No	No	Yes	No
	Powers 2013	No	No	No	No	No	No	No	No	Yes

Aldwikat 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of CAM-ICU delivered by research assistnats in post-anaesthetic care units compared to DSM-IV. A sesntivity of 28% and a specificity of 98% were reported. Study designs considered for this review included prospective and retrospective cohort studies, randomised and non-randomised controlled trials.

Brefka 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of CAM-ICU delivered by trained lay raters or clinicians in acute care and emergency settings compared to DSM-IV. A sensitivity range of 95-100% and specificity range 89–93% were reported. Outcomes sub-groups of patients ≥65 years and patients with dementia were also reported (see Table 5 for details). The study design was not reported.

Calf 2021 (RoB [assessed with ROBIS]: Low) included 2 studies that looked at the diagnostic accuracy of CAM-ICU in emergency departments compared to DSM-IV-TR. The person administering the test was not reported. Outcomes were reported on an individual study basis. A pooled sensitivity of 85% (95% CI 39-98) and specificity of 98% (95% CI 94-99) were reported. Cohort and case-control studies were considered for this review.

Chen 2021 (RoB [assessed with ROBIS]: Low) included 29 studies (22 cross-sectional and 7 trials) that looked at the diagnostic accuracy of CAM-ICU in intensive care units. The person administering the test was not reported. Reference standards were not reported for each study, but it was noted that DSM was the most commonly used reference in the included studies (n=29). A pooled sensitivity of 84% (95% CI 77-88) a specificity of 95% (95% CI 91-97).

Ho 2020 (RoB [assessed with ROBIS]: High) included studies 23 (17 prospective and 6 cross-sectional) that looked at the diagnostic accuracy of CAM-ICU in intensive care units. Those administering the test included nurses, doctors, independent investigators, intensivists, physician or nurse investigators and examiners. All studies used either DSM-IV, DSM-IV-TR or DSM-5 as a reference standard. A pooled sensitivity of 85% (95% CI 77-91) and specificity of 95 (95% CI 90-97). An AUC of 0.96 was also found.

Kim 2021 (RoB [assessed with ROBIS]: High) included 2 studies that looked at the diagnostic accuracy of CAM-ICU, but outcomes were reported under CAM and variants. See the section on CAM above for details.

Mansutti 2019 (RoB [assessed with ROBIS]: Low) included 1 diagnostic and observational study that looked at the diagnostic accuracy delivered by a trained junior physician compared to a DSM evaluation. A sensitivity of 76% (95% CI 55-91) and specificity of 98% (95% CI 93-100) were reported. Positive and negative predictive values of 63% (95% CI 45-78) and 94% (95% CI 88-98) respectively were also found. Finally, a likelihood ratio was reported, however the review did not state whether this was a positive or negative likelihood ratio (see Appendix D for details).

Patel 2018 (RoB [assessed with ROBIS]: Low) included 2 studies that looked at the diagnostic accuracy of CAM-ICU in ICUs compared to DSM-IV. The person administering the test was not reported. Outcomes were reported on an individual study basis A sensitivity range of 62-76% and specificity range of 74-98% were reported. Positive and negative predictive values ranged from of 63-91% and 70-94% respectively. The study design was not reported.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 6 studies that looked at the diagnostic accuracy of CAM-ICU compared to DSM-IV or DSM-IV-TR. Settings included emergency departments, ICUs, general hospitals, post-operative and geriatric units. Those administering the tests included doctors, nurses, researchers and psychologists. Outcomes were reported on an individual study basis. A sensitivity range of 28-92% and a specificity range of 89-99% were found (see Table 5 for details). The study designs were not reported.

Evidence for B-CAM

Table 14 Matrix of primary studies focussing on B-CAM captured within systematic reviews

		Systematic reviews			
		Brefka 2022	**Calf 2021**	**van Velthuijsen 2016**	**Watt 2021**
Primary studies	Han 2013	Yes	Yes	Yes	No
	Baten 2018	No	Yes	No	No
	Wilson 2019	No	No	No	Yes

Brefka 2022 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of B-CAM delivered by trained raters or physicians in acute care and emergency settings compared to DSM-IV. A sensitivity range of 78-84% and specificity range of 96-97% were found. The study design was not reported.

Calf 2021 (RoB [assessed with ROBIS]: Low) included 2 diagnostic studies that looked at the diagnostic accuracy of B-CAM in emergency departments compared to DSM-IV or DSM-IV-TR. The person administering the test was not reported. Outcomes were reported on an individual study basis. A sensitivity range of 65-84% and specificity range of 94-96% were reported (see Table 5 for details). Cohort and case-control studies were considered for this review.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 1 study that looked at the diagnostic accuracy of B-CAM delivered by researchers in emergency departments compared to DSM-IV-TR. A sensitivity range of 78-84% and specificity range of 96-97% were found. The study designs were not reported.

Watt 2021 (RoB [assessed with ROBIS]: High) included 1 study that that looked at the diagnostic accuracy of B-CAM in palliative care unit and general inpatient settings compared to DSM-5. The person administering the test was not reported. A sensitivity of 80% (95% CI 40-96) an a specificity of 87% (95% CI 67-96) were found. Primary quantitative research studies were considered for this review.

Evidence for UB2-CAM

Table 15 Matrix of primary studies focussing on UB2-CAM captured within systematic reviews

		Systematic reviews
		Brefka 2022
Primary studies	Armstrong 2021	Yes
	Husser 2021	Yes
	Motyl 2020	Yes

Brefka 2022 (RoB [assessed with ROBIS]: High) inluded 3 studies that looked at the diagnostic accuracy of UB2-CAM delivered by trained physicians or nurses in in acute care and emergency settings compared to 3D-CAM. A pooled sensitivity of 93% and specificity of 95% were found. The study designs were not reported.

Evidence for NuDESC

Table 16 Matrix of primary studies focussing on NuDESC captured within systematic reviews

		Systematic reviews						
		Aldwikat 2022	Brefka 2022	Ho 2022	Jeong 2020b	Kim 2021	van Velthuijsen 2016	Watt 2021
Primary studies	Gaudreau 2005a	No	Yes	No	Yes	No	No	No
	Gaudreau 2005b	No	Yes	No	No	No	No	No
	Radtke 2008	Yes	No	Yes	Yes	Yes	No	No
	Saller 2019	Yes	No	Yes	Yes	Yes	No	No
	Neufeld 2013	Yes	No	Yes	Yes	Yes	Yes	No
	Abelha 2013	No	No	Yes	No	No	No	No
	Çınar and Eti Aslan 2019	No	No	Yes	No	No	No	No
	Ding 2016	No	No	Yes	No	No	No	No
	Lingehall 2013	No	No	Yes	Yes	Yes	Yes	No
	Luetz 2010	No	No	Yes	Yes	No	Yes	No
	Mei 2010	No	No	Yes	No	No	No	No
	Ning 2014	No	No	Yes	No	No	No	No
	Radtke 2010	No	No	Yes	Yes	Yes	No	No
	Heinrich 2019	No	No	No	Yes	No	No	No
	Birch 2018	No	No	No	Yes	No	No	No
	Hargrave 2017	No	No	No	Yes	Yes	No	No
	Leung 2008	No	No	No	Yes	No	Yes	No
	de la Cruz 2015	No	No	No	No	No	No	Yes

Aldwikat 2022 (RoB [assessed with ROBIS]: High) included 3 studies that looked at the diagnostic accuracy of NuDESC delivered by research assisants in post-anaesthetic care units compared to DSM-IV or DSM-5. Outcomes were reported on an individual study basis. A sensitivity range of 27-95% and specificity range of 69-99% were found. Study designs considered for this review included prospective and retrospective cohort studies, randomised and non-randomised controlled trials.

Brefka 2022 (RoB [assessed with ROBIS]: High) included 2 studies studies that looked at the diagnostic accuracy of NuDESC delivered by trained clinicians, nurses or lay raters in acute care and emergency settings compared to CAM. A pooled sensitivity of 86% and specificity of 87% were found. The study designs were not reported.

Ho 2022 (RoB [assessed with ROBIS]: Low) included 11 studies that looked at the diagnostic accuracy of NuDESC.in a post-surgery setting. Those administering the test included nurses, researchers or research assistants, physicians and psychiatrists. In 8 studies, DSM-IV or DSM-IV-TR were used as reference standards. The remaining 3 studies used ICDSC, CAM-ICU and DSM-5 and CAM-ICU as reference standards. A pooled sensitivity of 73% (95% CI: 44-90) and specificity of 93% (95% CI: 87–96) were reported. cohort studies, including prospective, cross-sectional,case-control, and retrospective were considered for this review.

Jeong 2020b (RoB [assessed with ROBIS]: High) included 11 prospective cohort studies that looked at the diagnostic accuracy of NuDESC. The person administering the test was not reported. Six of the reviewed studies were conducted in general wards, including rehabilitation units, medical units, and surgical units in a hospital; three in a recovery room or post-anaesthesia care unit; one in an intensive care unit; and one in an emergency department. In 10 studies, DSM-IV or DSM5 were used as reference standards. CAM was used as a reference standarad in the 1 remaining study. A pooled sensitvity of 68.6% (95% CI 55.3-79.5) and specificity of 89.4% (95% CI 83.3-93.5). An AUC of 0.882 was also found.

Kim 2021 (RoB [assessed with ROBIS]: High) included 6 prospective cohort studies that looked at the diagnostic accuracy of NuDESC delivered by registered nurses in a post-surgery setting compared to DSM-IV. Separate pooled sensitvities, specificities and likelihood ratios were reported for studies that used a cut-off score ≥2 (n=5) and a cut-off score of ≥1 (n=3). For studies that used a cut-off score of ≥2, a pooled sensitvity of 63% (95% CI 56-69) and specificity of 93% (95% CI 91-94) were reported. Positive and negative likelihood rations of 7.97 (95% CI 4.38-14.49) and 0.33 (95% CI 0.16-0.67) respectively were also found. For studies that used a cut-off score of ≥1, a pooled sensitvity 69% (95% CI 60-76) and specificity of 94% (95% CI 92-96) were reported. Positive and negative likelihood rations of 7.76 (95% CI 2.58[a]-23.32) and 0.38 (95% CI 0.29-0.48) respectively were also found. It was assumed that the lower range of the 95% CI for positive likelihood ratio was reported in error.

Van Velthuijsen 2016 (RoB [assessed with ROBIS]: Low) included 4 studies that looked at the diagnostic accuracy of NuDESC delivered by nurses compared to DSM-IV or DSM-IV-TR. Settings included geriatric units, cardiac surgery, post-operative units and ICUs. Outcomes were reported on an individual study basis. A sensitivity range of 32-96% and a specificity range of 67-92% were found. The study designs were not reported.

Watt 2021 (RoB [assessed with ROBIS]: High) included 1 study that looked at the diagnostic accuracy of NuDESC delivered by nurses or caregivers in a community hospital setting compared to MDAS. When NuDESC was delivered by a nurse, a sensitvity of 63% and a specificity of 67% were reported. When NuDESC was delivered by a caregiver, separate sensitvities and specificities were reported depending on whether the test was administered in the evening or at night. When NuDESC was delivered in the evening, a sensitivity of 63% and a specificity of 67% were reported. When NuDESC was delivered at night, a sensitivity

[a] Reported in Kim (2021) as 2,058 and assumed to be a typographical error.

of 63% and a specificity of 67% were reporte. Primary quantitative research studies were considered for this review.

1.1.7 Economic evidence

1.1.7.1 Included studies

A search was performed to identify published economic evaluations of relevance to this review question. This search retrieved 179 studies. Based on title and abstract screening, 178 of the studies could confidently be excluded for this question and following full text screening no further studies were excluded. Thus, the review for this question includes only one study from the existing literature.

1.1.7.2 Excluded studies

No studies were excluded at the full text review stage.

1.1.8 Summary of included economic evidence

Table 17: Economic evidence profile

Study	Applicability	Limitations	Other comments	Cost (£)	Effects (QALYs)	ICER (£/QALY)	Uncertainty
MacLullich et al (2019) The 4 'A's test for detecting delirium in acute medical patients: a diagnostic accuracy study	Directly applicable. The guideline update was triggered by this study, in a directly applicable population from the NHS and PSS perspective.	Minor limitations. The model structure was appropriate to answer the research question, but some parameter values were based on expert opinion.	Although the utility values were not derived using preferred methods, these were estimated from a diverse sample of experts, and sensitivity analyses were conducted.	Scottish: 4AT: £4,680 CAM: £4,770 Incr: -£90.35 English: 4AT: £4,416 CAM: £4,478 Incr: -£61.52	Scottish: 4AT: 0.14050 CAM: 0.14103 Incr: -0.00053 English: 4AT: 0.14050 CAM: 0.14103 Incr: -0.00053	Scottish: £170,553 (SWQ) English: £116,133 (SWQ)	The results of the probabilistic sensitivity analysis indicated that there is considerable uncertainty in the estimated cost-effectiveness due to clustering of incremental costs and QALYs around zero.

Incr: Incremental; SWQ: South West Quadrant

1.1.9 Economic model

No original economic modelling was conducted for this review.

1.1.10 Evidence statements

Diagnostic evidence statements

- Diagnostic evidence statements have not been provided because key diagnostic data has previously been covered in Table 3, Table 4, Table 5 and the narrative synthesis.

Economic evidence statement

- One published cost-utility study was identified comparing the 4AT with the CAM in older patients aged ≥70 years in emergency departments or acute general medical wards. The study found that the 4AT was more cost effective as although it was associated with marginally lower total QALYs than the CAM, it was also less costly. The study had some uncertainty, and since the incremental QALYs were so small, the ICER and the resulting conclusion of cost-effectiveness fluctuated significantly during probabilistic sensitivity analysis.

1.1.11 The committee's discussion and interpretation of the evidence

1.1.11.1. The outcomes that matter most

The committee agreed that the key outcome of interest for recommending tools for the assessment of delirium was diagnostic accuracy. Although the committee was interested in all measures of diagnostic accuracy, the majority of the included studies reported sensitivity and specificity. They agreed that both sensitivity and specificity were important but that while a test needed to be sensitive enough to detect delirium, it was also very important that tests could differentiate between delirium and other conditions that present in a similar way, for example depression or dementia. They also noted a need to establish whether tools were able to distinguish delirium superimposed on other conditions (for example dementia). The committee agreed that the evidence base was sufficient to recommend the use of specific delirium tools.

The committee were clear that diagnostic accuracy is not the main parameter of interest when deciding which assessment tool to use because several of the tools had high sensitivity and specificity. It was also key to consider implementation in practice, who could use the tool (and how much training they would require to do so), how long the assessment took, and what settings it could be delivered in (for example emergency departments, long term residential settings, or post-operative settings).

1.1.11.2 The quality of the evidence

The committee noted that the variation in numbers of papers included in the various systematic reviews could not be accounted for simply by the publication dates of the primary studies and the search dates of the systematic reviews. They agreed that the main drivers of these differences were the populations and settings that formed the inclusion criteria for the different systematic reviews. The committee agreed that it would be helpful if this were reported in a way that could usefully be disambiguated, but that a review of reviews methodology made that task impossible since each review included its own cluster of setting and populations.

No included systematic reviews reported GRADE assessments of their outcomes. The development team conducted a review of systematic reviews and did not conduct their own meta-analysis. Therefore, the methodological quality of the included systematic reviews was assessed using the ROBIS checklist. Out of the 17 systematic reviews included in the evidence review, 11 were rated as having a low risk of bias, with the remaining 6 having a high risk of bias (most commonly because of concerns about the data synthesis, for example not exploring bias in included primary studies or not addressing heterogeneity in the synthesis). The committee noted that diagnostic data for each test were reported in systematic reviews with a mixture of high and low risk of bias ratings. Therefore, they did not believe that the risk of bias ratings would have a substantial effect on the quality of evidence for any given test. The one exception was UB2-CAM, where diagnostic data was only reported in one systematic review with a high risk of bias (Brefka 2022). The committee did not make any recommendations regarding UB2-CAM.

All included systematic reviews were considered to be fully applicable to the review protocol, although several of them included reference standards that were not specified in the review protocol as well as those that were. The committee were aware that it was often not possible to disambiguate the results that used a reference standard specified in the review protocol from results that used a different reference standard, particularly when results were pooled in the systematic review. They were also aware that the majority of data used a reference standard that was included in the protocol. The committee agreed that the possible impact of this would be to give less consistent point estimates and wider confidence intervals. In spite of this they agreed that the systematic reviews included in the evidence review were all fully applicable. The committee considered the evidence base as a whole and deemed the size and quality sufficient to make recommendations.

When considering the evidence base, the committee noted that a number of the screening tests had high median sensitivities and specificities derived from a range of systematic reviews and primary studies. Therefore, when comparing the tools, more emphasis was placed on the ability of a test to be used across multiple settings, by several occupations in a short amount of time.

The committee agreed that the evidence for 4AT demonstrated good median sensitivity (87% (IQR: 3.25%)) and specificity (88% (IQR: 0.5%)) derived from a relatively large number of primary studies (N=19) reported in 6 reviews, 2 at low risk of bias and 4 at high risk of bias. Additionally, the tool had been used in a wide range of settings, including both hospital (post-surgical, emergency and acute, ICU and stroke units) and long-term residential settings (nursing homes and daily care settings) and could be administered by a wide range of healthcare practitioners. Finally, the assessment could be done quickly (<2 min – <5 min) and would be suitable for use in time pressured environments. For these reasons, the committee recommended that if indicators of delirium are identified, a healthcare practitioner who is trained and competent should carry out an assessment based on 4AT.

Furthermore, the committee recommended the use of different tools for delirium screening in critical care and post-surgery recovery settings. Both CAM-ICU and ICDSC showed good median sensitivities and specificities when used by a range of healthcare practitioners. Moreover, the tools were both specifically designed for use in intensive care settings. Both tools ranged between 2 min – <5 min to administer, although one study did note that CAM-ICU may take up to 10 min when users are unfamiliar with the content. The committee determined a 2 min – <5 min administration time to be appropriate. Taking these factors into account the committee determined that either CAM-ICU or ICDSC should be recommended as delirium screening tools in critical care or in the recovery room after surgery use.

1.1.11.3 Benefits and harms

The committee noted that the previous wording of recommendation 1.6.1, in the 2010 version of the guideline, did not reflect current practice in the NHS. The recommendation focussed on how to carry out a clinical assessment to confirm diagnosis of delirium. They agreed that, in themselves, delirium assessment tools are not intended to definitively diagnose delirium (although they may be used by specialist healthcare professionals to support their diagnosis) but are to allow a broad range of healthcare practitioners to assess the indicators of delirium. If people meet the threshold in the assessment tool, then diagnosis should be made by a trained, competent healthcare professional, for example a geriatrician or a psychiatrist. They agreed that DSM-5 criteria did not need to be specified as the basis for diagnostic assessment, as this should be at the discretion of the healthcare professional conducting the diagnosis.

The committee split the previous recommendation into three distinct recommendations: one recommendation to address screening using a tool, one to address formal diagnosis by a relevant specialist, and a third recommendation for the part of the previous recommendation that covered delirium superimposed on dementia. They agreed that along with recommendation 1.5.1 this represented the 'best practice' diagnostic pathway for delirium – healthcare practitioners should be alert to signs of changes in behaviour (identified through observation, or through tools such as NEWS2 or SQiD) as recommended in 1.5.1 in the 2010 version of the guideline. If signs are found, this should be recorded and the person should undergo a formal assessment by a healthcare practitioner for delirium (1.6.1) and if this indicates delirium is likely to be present, a healthcare professional should make a final diagnosis. Due to this change in emphasis, the committee also changed the wording of recommendations 1.3.1 and 1.5.1, from the 2010 version of the guideline, to clarify that if changes in the patient that might indicate delirium are observed then a formal assessment should be carried out by a healthcare practitioner using an assessment tool (that they are trained and competent to use) before referring for final diagnosis by a healthcare professional.

The committee agreed that separating the previous recommendation 1.6.1, in the 2010 version of the guideline, into recommendations about screening and diagnosis would prevent delirium diagnoses from being based on a screening tool that was not intended to be used to make a final diagnosis. Furthermore, it would give a clearer escalation pathway from daily observations to formal screening then final diagnosis of delirium. The committee also recommended that healthcare practitioners can conduct the screening process, whereas a healthcare professional with the specialist skills to do so should give the final diagnosis.

The committee agreed that recommending delirium screening tools such as 4AT, CAM-ICU and ICDSC would benefit health practitioners by giving them an evidence-based assessment tool upon which to base their assessments. However, some members of the committee noted that this could suggest that health practitioners should only use the recommended tests. The committee agreed that the recommendation was not intended to restrict the use of other screening tools if they would be more appropriate in a given setting. Furthermore, the committee stated that healthcare professionals highly experienced in the diagnosis of delirium may not need to use a tool at all.

1.1.11.4 Cost effectiveness and resource use

Cost-effectiveness evidence was available from only one study (MacLullich et al, 2019), which was a prospective, double-blind diagnostic test accuracy study in emergency departments or in acute general medical wards in three UK sites. This study found that the use of the 4AT as a rapid delirium assessment tool was more cost effective compared with the CAM for patients aged ≥70 years as although the 4AT was associated with slightly fewer QALYs it was also less costly, and the ICER indicated cost-effectiveness. Although the quality of life (utility) values were not derived using preferred methods, these were estimated from a diverse sample of experts, and sensitivity analyses were conducted.

The committee considered this economic evidence on the 4AT and noted that the evidence was obtained from patients aged ≥70 years which could be a limitation as delirium can present in people much younger than this. Despite this limitation and the uncertainty indicated by the sensitivity analysis around the reported ICERs, the committee felt that the 4AT would be a cost-effective assessment tool and recommended it for use prior to a formal diagnosis of delirium.

The committee agreed that the new recommendations would improve resource use and more accurately reflect current practice and the intended use of the screening tools, so that rather than only being tools to support the formal diagnosis of delirium they could also be used by any healthcare practitioner (for example healthcare assistants) trained to use them as a quick and easy way to assess whether a person was likely to have delirium. They agreed this would free up healthcare professionals time to allow them to undertake other duties.

1.1.11.5 Other factors the committee took into account

The committee were aware that 4AT was a newer screening tool, and as such, the evidence base surrounding it was considerably smaller than a tool such as CAM. However, there is enough evidence in the extended literature to support the diagnostic accuracy and implementability of 4AT, and its ability to detect delirium superimposed on dementia. The committee noted recommendations in the NICE dementia guideline related to telling the difference between delirium and dementia. Furthermore, the committee noted certain nuances with CAM that can cause some issues when using it as a screening tool. Firstly, long CAM is too time consuming to be used as a screening tool and short CAM is not designed to be used as a standalone tool. Researchers have tried to improve these shortcomings by the introduction of new variants e.g. b-CAM, 3D-CAM etc. However, CAM generally relies on conditional logic, which makes it difficult to use in practice. For these reasons, the committee decided to remove CAM from recommendation 1.6.1 and replace it with 4AT.

The committee considered whether there would be any correlation between tools that had been derived from other tests, for example CAM-ICU being derived from CAM. The committee agreed that there were considerable differences between CAM and its variants and that correlation between the tests would be minimal.

The committee noted there was a lack of evidence regarding the implementation of delirium screening tools and their use in different patient groups, for example people who might struggle to understand or respond because they are not fluent English speakers. The evidence base did not indicate how easy or difficult delirium tools were to deliver across various settings or by different occupations. Furthermore, the committee were also aware of the difficulties in identifying delirium in people with dementia or the cognitive impairments, learning disabilities or affective disorders. Based on the evidence, they were unable to make recommendations about this and they did not make a consensus recommendation because they agreed that it was an area where further research was needed.. The committee made a research recommendation to address these gaps in the evidence (see appendix K).

Finally, the committee suggested a new recommendation to section 1.5 stating that any changes are documented in the person’s record (recommendation 1.5.2) to help ensure they are not overlooked, particularly in the residential sector where, in the committees experience, recording such observations could be haphazard. They agreed that this documentation was an important baseline to enable different staff (for example on a different shift) to be able to check whether their patients were showing any new changes that might indicate delirium.

1.1.12 Recommendations supported by this evidence review

This evidence review supports recommendations 1.5.2, 1.6.1 and 1.6.2 and the research recommendation on validating the tools in different settings and populations.

1.1.13 References – included studies

1.1.13.1 Diagnostic accuracy reviews

Aldwikat, Rami K, Manias, Elizabeth, Tomlinson, Emily et al. (2022) Delirium screening tools in the post-anaesthetic care unit: a systematic review and meta-analysis. Aging clinical and experimental research 34(6): 1225-1235

Brefka, Simone, Eschweiler, Gerhard Wilhelm, Dallmeier, Dhayana et al. (2022) Comparison of delirium detection tools in acute care : A rapid review. Zeitschrift fur Gerontologie und Geriatrie 55(2): 105-115

Calf, Agneta H, Pouw, Maaike A, van Munster, Barbara C et al. (2021) Screening instruments for cognitive impairment in older patients in the Emergency Department: a systematic review and meta-analysis. Age and ageing 50(1): 105-112

Chen, Ting-Jhen, Chung, Yi-Wei, Chang, Hui-Chen Rita et al. (2021) Diagnostic accuracy of the CAM-ICU and ICDSC in detecting intensive care unit delirium: A bivariate meta-analysis. International journal of nursing studies 113: 103782

Ho, MH, Montgomery, A, Traynor, V et al. (2020) Diagnostic Performance of Delirium Assessment Tools in Critically Ill Patients: A Systematic Review and Meta-Analysis. Worldviews on evidence-based nursing 17(4): 301-310

Ho, Mu-Hsing, Choi, Edmond Pui Hang, Chiu, Hsiao-Yean et al. (2022) Using the nursing delirium screening scale in assessing postoperative delirium: A meta-regression. Research in nursing & health 45(1): 23-33

Jeong, E; Park, J; Lee, J (2020) Diagnostic test accuracy of the Nursing Delirium Screening Scale: A systematic review and meta-analysis. Journal of advanced nursing 76(10): 2510-2521

Jeong, E; Park, J; Lee, J (2020) Diagnostic Test Accuracy of the 4AT for Delirium Detection: A Systematic Review and Meta-Analysis. International journal of environmental research and public health 17(20): 1-15

Kim, Sujeong, Choi, Eunju, Jung, Youngsun et al. (2021) Postoperative delirium screening tools for post-anaesthetic adult patients in non-intensive care units: A systematic review and meta-analysis. Journal of clinical nursing

Mansutti, I; Saiani, L; Palese, A (2019) Detecting delirium in patients with acute stroke: a systematic review of test accuracy. BMC neurology 19(1): 310

Park, J.; Jeong, E.; Lee, J. (2021) The Delirium Observation Screening Scale: A Systematic Review and Meta-Analysis of Diagnostic Test Accuracy. Clinical nursing research 30(4): 464-473

Patel, MB, Bednarik, J, Lee, P et al. (2018) Delirium Monitoring in Neurocritically Ill Patients: A Systematic Review. Critical care medicine 46(11): 1832-1841

Quispel-Aggenbach, DWP, Holtman, GA, Zwartjes, HAHT et al. (2018) Attention, arousal and other rapid bedside screening instruments for delirium in older patients: a systematic review of test accuracy studies. Age and ageing 47(5): 644-653

Rosgen, B, Krewulak, K, Demiantschuk, D et al. (2018) Validation of Caregiver-Centered Delirium Detection Tools: A Systematic Review. Journal of the American Geriatrics Society 66(6): 1218-1225

Tieges, Zoe, Maclullich, Alasdair M J, Anand, Atul et al. (2021) Diagnostic accuracy of the 4AT for delirium detection in older adults: systematic review and meta-analysis. Age and ageing 50(3): 733-743

van Velthuijsen, EL, Zwakhalen, SM, Warnier, RM et al. (2016) Psychometric properties and feasibility of instruments for the detection of delirium in older hospitalized patients: a systematic review. International journal of geriatric psychiatry 31(9): 974-89

Watt, Christine L, Scott, Mary, Webber, Colleen et al. (2021) Delirium screening tools validated in the context of palliative care: A systematic review. Palliative medicine 35(4): 683-696

1.1.13.2 Economic studies

MacLullich, Alasdair Mj, Shenkin, Susan D, Goodacre, Steve et al. (2019) The 4 'A's test for detecting delirium in acute medical patients: a diagnostic accuracy study. Health technology assessment (Winchester, England) 23(40): 1-194

Appendices

Appendix A – Review protocols

Review protocol for the diagnostic accuracy of diagnostic tests compared with the reference standard DSM-5 to identify delirium in people in hospital and long-term residential care settings

ID	Field	Content
0.	PROSPERO registration number	CRD42022353773
1.	Review title	The diagnostic accuracy of diagnostic tests compared with the reference standard DSM-5 to identify delirium in people in hospital and long-term residential care settings
2.	Review question	What is the diagnostic accuracy of diagnostic tests compared with the reference standard DSM-5, to identify delirium in people in hospital and long-term residential care settings?
3.	Objective	To determine whether newer diagnostic tests are superior to CAM and CAM-ICU as recommended in the current guideline.
4.	Searches	There will be a two-stage search process, first for SRs of the index tests, then a later search for cross-sectional studies of diagnostic accuracy for index tests where no SRs were identified. The initial search for SRs of index tests will be carried out in Epistemonikos and the Cochrane Database of Systematic Reviews. Epistemonikos is a large collection of systematic reviews, compiled from frequent searches of primary literature databases including PubMed, Embase, PsycINFO and Cinahl. It is currently estimated to incorporate around 97% of health-related systematic reviews with abstracts and 94% of those without (Rada et al, 2020). To allow for the possibility of a time lag between systematic reviews being published and being screened for inclusion in in Epistemonikos, we will conduct an additional, focused (high precision) search for systematic reviews published since 2021 in Medline, Embase and PsycInfo. Searches for systematic reviews will be stepped. In the first instance we will search for systematic reviews on all index tests from 2019-date. For index tests that are not included in a systematic review published since 2019, or where the search dates of systematic reviews do not reach back to 2010 (the date of the previous search) we will extend the search back to 2016.

ID	Field	Content
		For index tests with no available systematic review-level evidence the search for cross-sectional studies will be carried out in Medline, Embase, PsycInfo and CENTRAL. For the review of health economic evidence we will use the same basic search strategy as the main diagnostic test accuracy review with an additional filter, developed at NICE, which is designed to retrieve cost utility studies with high recall (Hubbard et al, in press). We will look for cost utility analyses from August 2009 (date of the previous guideline searches) to the present day in the following databases: Medline; Embase; PsycInfo; Econlit and the INAHTA international HTA database.
5.	Condition or domain being studied	Delirium
6.	Population	Adults (18 years and older) in: hospital, including surgical, medical, ICU, and accident and emergency departments long-term residential care settings Exclusion: People receiving end-of-life care (within the last few days of life) People with intoxication and/or withdrawing from drugs or alcohol, and people with delirium associated with these states.
7.	Intervention/Exposure/Test	Index tests, including the people operating them, subdivided by setting: 4AT Confusion Assessment Method Instrument (CAM) 3D CAM Delirium Observation Screening Scale Single Question to Identify Delirium (SQID) Recognizing acute delirium as part of your routine (RADAR) Intensive Care Delirium Screening Checklist (ICD-SC) CAM-ICU Brief CAM (B-CAM) Ultra Brief CAM (UB2-CAM) NuDESC
8.	Comparator/Reference standard/Confounding factors	DSM-IV/5 or ICD-10/11, applied by a trained specialist.

ID	Field	Content
9.	Types of study to be included	A 2-stage approach to addressing this question will be taken: A narrative review of systematic reviews for the index tests of interest For tests where no SRs are identified, primary cross-sectional studies will be searched. Review of reviews Systematic reviews of diagnostic studies. Review of primary studies Cross-sectional studies
10.	Other exclusion criteria	Non-English language Conference abstracts Theses/dissertations Primary studies that do not report data that can be used to easily establish 2x2 tables of diagnostic accuracy.
11.	Context	NICE is updating the guideline on delirium: prevention, diagnosis and management (CG103). The guideline was originally published in July 2010 and last updated in March 2019. It was developed as set out in the original scope (2008). New evidence about diagnostic tests for delirium suggests that recommendations on diagnosis (specialist clinical assessment) may need updating. Full details are set out in the 2020 exceptional surveillance review decision.
12.	Primary outcomes (critical outcomes)	Sensitivity and specificity Likelihood ratios
13.	Secondary outcomes (important outcomes)	Positive and negative predictive values if these are reported by SRs ROC/AUC, c-statistic Ease of use (for example, time taken, range of staff who can use)
14.	Data extraction (selection and coding)	All references identified by the searches and from other sources will be uploaded into EPPI reviewer and de-duplicated. 10% of the abstracts will be reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer. The full text of potentially eligible studies will be retrieved and will be assessed in line with the criteria outlined above. A standardised form will be used to extract data from studies (see Developing NICE guidelines: the manual section 6.4). Study investigators may be contacted for missing data where time and resources allow.
15.	Risk of bias (quality) assessment	Risk of bias will be assessed using the appropriate checklist as described in Developing NICE guidelines: the manual.

ID	Field	Content
		For systematic reviews of diagnostic studies the ROBIS tool will be used. For diagnostic test accuracy studies the QUADAS-2 tool will be used.
16.	Strategy for data synthesis	Systematic reviews Whole systematic reviews be used. Data for individual studies will not be extracted from the reviews. The results of all systematic reviews identified in the past 3 (or 6) years (see box 4. Searches) will be reported narratively by test and setting. The narrative will report the main outcomes of the systematic review alongside any assessment of confidence in the outcome (for example GRADE). Where GRADE is not reported, the RoB of included studies will be reported instead. A matrix will be constructed to show the spread of primary studies across systematic reviews for the same test to enable the committee to take into account the duplication of primary studies in several systematic reviews. Systematic reviews less than 5 years old will be considered to be up to date. For systematic reviews older than 5 years the committee will be asked whether they consider the results to be up to date or whether there has been significant development in that tool since 2015. Diagnostic test accuracy data for primary studies where SRs are not available for a particular test Meta-analysis of diagnostic accuracy data will be conducted with reference to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Deeks et al. 2010). Where five or more studies were available for all included strata, a bivariate model will be fitted using the mada package in R v3.4.0, which accounts for the correlations between positive and negative likelihood ratios, and between sensitivities and specificities. Where sufficient data are not available (2-4 studies), separate independent pooling will be performed for positive likelihood ratios, negative likelihood ratios, sensitivity and specificity, using Microsoft Excel. This approach is conservative as it is likely to somewhat underestimate test accuracy, due to failing to account for the correlation and trade-off between sensitivity and specificity (see Deeks 2010). Random-effects models (der Simonian and Laird) will be fitted for all syntheses, as recommended in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Deeks et al. 2010). Confidence in all outcomes will be assessed using modified GRADE for diagnostic studies.
17.	Analysis of sub-groups	If sub-groups of interest are reported in included SRs, they will be reported in the narrative. In primary studies of diagnostic test accuracy, where data can be disambiguated it will be stratified by sub-groups on interest. Sub groups of interest are Setting

ID	Field	Content		
		Age Comorbidity (dementia/depression) Language fluency (including communication problems resulting from disability etc)		
18.	Type and method of review	☐	Intervention	
		☒	Diagnostic	
		☐	Prognostic	
		☐	Qualitative	
		☐	Epidemiologic	
		☐	Service Delivery	
		☐	Other (please specify)	
19.	Language	English		
20.	Country	England		
21.	Anticipated or actual start date	15 July 2022		
22.	Anticipated completion date	9 February 2023		
23.	Stage of review at time of this submission	Review stage	Started	Completed
		Preliminary searches	☐	☐
		Piloting of the study selection process	☐	☐
		Formal screening of search results against eligibility criteria	☐	☐
		Data extraction	☐	☐
		Risk of bias (quality) assessment	☐	☐
		Data analysis	☐	☐
24.	Named contact	5a. Named contact Guideline Development Team B 5b Named contact e-mail deliriumdev@nice.org.uk 5e Organisational affiliation of the review National Institute for Health and Care Excellence (NICE) Guideline Development Team B		
25.	Review team members	From the Guideline Development Team B: Mr Chris Carmona, Technical Adviser Mr Giacomo De Guisa, Technical analyst Syed Mohiuddin, Technical Adviser (Economics Hannah Tebbs, Technical Analyst (Economics) Mr Tom Hudson, Information Specialist		
26.	Funding sources/sponsor	This systematic review is being completed by the Guideline Development Team B which is part of NICE.		

<table>
<tr><th>ID</th><th>Field</th><th colspan="2">Content</th></tr>
<tr><td>27.</td><td>Conflicts of interest</td><td colspan="2">All guideline committee members and anyone who has direct input into NICE guidelines (including the evidence review team and expert witnesses) must declare any potential conflicts of interest in line with NICE's code of practice for declaring and dealing with conflicts of interest. Any relevant interests, or changes to interests, will also be declared publicly at the start of each guideline committee meeting. Before each meeting, any potential conflicts of interest will be considered by the guideline committee Chair and a senior member of the development team. Any decisions to exclude a person from all or part of a meeting will be documented. Any changes to a member's declaration of interests will be recorded in the minutes of the meeting. Declarations of interests will be published with the final guideline.</td></tr>
<tr><td>28.</td><td>Collaborators</td><td colspan="2">Development of this systematic review will be overseen by an advisory committee who will use the review to inform the development of evidence-based recommendations in line with section 3 of Developing NICE guidelines: the manual. Members of the guideline committee are available on the NICE website: https://www.nice.org.uk/guidance/indevelopment/gid-ng10332</td></tr>
<tr><td>29.</td><td>Other registration details</td><td colspan="2"></td></tr>
<tr><td>30.</td><td>Reference/URL for published protocol</td><td colspan="2"></td></tr>
<tr><td>31.</td><td>Dissemination plans</td><td colspan="2">NICE may use a range of different methods to raise awareness of the guideline. These include standard approaches such as:
notifying registered stakeholders of publication
publicising the guideline through NICE's newsletter and alerts
issuing a press release or briefing as appropriate, posting news articles on the NICE website, using social media channels, and publicising the guideline within NICE.</td></tr>
<tr><td>32.</td><td>Keywords</td><td colspan="2">Delirium, Diagnostic test.</td></tr>
<tr><td>33.</td><td>Details of existing review of same topic by same authors</td><td colspan="2">None</td></tr>
<tr><td rowspan="5">34.</td><td rowspan="5">Current review status</td><td>☒</td><td>Ongoing</td></tr>
<tr><td>☐</td><td>Completed but not published</td></tr>
<tr><td>☐</td><td>Completed and published</td></tr>
<tr><td>☐</td><td>Completed, published and being updated</td></tr>
<tr><td>☐</td><td>Discontinued</td></tr>
<tr><td>35..</td><td>Additional information</td><td colspan="2"></td></tr>
<tr><td>36.</td><td>Details of final publication</td><td colspan="2">www.nice.org.uk</td></tr>
</table>

Appendix B – Literature search strategies

Background and development

Search design and peer review

A NICE information specialist conducted the literature searches for the evidence reviews. The searches were run on the 19th and 20th July 2022. This search report is compliant with the requirements of PRISMA-S (Rethlefsen et al, 2021).

The MEDLINE and Epistemonikos strategies below were quality assured (QA) by a trained NICE information specialist. All translated search strategies were peer reviewed to ensure their accuracy. Both procedures were adapted from the 2015 PRESS Checklist (McGowan et al, 2016)

The principal search strategy was developed in MEDLINE (Ovid interface) and adapted, as appropriate, for use in the other sources listed in the protocol, taking into account their size, search functionality and subject coverage.

Review management

The search results were managed in EPPI-Reviewer v5. Duplicates were removed in EPPI-R5 using a two-step process. First, automated deduplication is performed using a high-value algorithm. Second, manual deduplication is used to assess 'low-probability' matches. All decisions made for the review can be accessed via the deduplication history.

Prior work

The terminology used in the searches was informed by the searches carried out for the previous NICE delirium guideline (CG103, 2010, updated 2019), and the current SIGN guideline on risk reduction and management of delirium (SIGN 157, 2019). Additional terms were included to reflect the index tests mentioned in the review protocol.

Limits and restrictions

English language limits were applied in adherence to standard NICE practice and the review protocol.

Limits to exclude editorials, letters and conference abstracts were applied in adherence to standard NICE practice and the review protocol.

The Epistemonikos and Cochrane systematic review searches for the main diagnostic test accuracy review were limited to references published from 2016. The searches for systematic reviews in Medline ALL, Embase and PsycInfo were limited to references published from 2021, as defined in the review protocol.

Searches for the review of cost-effectiveness (cost-utility) studies were limited to references added to bibliographic databases since 13th August 2009, where possible, so as to lead on from the previous NICE guideline searches.

Search filters

Diagnostic test accuracy searches

The systematic review filter used in MEDLINE was the “Health-evidence.ca Systematic review search filter” from Lee et al. (2012).

The standard NICE modifications were used: pubmed.tw added; systematic review.pt added from MeSH update 2019.

The Embase systematic review filter was the “Health-evidence.ca Systematic review search filter”, also from Lee et al. (2012).

The standard NICE modifications were used: pubmed.tw added to line medline.tw.

A similar approach to the Lee et al filters was adopted for PsycInfo but this set of search terms has not previously been formally validated.

Cost effectiveness searches

The searches for the review of cost-effectiveness studies used filters developed at NICE. In testing the Medline and Embase versions retrieved all of a validation set containing 370 cost-utility study references, which had previously been included in NICE evidence reviews. They are currently awaiting publication as Hubbard et al, 2022.

Key decisions

Note that the search for diagnostic test accuracy studies was designed to find recent systematic reviews that cover the index tests mentioned in the review protocol. The search of Epistemonikos can be seen as the core of the search approach, given Epistemonikos’ extensive coverage of systematic reviews from multiple sources (Rada et al, 2020). The searches of Medline, Embase and PsycInfo were designed to supplement to the Epistemonikos search by covering any indexing lag, rather than as comprehensive, stand-alone approaches. Further details of the rationale for this approach are given in the review protocol.

References

Hubbard W, Walsh N, Hudson T et al (2022) Development and validation of paired MEDLINE and Embase search filters for cost-utility studies *[submitted manuscript].*

Lee E, Dobbins M, DeCorby K et al (2012) An optimal search filter for retrieving systematic reviews and meta-analyses. *BMC Medical Research Methodology,* 12 (1) 51.

McGowan J, Sampson M, Salzwedel DM et al (2016) PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. *Journal of Clinical Epidemiology,* 75 p40-6.

Rada G, Pérez D, Araya-Quintanilla F et al (2020) Epistemonikos: a comprehensive database of systematic reviews for health decision-making. *BMC Medical Research Methodology,* 20 (286).

Rethlefsen ML, Kirtley S, Waffenschmidt S (2021) PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews,* 10 (39).

Diagnostic test accuracy searches

Main search – Databases

Database	Date searched	Database platform	Database segment or version	No. of results downloaded
Epistemonikos	19th July 2022	-	-	336
Cochrane Database of Systematic Reviews	19th July 2022	Wiley	Issue 7 of 12, July 2022	70
Medline ALL	19th July 2022	Ovid	1946 to 18th July 2022	138
Embase	19th July 2022	Ovid	1974 to 18th July 2022	158
PsycInfo	19th July 2022	Ovid	1806 to July week 2, 2022	13

Search strategy history

Database name: Epistemonikos

Strategy as single block of text (pasted into Epistemonikos using the option to run as title/abstract search). Filters to limit to systematic reviews, published from 2016, were applied on-screen.

(deliri* OR (acute* AND confus*) OR (confus* AND state*) OR (acute* AND "brain syndrome") OR (acute* AND "brain failure") OR "organic psychosyndrome" OR "psychoorganic syndrome" OR "psycho-organic syndrome" OR "toxic confusion" OR "toxic confusional" OR "toxic psychosis") AND (("4AT" OR "4 AT" OR "4-AT" OR 4AST OR "4AS" OR "4 AS" OR "4-AS" OR "4 A S test" OR "4-A S test" OR "4 'A's Test" OR "4-'A's Test" OR "4 assessment test" OR "4-assessment test" OR "4 assessments test" OR "4-assessments test") OR ("confusion assessment method" OR CAM OR CAMICU OR "CAM-ICU" OR 3DCAM OR "3-D-CAM" OR "3D-CAM" OR BCAM OR "B-CAM" OR UBCAM OR "UB-CAM" OR UB2CAM OR "UB2-CAM") OR ("delirium observation screening scale" OR DOSS) OR ("Single Question" OR SQID OR squid) OR ("recognizing acute delirium" OR "recognising acute delirium" OR RADAR) OR ("intensive care delirium screening checklist" OR ICDSC OR "ICD SC" OR "ICD-SC") OR ("nursing delirium screening scale" OR ndss OR NUDESC OR "NU DESC" OR "NU-DESC" OR NDESC OR "N DESC" OR "N-DESC") OR (scale OR scales OR score OR scores OR tool OR tools OR sensitivity OR specificity OR "likelihood ratio" OR "likelihood ratios" OR "positive predictive value" OR "positive predictive values" OR "negative predictive value" "negative predictive values" OR PPV OR NPV OR "roc curve" OR "roc curves" OR usability OR "easy to use" OR "easy-to-use" OR "ease of use" OR "ease-of-use" OR "diagnostic test accuracy" OR dta))

Easier-to-read version of the above strategy

(deliri* OR

(acute* AND confus*) OR

(confus* AND state*) OR

(acute* AND "brain syndrome") OR

(acute* AND "brain failure") OR

"organic psychosyndrome" OR

"psychoorganic syndrome" OR

"psycho-organic syndrome" OR

"toxic confusion" OR

"toxic confusional" OR

"toxic psychosis")

AND

(("4AT" OR "4 AT" OR "4-AT" OR 4AST OR "4AS" OR "4 AS" OR "4-AS" OR "4 A S test" OR "4-A S test" OR "4 'A's Test" OR "4-'A's Test" OR "4 assessment test" OR "4-assessment test" OR "4 assessments test" OR "4-assessments test") OR

("confusion assessment method" OR CAM OR CAMICU OR "CAM-ICU" OR 3DCAM OR "3-D-CAM" OR "3D-CAM" OR BCAM OR "B-CAM" OR UBCAM OR "UB-CAM" OR UB2CAM OR "UB2-CAM") OR

("delirium observation screening scale" OR DOSS) OR

("Single Question" OR SQID OR squid) OR

("recognizing acute delirium" OR "recognising acute delirium" OR RADAR) OR

("intensive care delirium screening checklist" OR ICDSC OR "ICD SC" OR "ICD-SC") OR

("nursing delirium screening scale" OR ndss OR NUDESC OR "NU DESC" OR "NU-DESC" OR NDESC OR "N DESC" OR "N-DESC") OR

(scale OR scales OR score OR scores OR tool OR tools OR sensitivity OR specificity OR "likelihood ratio" OR "likelihood ratios" OR "positive predictive value" OR "positive predictive values" OR "negative predictive value" "negative predictive values" OR PPV OR NPV OR "roc curve" OR "roc curves" OR usability OR "easy to use" OR "easy-to-use" OR "ease of use" OR "ease-of-use" OR "diagnostic test accuracy" OR dta))

As interpreted by Epistemonikos, with filters applied

(advanced_title_en:((deliri* OR (acute* AND confus*) OR (confus* AND state*) OR (acute* AND "brain syndrome") OR (acute* AND "brain failure") OR "organic psychosyndrome" OR "psychoorganic syndrome" OR "psycho-organic syndrome" OR "toxic confusion" OR "toxic confusional" OR "toxic psychosis") AND (("4AT" OR "4 AT" OR "4-AT" OR 4AST OR "4AS" OR "4 AS" OR "4-AS" OR "4 A S test" OR "4-A S test" OR "4 'A's Test" OR "4-'A's Test" OR "4 assessment test" OR "4-assessment test" OR "4 assessments test" OR "4-assessments test") OR ("confusion assessment method" OR CAM OR CAMICU OR "CAM-ICU" OR 3DCAM OR "3-D-CAM" OR "3D-CAM" OR BCAM OR "B-CAM" OR UBCAM OR "UB-CAM" OR UB2CAM OR "UB2-CAM") OR ("delirium observation screening scale" OR DOSS) OR ("Single Question" OR SQID OR squid) OR ("recognizing acute delirium" OR "recognising acute delirium" OR RADAR) OR ("intensive care delirium screening checklist" OR ICDSC OR "ICD SC" OR "ICD-SC") OR ("nursing delirium screening scale" OR ndss OR NUDESC OR "NU DESC" OR "NU-DESC" OR NDESC OR "N DESC" OR "N-DESC") OR (scale OR scales

OR score OR scores OR tool OR tools OR sensitivity OR specificity OR "likelihood ratio" OR "likelihood ratios" OR "positive predictive value" OR "positive predictive values" OR "negative predictive value" "negative predictive values" OR PPV OR NPV OR "roc curve" OR "roc curves" OR usability OR "easy to use" OR "easy-to-use" OR "ease of use" OR "ease-of-use" OR "diagnostic test accuracy" OR dta))) OR advanced_abstract_en:((deliri* OR (acute* AND confus*) OR (confus* AND state*) OR (acute* AND "brain syndrome") OR (acute* AND "brain failure") OR "organic psychosyndrome" OR "psychoorganic syndrome" OR "psycho-organic syndrome" OR "toxic confusion" OR "toxic confusional" OR "toxic psychosis") AND (("4AT" OR "4 AT" OR "4-AT" OR 4AST OR "4AS" OR "4 AS" OR "4-AS" OR "4 A S test" OR "4-A S test" OR "4 'A's Test" OR "4-'A's Test" OR "4 assessment test" OR "4-assessment test" OR "4 assessments test" OR "4-assessments test") OR ("confusion assessment method" OR CAM OR CAMICU OR "CAM-ICU" OR 3DCAM OR "3-D-CAM" OR "3D-CAM" OR BCAM OR "B-CAM" OR UBCAM OR "UB-CAM" OR UB2CAM OR "UB2-CAM") OR ("delirium observation screening scale" OR DOSS) OR ("Single Question" OR SQID OR squid) OR ("recognizing acute delirium" OR "recognising acute delirium" OR RADAR) OR ("intensive care delirium screening checklist" OR ICDSC OR "ICD SC" OR "ICD-SC") OR ("nursing delirium screening scale" OR ndss OR NUDESC OR "NU DESC" OR "NU-DESC" OR NDESC OR "N DESC" OR "N-DESC") OR (scale OR scales OR score OR scores OR tool OR tools OR sensitivity OR specificity OR "likelihood ratio" OR "likelihood ratios" OR "positive predictive value" OR "positive predictive values" OR "negative predictive value" "negative predictive values" OR PPV OR NPV OR "roc curve" OR "roc curves" OR usability OR "easy to use" OR "easy-to-use" OR "ease of use" OR "ease-of-use" OR "diagnostic test accuracy" OR dta)))) [Filters: classification=systematic-review, protocol=no, min_year=2016, max_year=2022]

Database name: Cochrane Database of Systematic Reviews

ID	Search
#1	[mh delirium]
#2	[mh ^confusion]
#3	deliri*:ti,ab
#4	(confus* NEAR/3 state*):ti,ab
#5	(acute* NEAR/3 (confus* or "brain syndrome" or "brain failure")):ti,ab
#6	"organic psychosyndrome":ti,ab
#7	"psychoorganic syndrome":ti,ab
#8	"psycho-organic syndrome":ti,ab
#9	(toxic NEXT confusion*):ti,ab
#10	"toxic psychosis":ti,ab
#11	{or #1-#10}
#12	("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or (4 NEXT assessment* NEXT test)):ti,ab
#13	("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM"):ti,ab
#14	("delirium observation screening scale" or DOSS):ti,ab
#15	("Single Question" or SQID or squid):ti,ab
#16	("recognizing acute delirium" or "recognising acute delirium" or RADAR):ti,ab
#17	("intensive care delirium screening checklist" or ICDSC or "ICD-SC"):ti,ab
#18	("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC"):ti,ab
#19	{or #12-#18}
#20	((assess* or screen* or diagnos* or identif* or check*) NEAR/2 (scale or scales or score or scores or tool or tools)):ti,ab
#21	[mh ^"Mass Screening"]
#22	[mh ^"Psychiatric Status Rating Scales"]
#23	[mh ^"Sensitivity and Specificity"]

```
#24	sensitivity:ti,ab
#25	specificity:ti,ab
#26	(likelihood NEAR ratio*):ti,ab
#27	((positive NEXT predictive NEXT value*) or (negative NEXT predictive NEXT value*) or PPV or NPV):ti,ab
#28	[mh ^"predictive value of tests"]
#29	[mh ^"roc curve"]
#30	(roc NEXT curve*):ti,ab
#31	usability:ti,ab
#32	((easy or ease) NEAR/2 "use"):ti,ab
#33	("diagnostic test accuracy" or dta):ti,ab
#34	[mh ^diagnosis]
#35	[mh ^"diagnostic techniques and procedures"]
#36	[mh ^"diagnostic tests, routine"]
#37	{or #20-#36}
#38	#19 or #37
#39	#11 and #38
```

Database name: Medline ALL

1 exp Delirium/ (11752)
2 Confusion/ (4997)
3 deliri*.tw. (19133)
4 (confus* adj3 state*).tw. (1666)
5 (acute* adj3 (confus* or "brain syndrome" or "brain failure")).tw. (1515)
6 organic psychosyndrome.tw. (110)
7 psychoorganic syndrome.tw. (83)
8 psycho-organic syndrome.tw. (105)
9 toxic confusion*.tw. (27)
10 toxic psychosis.tw. (159)
11 or/1-10 (27571)
12 ("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or "4 assessment* test").tw. (3145379)
13 ("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM").tw. (31778)
14 ("delirium observation screening scale" or DOSS).tw. (380)
15 ("Single Question" or SQID or squid).tw. (7697)
16 ("recognizing acute delirium" or "recognising acute delirium" or RADAR).tw. (6530)
17 ("intensive care delirium screening checklist" or ICDSC or "ICD-SC").tw. (192)
18 ("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC").tw. (254)
19 or/12-18 (3185954)
20 ((assess* or screen* or diagnos* or identif* or check*) adj2 (scale or scales or score or scores or tool or tools)).tw. (222409)
21 Mass Screening/ (114130)
22 Psychiatric Status Rating Scales/ (79287)
23 "Sensitivity and Specificity"/ (365415)
24 sensitivity.tw. (916751)
25 specificity.tw. (523767)
26 likelihood ratio*.tw. (18284)
27 ("positive predictive value*" or "negative predictive value*" or PPV or NPV).tw. (84690)
28 "predictive value of tests"/ (221241)
29 roc curve/ (69203)
30 "roc curve*".tw. (47637)
31 usability.tw. (18153)
32 ((easy or ease) adj2 "use").tw. (32281)
33 ("diagnostic test accuracy" or dta).tw. (4823)
34 diagnosis/ or "diagnostic techniques and procedures"/ or diagnostic tests, routine/ (35852)
35 or/20-34 (2017377)
36 19 or 35 (4947656)
37 11 and 36 (6054)
38 limit 37 to (editorial or letter) (125)
39 37 not 38 (5929)
40 limit 39 to english language (5461)
41 (MEDLINE or pubmed).tw. (283696)
42 systematic review.tw. (230128)
43 systematic review.pt. (202559)
44 meta-analysis.pt. (164608)
45 intervention$.ti. (182990)
46 or/41-45 (608378)
47 40 and 46 (450)

48 limit 47 to yr="2021 -Current" (138)

Database name: Embase

1 delirium assessment/ (75)
2 deliri*.tw. (29094)
3 delirium/ or emergence agitation/ or hyperactive delirium/ or hypoactive delirium/ or postoperative delirium/ (35614)
4 (confus* adj3 state*).tw. (2653)
5 (acute* adj3 (confus* or "brain syndrome" or "brain failure")).tw. (2372)
6 organic psychosyndrome.tw. (179)
7 psychoorganic syndrome.tw. (135)
8 psycho-organic syndrome.tw. (116)
9 toxic confusion*.tw. (33)
10 toxic psychosis.tw. (153)
11 or/2-10 (45527)
12 ("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or "4 assessment* test").tw. (4480742)
13 ("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM").tw. (39154)
14 exp confusion assessment method/ (614)
15 ("delirium observation screening scale" or DOSS).tw. (548)
16 ("Single Question" or SQID or squid).tw. (7810)
17 ("recognizing acute delirium" or "recognising acute delirium" or RADAR).tw. (6149)
18 ("intensive care delirium screening checklist" or ICDSC or "ICD-SC").tw. (336)
19 ("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC").tw. (354)
20 nursing delirium screening scale/ (82)
21 or/12-20 (4526900)
22 ((assess* or screen* or diagnos* or identif* or check*) adj2 (scale or scales or score or scores or tool or tools)).tw. (323395)
23 sensitivity.tw. (1181471)
24 specificity.tw. (675446)
25 "sensitivity and specificity"/ (437620)
26 likelihood ratio*.tw. (24808)
27 ("positive predictive value*" or "negative predictive value*" or PPV or NPV).tw. (130241)
28 predictive value/ (218664)
29 "roc curve*".tw. (80460)
30 receiver operating characteristic/ (172713)
31 usability.tw. (22200)
32 usability/ (2378)
33 ((easy or ease) adj2 "use").tw. (43913)
34 ("diagnostic test accuracy" or dta).tw. (6317)
35 diagnostic accuracy/ (284731)
36 or/22-35 (2326417)
37 21 or 36 (6500902)
38 11 and 37 (11602)
39 1 or 38 (11627)
40 (MEDLINE or pubmed).tw. (351496)
41 exp systematic review/ or systematic review.tw. (425275)
42 meta-analysis/ (251036)
43 intervention$.ti. (240544)
44 or/40-43 (846973)
45 39 and 44 (839)
46 45 (839)
47 limit 46 to yr="2021 -Current" (195)
48 (conference abstract* or conference review or conference paper or conference proceeding).db,pt,su. (5233367)

49 47 not 48 (161)
50 limit 49 to english language (159)
51 limit 50 to (editorial or letter) (1)
52 50 not 51 (158)

Database name: PsycInfo

1 delirium/ (3807)
2 mental confusion/ (1212)
3 deliri*.tw. (8124)
4 (confus* adj3 state*).tw. (1050)
5 (acute* adj3 (confus* or "brain syndrome" or "brain failure")).tw. (530)
6 organic psychosyndrome.tw. (63)
7 psychoorganic syndrome.tw. (33)
8 psycho-organic syndrome.tw. (35)
9 toxic confusion*.tw. (16)
10 toxic psychosis.tw. (110)
11 or/1-10 (10274)
12 ("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or "4 assessment* test").tw. (407879)
13 ("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM").tw. (2827)
14 ("delirium observation screening scale" or DOSS).tw. (106)
15 ("Single Question" or SQID or squid).tw. (730)
16 ("recognizing acute delirium" or "recognising acute delirium" or RADAR).tw. (1056)
17 ("intensive care delirium screening checklist" or ICDSC or "ICD-SC").tw. (26)
18 ("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC").tw. (87)
19 or/12-18 (412231)
20 diagnosis/ or exp psychodiagnosis/ (87477)
21 screening/ or exp screening tests/ (17994)
22 ((assess* or screen* or diagnos* or identif* or check*) adj2 (scale or scales or score or scores or tool or tools)).tw. (61968)
23 sensitivity.tw. (104120)
24 Test Sensitivity/ (378)
25 specificity.tw. (41442)
26 Test Specificity/ (306)
27 likelihood ratio*.tw. (2212)
28 ("positive predictive value*" or "negative predictive value*" or PPV or NPV).tw. (3446)
29 "roc curve*".tw. (2673)
30 usability.tw. (6509)
31 ((easy or ease) adj2 "use").tw. (6479)
32 ("diagnostic test accuracy" or dta).tw. (210)
33 exp Test Validity/ or exp Diagnosis/ or exp Test Reliability/ (339512)
34 or/20-33 (511058)
35 19 or 34 (874257)
36 11 and 35 (3039)
37 limit 36 to (editorial or letter) (98)
38 36 not 37 (2941)
39 limit 38 to english language (2512)
40 (MEDLINE or pubmed).tw. (29725)
41 systematic review.tw. (38158)
42 intervention$.ti. (81883)
43 "systematic review"/ or meta analysis/ (5829)
44 or/40-43 (134279)
45 39 and 44 (123)
46 limit 45 to yr="2021 - 2022" (13)

Cost-effectiveness searches

Main search – Databases

Database	Date searched	Database Platform	Database segment or version	No. of results downloaded
Medline ALL	20th July 2022	Ovid	1946 to 19th July 2022	100
Embase	20th July 2022	Ovid	1974 to 19th July 2022	122
PsycInfo	20th July 2022	Ovid	1806 to July Week 2, 2022	12
Econlit	20th July 2022	Ovid	1886 to 14th July 2022	0
INAHTA International HTA database	20th July 2022	-	-	13

Search strategy history

Database name: Medline ALL

1 exp Delirium/ (11784)
2 Confusion/ (4998)
3 deliri*.tw. (19179)
4 (confus* adj3 state*).tw. (1667)
5 (acute* adj3 (confus* or "brain syndrome" or "brain failure")).tw. (1516)
6 organic psychosyndrome.tw. (110)
7 psychoorganic syndrome.tw. (83)
8 psycho-organic syndrome.tw. (105)
9 toxic confusion*.tw. (27)
10 toxic psychosis.tw. (159)
11 or/1-10 (27619)
12 ("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or "4 assessment* test").tw. (3150392)
13 ("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM").tw. (31843)
14 ("delirium observation screening scale" or DOSS).tw. (382)
15 ("Single Question" or SQID or squid).tw. (7705)
16 ("recognizing acute delirium" or "recognising acute delirium" or RADAR).tw. (6563)
17 ("intensive care delirium screening checklist" or ICDSC or "ICD-SC").tw. (193)
18 ("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC").tw. (256)
19 or/12-18 (3191063)
20 ((assess* or screen* or diagnos* or identif* or check*) adj2 (scale or scales or score or scores or tool or tools)).tw. (223101)
21 Mass Screening/ (114252)
22 Psychiatric Status Rating Scales/ (79302)
23 "Sensitivity and Specificity"/ (365641)

24 sensitivity.tw. (918870)
25 specificity.tw. (524903)
26 likelihood ratio*.tw. (18328)
27 ("positive predictive value*" or "negative predictive value*" or PPV or NPV).tw. (84884)
28 "predictive value of tests"/ (221287)
29 roc curve/ (69321)
30 "roc curve*".tw. (48011)
31 usability.tw. (18240)
32 ((easy or ease) adj2 "use").tw. (32401)
33 ("diagnostic test accuracy" or dta).tw. (4842)
34 diagnosis/ or "diagnostic techniques and procedures"/ or diagnostic tests, routine/ (35868)
35 or/20-34 (2021163)
36 19 or 35 (4956129)
37 11 and 36 (6071)
38 limit 37 to (editorial or letter) (126)
39 37 not 38 (5945)
40 limit 39 to english language (5477)
41 Cost-Benefit Analysis/ (90447)
42 Quality-Adjusted Life Years/ (15027)
43 Markov Chains/ (15766)
44 exp Models, Economic/ (16133)
45 cost*.ti. (136633)
46 (cost* adj2 utilit*).tw. (7042)
47 (cost* adj2 (effective* or assess* or evaluat* or analys* or model* or benefit* or threshold* or quality or expens* or saving* or reduc*)).tw. (252998)
48 (economic* adj2 (evaluat* or assess* or analys* or model* or outcome* or benefit* or threshold* or expens* or saving* or reduc*)).tw. (42484)
49 (qualit* adj2 adjust* adj2 life*).tw. (16232)
50 QALY*.tw. (13088)
51 (incremental* adj2 cost*).tw. (15828)
52 ICER.tw. (5296)
53 utilities.tw. (8566)
54 markov*.tw. (29197)
55 (dollar* or USD or cents or pound or pounds or GBP or sterling* or pence or euro or euros or yen or JPY).tw. (50687)
56 ((utility or effective*) adj2 analys*).tw. (22864)
57 (willing* adj2 pay*).tw. (8636)
58 (EQ5D* or EQ-5D*).tw. (11651)
59 ((euroqol or euro-qol or euroquol or euro-quol or eurocol or euro-col) adj3 ("5" or five)).tw. (3289)
60 (european* adj2 quality adj3 ("5" or five)).tw. (600)
61 or/41-60 (463116)
62 40 and 61 (122)
63 limit 62 to dt=20090813-20220720 (100)

Database name: Embase

1 delirium assessment/ (75)
2 deliri*.tw. (29106)
3 delirium/ or emergence agitation/ or hyperactive delirium/ or hypoactive delirium/ or postoperative delirium/ (35626)
4 (confus* adj3 state*).tw. (2653)
5 (acute* adj3 (confus* or "brain syndrome" or "brain failure")).tw. (2372)
6 organic psychosyndrome.tw. (179)
7 psychoorganic syndrome.tw. (135)
8 psycho-organic syndrome.tw. (116)
9 toxic confusion*.tw. (33)
10 toxic psychosis.tw. (153)
11 or/2-10 (45539)
12 ("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or "4 assessment* test").tw. (4481731)
13 ("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM").tw. (39162)
14 exp confusion assessment method/ (614)
15 ("delirium observation screening scale" or DOSS).tw. (548)
16 ("Single Question" or SQID or squid).tw. (7811)
17 ("recognizing acute delirium" or "recognising acute delirium" or RADAR).tw. (6151)
18 ("intensive care delirium screening checklist" or ICDSC or "ICD-SC").tw. (336)
19 ("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC").tw. (354)
20 nursing delirium screening scale/ (82)
21 or/12-20 (4527900)
22 ((assess* or screen* or diagnos* or identif* or check*) adj2 (scale or scales or score or scores or tool or tools)).tw. (323533)
23 sensitivity.tw. (1181803)
24 specificity.tw. (675622)
25 "sensitivity and specificity"/ (437738)
26 likelihood ratio*.tw. (24812)
27 ("positive predictive value*" or "negative predictive value*" or PPV or NPV).tw. (130281)
28 predictive value/ (218717)
29 "roc curve*".tw. (80509)
30 receiver operating characteristic/ (172840)
31 usability.tw. (22217)
32 usability/ (2392)
33 ((easy or ease) adj2 "use").tw. (43929)
34 ("diagnostic test accuracy" or dta).tw. (6318)
35 diagnostic accuracy/ (284756)
36 or/22-35 (2327037)
37 21 or 36 (6502443)
38 11 and 37 (11605)
39 1 or 38 (11630)
40 (conference abstract* or conference review or conference paper or conference proceeding).db,pt,su. (5234319)
41 39 not 40 (7054)
42 limit 41 to english language (6470)
43 limit 42 to (editorial or letter) (239)
44 42 not 43 (6231)
45 cost utility analysis/ (11248)
46 quality adjusted life year/ (32003)
47 cost*.ti. (181397)
48 (cost* adj2 utilit*).tw. (11483)

49 (cost* adj2 (effective* or assess* or evaluat* or analys* or model* or benefit* or threshold* or quality or expens* or saving* or reduc*)).tw. (349787)
50 (economic* adj2 (evaluat* or assess* or analys* or model* or outcome* or benefit* or threshold* or expens* or saving* or reduc*)).tw. (59720)
51 (qualit* adj2 adjust* adj2 life*).tw. (24539)
52 QALY*.tw. (24067)
53 (incremental* adj2 cost*).tw. (25845)
54 ICER.tw. (11473)
55 utilities.tw. (13713)
56 markov*.tw. (36207)
57 (dollar* or USD or cents or pound or pounds or GBP or sterling* or pence or euro or euros or yen or JPY).tw. (66090)
58 ((utility or effective*) adj2 analys*).tw. (34037)
59 (willing* adj2 pay*).tw. (12892)
60 (EQ5D* or EQ-5D*).tw. (22412)
61 ((euroqol or euro-qol or euroquol or euro-quol or eurocol or euro-col) adj3 ("5" or five)).tw. (4368)
62 (european* adj2 quality adj3 ("5" or five)).tw. (819)
63 or/45-62 (576877)
64 44 and 63 (138)
65 limit 64 to dc=20090813-20220720 (122)

Database name: PsycInfo

1 delirium/ (3807)
2 mental confusion/ (1212)
3 deliri*.tw. (8124)
4 (confus* adj3 state*).tw. (1050)
5 (acute* adj3 (confus* or "brain syndrome" or "brain failure")).tw. (530)
6 organic psychosyndrome.tw. (63)
7 psychoorganic syndrome.tw. (33)
8 psycho-organic syndrome.tw. (35)
9 toxic confusion*.tw. (16)
10 toxic psychosis.tw. (110)
11 or/1-10 (10274)
12 ("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or "4 assessment* test").tw. (407879)
13 ("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM").tw. (2827)
14 ("delirium observation screening scale" or DOSS).tw. (106)
15 ("Single Question" or SQID or squid).tw. (730)
16 ("recognizing acute delirium" or "recognising acute delirium" or RADAR).tw. (1056)
17 ("intensive care delirium screening checklist" or ICDSC or "ICD-SC").tw. (26)
18 ("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC").tw. (87)
19 or/12-18 (412231)
20 diagnosis/ or exp psychodiagnosis/ (87477)
21 screening/ or exp screening tests/ (17994)
22 ((assess* or screen* or diagnos* or identif* or check*) adj2 (scale or scales or score or scores or tool or tools)).tw. (61968)
23 sensitivity.tw. (104120)
24 Test Sensitivity/ (378)
25 specificity.tw. (41442)
26 Test Specificity/ (306)
27 likelihood ratio*.tw. (2212)
28 ("positive predictive value*" or "negative predictive value*" or PPV or NPV).tw. (3446)
29 "roc curve*".tw. (2673)
30 usability.tw. (6509)
31 ((easy or ease) adj2 "use").tw. (6479)
32 ("diagnostic test accuracy" or dta).tw. (210)
33 exp Test Validity/ or exp Diagnosis/ or exp Test Reliability/ (339512)
34 or/20-33 (511058)
35 19 or 34 (874257)
36 11 and 35 (3039)
37 limit 36 to (editorial or letter) (98)
38 36 not 37 (2941)
39 limit 38 to english language (2512)
40 (cost* and ((qualit* adj2 adjust* adj2 life*) or qaly*)).tw. (1458)
41 ((incremental* adj2 cost*) or ICER).tw. (1342)
42 (cost adj2 utilit*).tw. (903)
43 (cost* and ((net adj benefit*) or (net adj monetary adj benefit*) or (net adj health adj benefit*))).tw. (389)
44 ((cost adj2 (effect* or utilit*)) and (quality adj of adj life)).tw. (2517)
45 (cost and (effect* or utilit*)).ti. (3382)
46 exp "quality of life measures"/ (750)
47 exp "Quality of Life"/ (48752)
48 exp "Costs and Cost Analysis"/ (46959)
49 health care economics/ (1093)
50 or/40-49 (97025)

51 39 and 50 (17)
52 limit 51 to up=20090813-20220720 (12)

Database name: Econlit

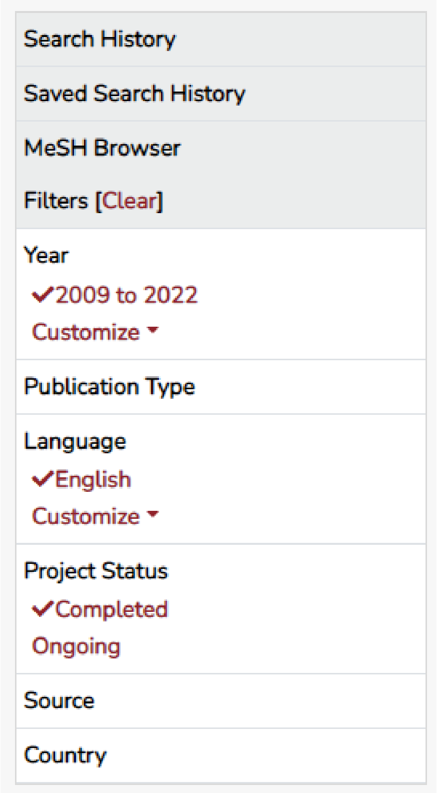
#	Searches	Results
1	deliri*.tw.	4
2	(confus* adj3 state*).tw.	27
3	(acute* adj3 (confus* or "brain syndrome" or "brain failure")).tw.	0
4	organic psychosyndrome.tw.	0
5	psychoorganic syndrome.tw.	0
6	psycho-organic syndrome.tw.	0
7	toxic confusion*.tw.	0
8	toxic psychosis.tw.	0
9	or/1-8	31
10	("4AT" or "4 AT" or 4AST or "4AS" or "4 AS" or "4 A S test" or "4 assessment* test").tw.	19389
11	("confusion assessment method" or CAM or CAMICU or 3DCAM or BCAM or UBCAM or "UB2CAM").tw.	103
12	("delirium observation screening scale" or DOSS).tw.	2
13	("Single Question" or SQID or squid).tw.	41
14	("recognizing acute delirium" or "recognising acute delirium" or RADAR).tw.	147
15	("intensive care delirium screening checklist" or ICDSC or "ICD-SC").tw.	0
16	("nursing delirium screening scale" or ndss or NUDESC or "NU-DESC" or NDESC or "N DESC").tw.	0
17	or/10-16	19669
18	((assess* or screen* or diagnos* or identif* or check*) adj2 (scale or scales or score or scores or tool or tools)).tw.	1604
19	sensitivity.tw.	14254
20	specificity.tw.	1835
21	likelihood ratio*.tw.	1413
22	("positive predictive value*" or "negative predictive value*" or PPV or NPV).tw.	609
23	"roc curve*".tw.	96
24	usability.tw.	292
25	((easy or ease) adj2 "use").tw.	600
26	("diagnostic test accuracy" or dta).tw.	62
27	or/18-26	20495
28	17 or 27	39696
29	9 and 28	3
30	(cost* and ((qualit* adj2 adjust* adj2 life*) or qaly*)).tw.	460
31	((incremental* adj2 cost*) or ICER).tw.	654
32	(cost adj2 utilit*).tw.	444
33	(cost* and ((net adj benefit*) or (net adj monetary adj benefit*) or (net adj health adj benefit*))).tw.	820
34	((cost adj2 (effect* or utilit*)) and (quality adj of adj life)).tw.	386
35	(cost and (effect* or utilit*)).ti.	2566
36	or/30-35	4143
37	29 and 36	0

Database name: INAHTA International HTA Database

(toxic psychosis) OR (toxic confusion*) OR (psycho organic syndrome) OR (psycho-organic syndrome) OR (psychoorganic syndrome) OR (organic psychosyndrome) OR (acute* AND (confus* OR "brain syndrome" OR "brain failure")) OR (confus* AND state*) OR (deliri*) OR ("Confusion"[mh]) OR ("Delirium"[mhe])

Search carried out in all fields (included titles, abstracts and MeSH headings)

On-screen limits used to limit to publication year, language and completed studies only.



Search History

Saved Search History

MeSH Browser

Filters [Clear]

Year
✓2009 to 2022
Customize ▾

Publication Type

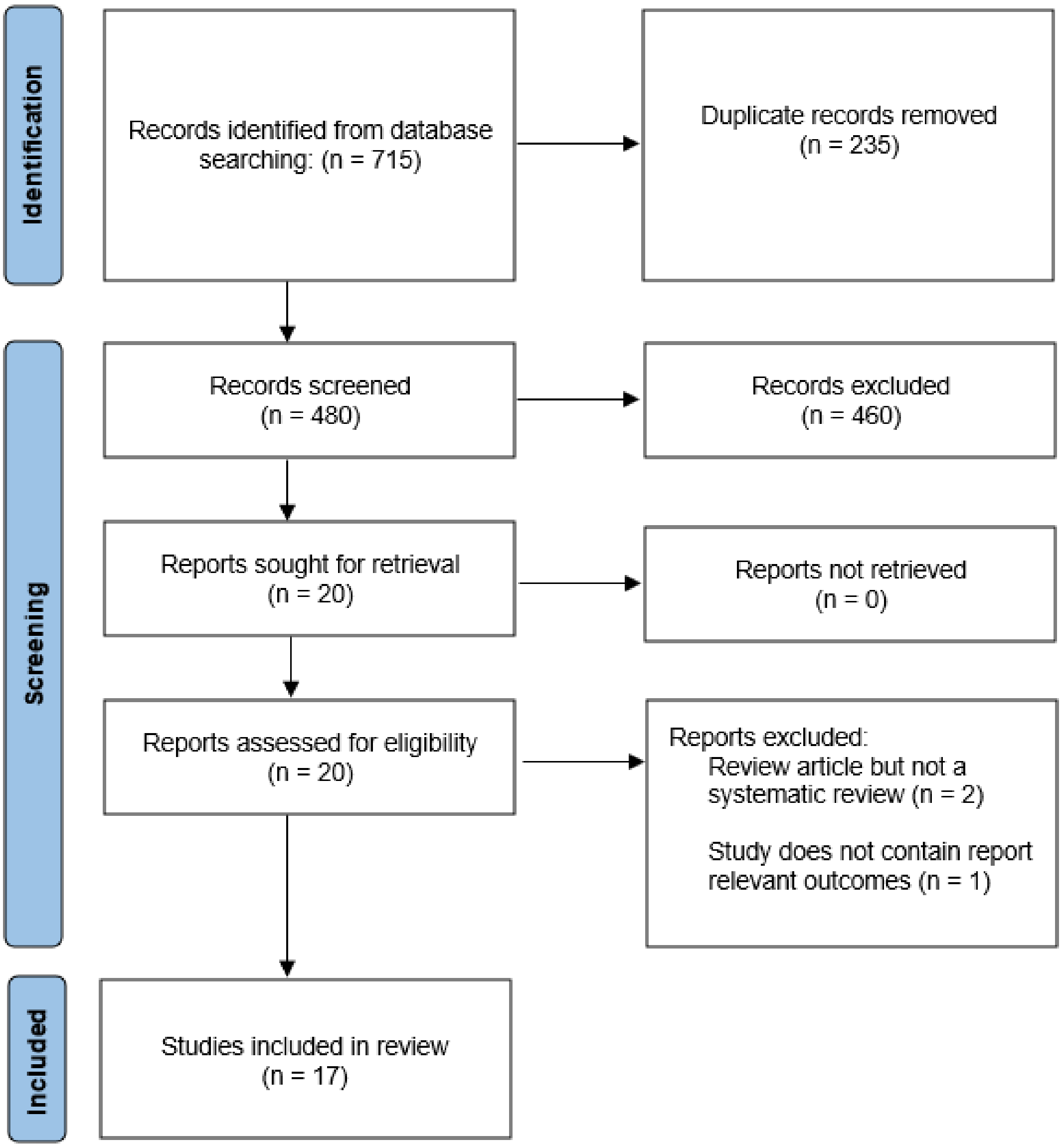
Language
✓English
Customize ▾

Project Status
✓Completed
Ongoing

Source

Country

Appendix C – Diagnostic evidence study selection


Identification
Records identified from database searching: (n = 715)
Duplicate records removed (n = 235)
Screening
Records screened (n = 480)
Records excluded (n = 460)
Reports sought for retrieval (n = 20)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 20)
Reports excluded:
Review article but not a systematic review (n = 2)
Study does not contain report relevant outcomes (n = 1)
Included
Studies included in review (n = 17)

Appendix D –Diagnostic evidence

Aldwikat, 2022

Bibliographic Reference	Aldwikat, Rami K; Manias, Elizabeth; Tomlinson, Emily; Amin, Mohammed; Nicholson, Patricia; Delirium screening tools in the post-anaesthetic care unit: a systematic review and meta-analysis.; Aging clinical and experimental research; 2022; vol. 34 (no. 6); 1225-1235

Study Characteristics

Study design	Systematic review
Countries of included studies	Germany, USA
Databases searched	• CINAHL • MEDLINE • Embase • PsycINFO • Scopus
Years searched	From database inception to September 2019 and again in April 2021
Inclusion criteria	• Studies concerning adults 18 years and over who were admitted to the PACU following surgery, including those with any pre-existing conditions • Studies that comprised evaluation of a delirium screening test which refers to the index test or the test to be evaluated ○ The tool must have been used at least once during patients' admission to the PACU by healthcare professionals, such as nurses, doctors and trained research assistants • Studies included in the review must have also used a reference standard test, which refers to the test whose results are considered the gold standard • Studies with research designs, such as prospective, retrospective cohort designs, randomised and non-randomised controlled trials, published in peer-reviewed journals with no language or time restrictions.
Exclusion criteria	Studies that were based on case reports, review articles, opinion papers or abstracts published in conference proceedings without full text
Number of studies included in the systematic review	4

Studies from the systematic review that are relevant for use in the current review	All studies
Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	Hospitals (post-anaesthetic care units (PACUs))
Patient population	Adults 18 years and over who were admitted to the PACU following surgery, including those with any pre-existing conditions
Tests included	4AT CAM 3D-CAM CAM-ICU NuDESC
Person delivering the test	Research assistants
Reference standard	• DSM-IV • DSM-5
Outcomes	CAM (Time to administer tool: 20 mins) (cut-off score: NR) • Sensitivity: 43% • Specificity: 98% NuDESC (Time to administer tool: 2-3 mins) (cut-off score: NR) Radtke 2008 • Sensitivity: 95% • Specificity: 87%

	Saller 2019 - Sensitivity: 27% - Specificity: 99% Nu-DESC (cut-off score: ≥ 2) - Sensitivity: 32% - Specificity: 92% Nu-DESC (cut-off score: ≥ 1) - Sensitivity: 80% - Specificity: 69% 3D-CAM (Time to administer tool: 3 min) (cut-off score: NR) - Sensitivity: 100% - Specificity: 88% 4AT (Time to administer tool: 2 min) (cut-off score: NR) - Sensitivity: 96% - Specificity: 99% CAM-ICU (Time to administer tool: 2 min) (cut-off score: NR) - Sensitivity: 28% - Specificity: 98%
Sources of funding	For this study, there was no funding source

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	High
Overall study ratings	Applicability as a source of data	Fully applicable

Brefka, 2022

Bibliographic Reference	Brefka, Simone; Eschweiler, Gerhard Wilhelm; Dallmeier, Dhayana; Denkinger, Michael; Leinert, Christoph; Comparison of delirium detection tools in acute care : A rapid review.; Zeitschrift fur Gerontologie und Geriatrie; 2022; vol. 55 (no. 2); 105-115

Study Characteristics

Study design	Systematic review
Countries of included studies	NR
Databases searched	• Medline • The Network for Investigation of Delirium: Unifying Scientists (NIDUS) website was used for an additional cross-reference search
Years searched	Between 2001 and 2021
Inclusion criteria	• Inclusion criteria were evaluation of assessment instruments in the context of delirium, English or German language and published between 2001 and 2021 • Publications were screened for existing tests and the psychometric properties. The cross-referenced literature of these reviews was examined in addition to the validation studies found to collect the primary literature of the respective assessments and extract further details • The Network for Investigation of Delirium: Unifying Scientists (NIDUS) website was used for an additional cross-reference search
Exclusion criteria	Not reported
Number of studies included in the systematic review	89
Studies from the systematic review that are relevant for use in the current review	• Bellelli 2014 • Inouye 1990 • Marcantonio 2014 • van Gemert 2007 • Koster 2009 • Schuurmans 2003 • Sands 2010 • Voyer 2015 • Bergeron 2001

	• Yu 2013 • Ely 2001 • Han 2013 • Armstrong 2021 • Husser 2021 • Motyl 2020 • Gaudreau 2005 (x2)
Studies from the systematic review that are not relevant for use in the current review	• Steis 2012 • Stillman 2000 • Grossmann 2014 • Hasemann 2018 • Han 2013 • O'Regan 2014 • Hasemann 2019 • Chester 2012 • Fick 2015 • Inouye 2014 • Lewis 1995 • Cacchione 2002 • Thomas 2012 • Dosa 2007 • O'Regan 2017 • Ní Chonchubhair 1995 • Linstedt 2002 • Lees 2013 • Vermeersch 1990 • Funk 1992 • Morita 2001 • Nagley 1986 • Emerson 2014 • Adamis 2005 • Inouye 2005 • Williams 1991 • Robertsson 1997 • Hart 1996 • O'keeffe 1994 • O'keeffe 1997 • Sala 1992 • Otter 2005 • Adamis 2016 • Kean 2010 • Franco 2020 • Tieges 2015 • Tieges 2020 • McCusker 2004 • Meagher 2008 • Grover 2013 • Garcia Nuñez 2019 • Meagher 2014 • de Jonghe 2005 • Trzepacz 1988 • Strub 1993 • Trzepacz 1999 • Rockwood 1996

	• Rosen 1994 • Trzepacz 2001 • Leung 2011 • O’Keeffe 1997 • Albert 1992 • Bettin 1998 • Christensen 1996 • Leonard 2016 • Rhodius-Meester 2013 • Salih 2012 • Breitbart 1997 • Neelon 1996 • van Gemert 2007 • Björkelund 2006 • Eriksson 2002 • Katzman 1983 • Richardson 2017 • Yadav 2020 • Han 2015 • Treloar 1997 • Miller 198 • Shulman 2016 • Erkinjuntti 1987 • Morandi 2012 • O’Regan 2014 • Lin 2015 • Hendry 2015 • Rosgen 2018 Studies did not concern a test of interest
Setting	Hospitals (focus on acute care and emergency settings)
Patient population	Older patients
Tests included	4AT CAM 3D-CAM DOS SQID RADAR ICDSC CAM-ICU

	B-CAM UB2-CAM NuDESC
Person delivering the test	• Physicians • Nurses • Trained lay-raters • Clinical staff / clinicians • Untrained geriatricians (Physicians and nurses were either trained or their training status was not specified)
Reference standard	• DSM-IV • 3D-CAM • CAM • Psychiatrist interview / diagnosis • Geriatric psychiatrist rating after comprehensive assessment • DSM-IV-TR
Outcomes	3D-CAM (Average duration: 3 min) (cut-off score: CAM algorithm:1+2+(3 or 4) positive = suspected delirium) Total sample • Sensitivity: 95% • Specificity: 94% Patients with dementia • Sensitivity: 96% • Specificity: 86% Patients without dementia • Sensitivity: 93% • Specificity: 96%

	CAM-ICU (Average duration: <5 min) (cut-off score: CAM algorithm: 1+2+(3 or 4) positive = suspected delirium) Total sample • Sensitivity: 95-100% • Specificity: 89-93% Patients ≥65 years • Sensitivity: 90-100% • Specificity: 83-100% Patients with dementia • Sensitivity: 100% • Specificity: 100% UB2-CAM (Average duration: 2 min) (cut-off score: CAM algorithm: 1+2+(3 or 4) positive = suspected delirium) • Sensitivity: 93% • Specificity: 95% 4AT (Average duration: <5 min) (cut-off score: ≥4 = possible delirium) • Sensitivity: 90% • Specificity: 84% • AUC: 0.89-0.93 DOS (Average duration: 5 min) (cut-off score: ≥3 = suspected delirium) Van Gemert 2007 • Sensitivity: 89% • Specificity: 88% Koster 2009

	• Sensitivity: 100% • Specificity: 97% NuDESC (Average duration: <2 min) (cut-off: ≥2 = suspected delirium) • Sensitivity: 86% • Specificity: 87% SQiD (Average duration: <1 min) (cut-off core: "yes" = suspected delirium) vs. psychiatrist interview • Sensitivity: 80% • Specificity: 71% vs. DSM-IV • Sensitivity: 77% • Specificity: 51% b-CAM (Average duration: <5 min) (cut-off score: CAM algorithm: 1+2+(3 or 4) positive = suspected delirium) • Sensitivity: 78-84% • Specificity: 96-97% CAM (Average duration: 5-10 min) (cut-off score: CAM algorithm: 1+2+(3 or 4) positive = suspected delirium) • Sensitivity: 94-100% • Specificity: 90-95% ICDSC (Average duration <5 min) (cut-off score: ≥4 = suspected delirium) • Sensitivity: 99% (estimated from ROC curve, at cut-off score of 4 points) • Specificity: 64% (estimated from ROC curve, at cut-off score of 4 points)

	RADAR (Average duration <1 min) (cut-off score: ≥1 “yes” = suspected delirium) • Sensitivity: 73% • Specificity: 67%
Sources of funding	Not reported

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	High
Overall study ratings	Applicability as a source of data	Fully applicable

Calf, 2021

Bibliographic Reference	Calf, Agneta H; Pouw, Maaike A; van Munster, Barbara C; Burgerhof, Johannes G M; de Rooij, Sophia E; Smidt, Nynke; Screening instruments for cognitive impairment in older patients in the Emergency Department: a systematic review and meta-analysis.; Age and ageing; 2021; vol. 50 (no. 1); 105-112

Study Characteristics

Study design	Systematic review
Countries of included studies	Germany, Brazil, Canada, USA, The Netherlands, UK
Databases searched	• MEDLINE • EMBASE • CINAHL • the Cochrane Central Register of Controlled trials (CENTRAL)
Years searched	Database inception to 3 March 2020
Inclusion criteria	• Cohort study or case-control study • Study population consisted of patients with a mean or median age 65 years or older, visiting an ED • The target condition was cognitive impairment irrespective of the aetiology. Ideally, the diagnosis was based on the Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria (version III, IV, IV-R, V) made by a specialist in geriatric care. The Confusion

	Assessment Method (CAM) and the Mini-Mental State Examination (MMSE) were accepted as a substitute gold standard because of their widely use in clinical practice • The index test was an instrument to assess cognition in the ED • The study provided sufficient data to construct a two-by-two table.
Exclusion criteria	Studies conducted in a different environment than the ED
Number of studies included in the systematic review	23 (14 for cognitive impairment specifically caused by delirium , 9 for cognitive impairment irrespective of the underlying aetiology)
Studies from the systematic review that are relevant for use in the current review	• Gagné 2018 • Shenkin 2019 • O'Sullivan 2018 • Fabri 2001 • Shenkin 2019 • Han 2018 • Han 2014 • Meeberg 2016 • Han 2013 • Baten 2018
Studies from the systematic review that are not relevant for use in the current review	• Bédard 2019 • Grossmann 2017 • Han 2015 • Hasemann 2018 • Hasemann 2019 • Marra 2018 • Barbic 2018 • Carpenter 2011 • Carpenter 2011 • Dyer 2016 • O’Sullivan 2017 • Schofield 2009 • Wilber 2005 • Wilber 2008 • Wilding 2016 Studies either did not contain an index test of interest or did not select for patients with delirium
Setting	Hospitals (emergency departments)
Patient population	Patients with a mean or median age 65 years or older, visiting an ED
Tests included	4AT

	CAM SQID CAM-ICU B-CAM
Person delivering the test	Not reported
Reference standard	• DSM-5 • DSM-IV • DSM-IV-TR • CAM
Outcomes	4AT - pooled data (duration not reported) (cut-off score: NR) • Sensitivity (95% CI): 87% (74-94%) • Specificity (95% CI): 87% (60-97%) CAM-ICU - pooled data (duration not reported) (cut-off score: NR) • Sensitivity (95% CI): 85% (39-98%) • Specificity (95% CI): 98% (94-99%) CAM - raw data (duration not reported) (cut-off score: NR) Fabbri 2001 • Sensitivity (95% CI): 94% (71-100%) • Specificity (95% CI): 96% (90-99%) Shenkin 2019 • Sensitivity (95% CI): 40% (26-57%) • Specificity (95% CI): 100% (98-100%) SQiD (patient) - raw data (duration not reported) (cut-off score: NR) Han 2018 • Sensitivity (95% CI): 62% (47-75%) • Specificity (95% CI): 79% (74-83%)

	SQiD (surrogate) - raw data (duration not reported) (cut-off score: NR) Han 2018 • Sensitivity (95% CI): 91% (76-98%) • Specificity (95% CI): 77% (71-82%) B-CAM - raw data (duration not reported) (cut-off score: NR) Baten 2018 • Sensitivity (95% CI): 65% (50-79%) • Specificity (95% CI): 94% (90-96%) Han 2013 • Sensitivity (95% CI): 84% (71-93%) • Specificity (95% CI): 96% (93-98%)
Sources of funding	This work is funded by the University Medical Center, Groningen.

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Chen, 2021

Bibliographic Reference	Chen, Ting-Jhen; Chung, Yi-Wei; Chang, Hui-Chen Rita; Chen, Pin-Yuan; Wu, Chia-Rung; Hsieh, Shu-Hua; Chiu, Hsiao-Yean; Diagnostic accuracy of the CAM-ICU and ICDSC in detecting intensive care unit delirium: A bivariate meta-analysis.; International journal of nursing studies; 2021; vol. 113; 103782

Study Characteristics

Study design	Systematic review
Countries of included studies	Not reported
Databases searched	• PubMed • Embase • CINAHL • ProQuest Dissertations • Theses A&I
Years searched	Database inception to April 26, 2019
Inclusion criteria	• Full-text studies assessing the sensitivity and specificity of the CAM-ICU or the ICDSC against reference standards (i.e., various editions of the DSM or ICD) in adult patients (aged≥18 years) who were admitted to an ICU • Incomplete published theses and dissertations were included if they met the aforementioned criteria • No language restrictions were applied
Exclusion criteria	Studies published in conference proceedings or book chapters without full text were excluded
Number of studies included in the systematic review	29 on CAM-ICU 12 on ICDSC 34 total (manually calculated)
Studies from the systematic review that are relevant for use in the current review	All studies
Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	Hospitals (intensive care units)

Patient population	Adult patients (aged≥18 years) who were admitted to an ICU
Tests included	ICDSC CAM-ICU
Person delivering the test	Not reported
Reference standard	- CAM-ICU: DSM was the most commonly used reference for delirium diagnosis (n = 28) - ICDSC: The DSM (i.e., fourth edition [DSM-IV], DSM-IV-Text Revision [TR], or fifth edition [DSM-5]) was the most frequently used reference standard for delirium diagnosis (n = 9)
Outcomes	CAM-ICU (Average duration: 2-3 min; may be up to 10 mins when users are unfamiliar with the content) (cut-off score: NR) - Sensitivity (95% CI): 84% (77-88%) - Specificity (95% CI): 95% (91-97%) ICDSC (Average duration: 2 mins) (cut off score: 4 (9 studies); 3 (1 study); NR (2 studies)) - Sensitivity (95% CI): 83% (74-90%) - Specificity (95% CI): 87% (78-93%)
Sources of funding	The authors declare no potential conflicts of interest regarding the authorship or publication of this article. This meta-analysis was supported by grants from Taipei Medical University Shuang Ho Hospital (108FRP-06) and the Ministry of Science and Technology, Taiwan (MOST 106-2314-B-038-058-MY3).

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Ho, 2020

Bibliographic Reference	Ho, MH; Montgomery, A; Traynor, V; Chang, CC; Kuo, KN; Chang, HR; Chen, KH; Diagnostic Performance of Delirium Assessment Tools in Critically Ill Patients: A Systematic Review and Meta-Analysis.; Worldviews on evidence-based nursing; 2020; vol. 17 (no. 4); 301-310

Study Characteristics

Study design	Systematic review
Countries of included studies	Not reported
Databases searched	• Cochrane Library • PubMed • Embase • CINAHL • Chinese Electronic Periodical Services
Years searched	Inception to October 2018
Inclusion criteria	• Papers written in English or Mandarin, published from inception to October 2018, published in a peer-reviewed journal, focused on the use of a delirium assessment tool in ICU, described appropriate reference criteria (DSM) by an expert in delirium and patients included were 18 years and older • Articles that evaluated the outcomes of sensitivity, specificity, receiver operating characteristics (ROC) curve, positive and negative likelihood ratio of the results of delirium assessment tools were included • Articles that adopted prospective, retrospective, observational (case-control, cross-sectional, cohort and longitudinal) research designs which met the inclusion criteria were considered eligible for inclusion
Exclusion criteria	Studies that were not published in full-text papers (i.e., abstract in conference proceedings) were excluded
Number of studies included in the systematic review	29
Studies from the systematic review that are relevant for use in the current review	All studies
Studies from the systematic review that are not relevant for use in the	N/A

current review	
Setting	Hospitals
Patient population	Patients 18 years and older
Tests included	ICDSC CAM-ICU
Person delivering the test	- Nurses - Doctors - Independent investigators - Intensivists - Physician / nurse investigators - Examiners
Reference standard	- DSM-IV - DSM-5 - DSM-IV-TR - Clinical diagnosis confirmed by a psychiatrist - DSM-III-R
Outcomes	Cam-ICU (duration not reported) (cut-off score: NR) - Sensitivity (95% CI): 85% (77-91%) - Specificity (95% CI): 95% (90-97%) - AUC: 0.96 ICDSC (duration not reported) (cut-off score: NR) - Sensitivity (95% CI): 87% (70-95%) - Specificity (95% CI): 91% (85-95%) - AUC: 0.95
Sources of funding	Not reported

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	High
Overall study ratings	Applicability as a source of data	Fully applicable

Ho, 2022

Bibliographic Reference	Ho, Mu-Hsing; Choi, Edmond Pui Hang; Chiu, Hsiao-Yean; Shen Hsiao, Shu-Tai; Traynor, Victoria; Using the nursing delirium screening scale in assessing postoperative delirium: A meta-regression.; Research in nursing & health; 2022; vol. 45 (no. 1); 23-33

Study Characteristics

Study design	Systematic review
Countries of included studies	Portugal, Turkey, China, Sweden, Germany, U.S.
Databases searched	• EMBASE (via OvidSP) • MEDLINE (via PubMed) • The Cochrane Library • CINAHL (via EBSCO) • A Chinese e-Journal database (via AirtiLibrary)
Years searched	January 2005 and June 2020
Inclusion criteria	• Studies reporting the diagnostic test accuracy of the Nu-DESC (irrespective of where the Nu-DESC was performed) were considered eligible. All cohort studies, including prospective, cross-sectional, case-control, and retrospective design, which compared the Nu-DESC with a reference standard (i.e., DSM criteria or valid delirium assessment tools) were included • Studies were included in which postoperative patients received an assessment for postoperative delirium. Postoperative delirium was defined as an acute change in attention, cognition, and levels of consciousness occurring post-anaesthesia and surgery. • Patients were adult (age ≥ 18 years) postoperative patients who received any type of surgery and any method of anaesthesia were considered for inclusion • The reference standard was a delirium diagnosis according to DSM criteria or a validated tool such as the CAM-ICU or the intensive care delirium screening checklist (ICDSC)
Exclusion criteria	Studies in which the diagnostic test accuracy of the Nu-DESC (i.e., sensitivity or specificity) was not reported or from which data could not be extracted were excluded
Number of studies included in the systematic review	11
Studies from the	All studies

systematic review that are relevant for use in the current review	
Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	Hospitals (post-surgery)
Patient population	Adult (age ≥ 18 years) postoperative patients who received any type of surgery and any method of anaesthesia
Tests included	NuDESC
Person delivering the test	• Nurses • Researchers / research assistants • Physicians • Psychiatrists
Reference standard	• DSM-5 • DSM-IV • DSM-IV-TR • ICDSC • CAM-ICU
Outcomes	Nu-DESC (Average duration: 2.13 (SD: 0.05)) (cut-off score: ≥2) • Sensitivity (95% CI): 73% (44–90%) • Specificity (95% CI): 93% (87–96%) • Positive likelihood ratio (95% CI): 10.2 (6.8–15.2), • Negative likelihood ratio (95% CI): 0.29 (0.12–0.69) • AUC (95% CI): 0.94 (0.91–0.96)
Sources of funding	University of Wollongong, Grant/Award Number: University Postgraduate Award

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low

Section	Question	Answer
Overall study ratings	Applicability as a source of data	Fully applicable

Jeong, 2020a

Bibliographic Reference	Jeong, E; Park, J; Lee, J; Diagnostic Test Accuracy of the 4AT for Delirium Detection: A Systematic Review and Meta-Analysis.; International journal of environmental research and public health; 2020; vol. 17 (no. 20); 1-15

Study Characteristics

Study design	Systematic review
Countries of included studies	Iran, Norway, Australia, UK, Thailand, Italy, Canada, Ireland, Germany
Databases searched	• EMBASE • MEDLINE • CINAHL • PsycINFO
Years searched	The literature was searched in February 2020 (date limits not reported)
Inclusion criteria	• Studies using the 4AT to detect delirium for identifying DTA of the tool • Studies applying a reference standard to diagnose delirium on the basis of a validated tool or standardized criteria of the Diagnostic and Statistical Manual of Mental Disorders (DSM) III, IV, or V; • Studies reporting estimates of DTA including true positive, true negative, false positive, and false negative, or sufficient information to derive them • Studies written in English • Prospective studies in the general clinical settings
Exclusion criteria	Purely observational studies that were inappropriate to test diagnostic accuracy were excluded.
Number of studies included in the systematic review	13
Studies from the systematic review that are relevant	All studies

for use in the current review	
Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	• Hospitals • One study conducted in Nursing homes and daily care centres
Patient population	Not reported
Tests included	4AT
Person delivering the test	Not reported
Reference standard	• DSM-5 • DSM-IV • CAM • CAM-ICU • 3D-CAM
Outcomes	4AT (duration not reported) (cut-off score: >3) • Sensitivity (95% CI): 81.5% (70.7–89.0%) • Specificity (95% CI): 87.5% (79.5–92.7%) • AUC: 0.911 • Positive and negative likelihood ratios not pooled
Sources of funding	Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A1A01072281).

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Jeong, 2020b

Bibliographic Reference	Jeong, E; Park, J; Lee, J; Diagnostic test accuracy of the Nursing Delirium Screening Scale: A systematic review and meta-analysis.; Journal of advanced nursing; 2020; vol. 76 (no. 10); 2510-2521

Study Characteristics

Study design	Systematic review
Countries of included studies	USA, Germany, Sweden, Hong Kong, Canada
Databases searched	• MEDLINE • EMBASE • PsycINFO • CINAHL
Years searched	The literature search was conducted in April 2019 (date limits were not reported)
Inclusion criteria	• Prospective cohort studies in the setting of general practices • Studies that used the Nu-DESC to screen for delirium for investigating Nu-DESC DTA and reported diagnostic accuracy estimates including true positive, false positive, true negative, and false negative, or had sufficient detail to derive these numbers • Studies that used a reference standard that was either a psychiatrist's or a neurologist's diagnosis based on diagnostic interviews or instruments that used the Diagnostic and Statistical Manual of Mental Disorders (DSM) III, IV, and V's diagnostic criteria which is the known gold standard of delirium diagnosis • Studies written in English.
Exclusion criteria	Other types of observational studies and studies enrolling participants with known delirious status (commonly referred to as 'case–control' designs in DTA literature) were excluded
Number of studies included in the systematic review	11
Studies from the systematic review that are relevant for use in the current review	All studies

Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	Six of the reviewed studies were conducted in general wards, including rehabilitation units, medical units, and surgical units in a hospital; three in a recovery room or post-anaesthesia care unit; one in an intensive care unit; and one in an emergency department
Patient population	Not reported
Tests included	NuDESC
Person delivering the test	Not reported
Reference standard	• DSM-5 • DSM-IV • CAM
Outcomes	NuDESC (duration not reported) (cut-off score: >1) • Sensitivity (95% CI): 68.6% (55.3–79.5%) • Specificity (95% CI): 89.4% (83.3–93.5%) • AUC: 0.882 • Positive and negative likelihood ratios not pooled
Additional comments	
Sources of funding	This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	High
Overall study ratings	Applicability as a source of data	Fully applicable

Kim, 2021

Bibliographic Reference	Kim, Sujeong; Choi, Eunju; Jung, Youngsun; Jang, Insil; Postoperative delirium screening tools for post-anaesthetic adult patients in non-intensive care units: A systematic review and meta-analysis.; Journal of clinical nursing; 2021

Study Characteristics

Study design	Systematic review
Countries of included studies	Germany, The Netherlands, Sweden, USA, China, Thailand
Databases searched	• MEDLINE • CINAHL • The Cochrane Library • EMBASE • KoreaMed
Years searched	The literature search was limited to studies published up to February 2020
Inclusion criteria	• Prospective validation studies reporting sensitivity and specificity values with sufficient data • Studies with the prognostic accuracy of a postoperative delirium screening tool reported by clinical nurses • Studies that included participants aged 20 years or older, who underwent general anaesthesia surgery
Exclusion criteria	• Studies designed as a review or meta-analysis • Studies conducted among ICU patients were excluded
Number of studies included in the systematic review	9
Studies from the systematic review that are relevant for use in the current review	All studies
Studies from the systematic review that	N/A

are not relevant for use in the current review	
Setting	Hospitals (post-surgery)
Patient population	Participants aged 20 years or older, who underwent general anaesthesia surgery
Tests included	4AT CAM DOS CAM-ICU NuDESC
Person delivering the test	• Nurses • Psychiatrists • Trainees (did not specify occupation)
Reference standard	DSM-IV
Outcomes	CAM (and variants) (Time taken: 5.27-14 min) (cut-off score: NR) • Sensitivity (95% CI): 47% (37–56%) • Specificity (95% CI): 99% (98–99%) • Positive likelihood ratio (95% CI): 32.10 (7.01–146.93) • Negative likelihood ratio (95% CI): 0.55 (0.34–0.87) NuDESC (Time taken: 1.27-13 min) (cut-off score: ≥2) • Sensitivity (95% CI): 63% (56–69%) • Specificity (95% CI): 93% (91–94%) • Positive likelihood ratio (95% CI): 7.97 (4.38–14.49) • Negative likelihood ratio (95% CI): 0.33 (0.16–0.67) NuDESC (Time taken: 1.27-13 min) (cut-off score ≥1) • Sensitivity (95% CI): 69% (60–76%) • Specificity (95% CI): 94% (92–96%) • Positive likelihood ratio (95% CI): 7.76 (2,058–23.32) • Negative likelihood ratio (95% CI): 0.38 (0.29–0.48)

	4AT (duration not reported) (cut-off score: ≥3) - Sensitivity (95% CI): 95% (77–99%) - Specificity (95% CI): 100% (99–100%) - Positive likelihood ratio (95% CI): 975.91 (60.94–15,614.60) - Negative likelihood ratio (95% CI): 0.06 (0.01–0.30) DOS (duration not reported) (cut-off score: ≥3) - Sensitivity (95% CI): 100% (86–100%) - Specificity (95% CI): 97% (90–99%) - Positive likelihood ratio (95% CI): 24.92 (8.91–69.69) - Negative likelihood ratio (95% CI): 0.02 (0.00–0.32)
Additional comments	CAM variants included CAM and CAM-ICU
Sources of funding	Chung-Ang University

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	High
Overall study ratings	Applicability as a source of data	Fully applicable

Mansutti, 2019

Bibliographic Reference	Mansutti, I; Saiani, L; Palese, A; Detecting delirium in patients with acute stroke: a systematic review of test accuracy.; BMC neurology; 2019; vol. 19 (no. 1); 310

Study Characteristics

Study design	Systematic review
Countries of included studies	Italy, Russia, UK, Czech Republic
Databases searched	- Medline - The Cumulative Index to Nursing and Allied - Health Literature (CINAHL)

	• Scopus
Years searched	Databases were searched up to September 2018
Inclusion criteria	• Diagnostic test accuracy studies • Studies evaluating tools detecting delirium among patients with acute stroke • Studies written in English • studies published up to September 2018.
Exclusion criteria	• Studies reporting protocols regarding diagnostic test accuracy studies • Studies evaluating tools aimed at screening other cognitive issues in patients with acute stroke (e.g., dementia, cognitive decline) • Studies analysing the association between post-stroke delirium and some risk factors or long-term consequences (e.g., dementia) • Studies not conducted in the acute phase of stroke, established as the first 48 h after the onset to the following two weeks
Number of studies included in the systematic review	4
Studies from the systematic review that are relevant for use in the current review	All studies
Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	Hospitals (stroke units / neurovascular departments)
Patient population	Patients with acute stroke
Tests included	4AT CAM-ICU
Person delivering the test	• Neurologists • Trained medical students / junior physicians

	• A panel of specialists, experts on delirium (two neurologists, two neuropsychologists, a psychiatrist and a speech therapist)
Reference standard	• DSM-IV • CAM
Outcomes	4AT (duration not reported) (cut-off score: not reported) Infante 2017 - at admission • Sensitivity: 90.2% • Specificity: 64.5% • AUC: 0.82 Infante 2017 - after 7 days • Sensitivity: 96.4% • Specificity: 76.7% • AUC: 0.88 Kutlubaev 2016 • Sensitivity: 93% • Specificity: 86% • Positive predicted value: 86% • Negative predicted value range: 85.6% Lees 2013 • Sensitivity (95% CI): 100% (74-100) • Specificity (95% CI): 82% (72-89) CAM-ICU (duration not reported) (cut-off score: not reported) • Sensitivity (95% CI): 76% (55–91%) • Specificity (95% CI): 98% (93–100%) • Positive predicted value: 91% (70–99%) • Negative predicted value: 94% (88–98%) • Likelihood ratio 0.47 (0.27–0.83)
Additional comments	Outcome data for CAM-ICU extracted from a single study (Mitasova 2012)

Sources of funding	This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Park, 2021

Bibliographic Reference	Park, J.; Jeong, E.; Lee, J.; The Delirium Observation Screening Scale: A Systematic Review and Meta-Analysis of Diagnostic Test Accuracy; Clinical nursing research; 2021; vol. 30 (no. 4); 464-473

Study Characteristics

Study design	Systematic review
Countries of included studies	The Netherlands, Switzerland, USA, Denmark, Belgium
Databases searched	• MEDLINE • CINAHL • EMBASE • PsycARTICLES
Years searched	Databases were searched in July 2019 (date limits were not reported)
Inclusion criteria	• Prospective cohort, cross-sectional, or controlled trial studies; • Studies using the 13-item DOS scale as a screening tool • Studies estimating diagnostic accuracy through sensitivity, specificity, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) calculations, or reporting sufficient details to derive these values • Studies using either the Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria or a neuropsychologist’s assessment as the reference standard • Studies written in English
Exclusion criteria	Abstracts presented at congresses, reviews, letters, editorials, and unpublished data

Number of studies included in the systematic review	8
Studies from the systematic review that are relevant for use in the current review	All studies
Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	• Hospitals (general, cardiac surgical and palliative wards) • Home hospice
Patient population	Not reported
Tests included	DOS
Person delivering the test	Not reported
Reference standard	• DSM-IV • DRS-R-98
Outcomes	DOS (duration not reported) (cut-off score: ≥3) • Sensitivity (95% CI): 90% (76-97%) • Specificity (95% CI): 92% (88-94%) • AUC: 0.94
Sources of funding	Juneyoung Lee’s research was partially supported by a grant from the College of Medicine, Korea University, Seoul, Republic of Korea (Grant number: K1922241)

Critical appraisal - GDT Crit App - ROBIS checklist

Section	Question	Answer
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Patel, 2018

Bibliographic Reference	Patel, MB; Bednarik, J; Lee, P; Shehabi, Y; Salluh, JI; Slooter, AJ; Klein, KE; Skrobik, Y; Morandi, A; Spronk, PE; Naidech, AM; Pun, BT; Bozza, FA; Marra, A; John, S; Pandharipande, PP; Ely, EW; Delirium Monitoring in Neurocritically Ill Patients: A Systematic Review.; Critical care medicine; 2018; vol. 46 (no. 11); 1832-1841

Study Characteristics

Study design	Systematic review
Countries of included studies	NR
Databases searched	• Cumulative Index to Nursing and Allied Health Literature (CINAHL) • Web of Science • PubMed from the National Center for Biotechnology Information
Years searched	The search was not restricted by date
Inclusion criteria	• Any type of study design investigating delirium monitoring in neuro-critically ill patients of any age • The definition of neuro-critically ill was restricted to and referred to ICU patients with acute intracranial injury (e.g., traumatic brain injury, haemorrhagic stroke), or ischemic stroke • Reference lists of potentially included studies and review articles were also reviewed for additional citations pertinent to the search • Delirium assessments should have occurred at least daily using a delirium screening assessment tool with reporting of rate • Only English-language studies and studies published in the peer-reviewed literature were eligible for inclusion
Exclusion criteria	Editorials, case reports, case-series, lay press articles, abstracts, and reviews
Number of studies included in the	7

systematic review	
Studies from the systematic review that are relevant for use in the current review	- Frenette 2016 - Lees 2013 - Mitasova 2012
Studies from the systematic review that are not relevant for use in the current review	- Naidech 2013 - Oldenbeuving 2011 - Rosenthal 2017 - Kostalova 2012 Studies did not report outcomes of interest
Setting	Hospitals (ICU)
Patient population	Neuro-critically ill patients of any age
Tests included	4AT ICDSC CAM-ICU
Person delivering the test	Not reported
Reference standard	- DSM-IV - CAM
Outcomes	CAM-ICU (duration not reported) (cut-off score: NR) Frenette 2016 - Sensitivity (95% CI): 62% (44-76%) - Specificity (95% CI): 74% (59-85%) - PPV (95% CI): 63% (45-78%) - NPV (95% CI): 70% (55-82%) - Delirium prevalence (by reference standard): 45.9% Mitasova 2012 - Sensitivity (95% CI): 76% (55-91%) - Specificity (95% CI): 98% (93-100%)

	• PPV (95% CI): 91% (70-99%) • NPV (95% CI): 94% (88-98%) • Delirium prevalence (by reference standard): 28% • Delirium prevalence (by index tool): 24% ICDSC (duration not reported) (cut-off score: NR) • Sensitivity (95% CI): 64% (49-77%) • Specificity (95% CI): 79% (63-89%) • PPV (95% CI): 74% (55-87%) • NPV (95% CI): 69 (54-81%) • Delirium prevalence (by reference standard): 45.9% 4AT (duration not reported) (cut-off score: NR) • Sensitivity (95% CI): 100% (74-100%) • Specificity (95% CI): 82% (72-8%9) • PPV (95% CI): 43% (NR) • NPV (95% CI): 100% (NR) • Delirium prevalence (by reference standard): 11% • Delirium prevalence (by index tool): 27%
Sources of funding	• Federal sources included the Veterans Affairs Tennessee Valley Geriatric Research, Education and Clinical Center (GRECC, Nashville, TN) and the National Institutes of Health AG027472, AG035117, HL111111, GM120484 (Bethesda, MD) • EWE has conducted CME activities sponsored by Abbott, Hospira, and Orion • PPP has received a research grant from Hospira Inc, in collaboration with the NIH

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Quispel-Aggenbach, 2018

Bibliographic Reference	Quispel-Aggenbach, DWP; Holtman, GA; Zwartjes, HAHT; Zuidema, SU; Luijendijk, HJ; Attention, arousal and other rapid bedside screening

instruments for delirium in older patients: a systematic review of test accuracy studies.; Age and ageing; 2018; vol. 47 (no. 5); 644-653

Study Characteristics

Study design	Systematic review
Countries of included studies	Not reported
Databases searched	• PubMed, Embase and PsycINFO were searched • Authors scrutinised references of the selected articles and four prior reviews • Authors performed a forward citation search in Google Scholar for each included article • Authors of the included studies were asked per email whether they knew unpublished studies
Years searched	The search was finalised in 12 December 2017, no restriction was made with respect to year of publication
Inclusion criteria	• A bedside screening instrument for delirium was tested • Administration time was <3 min as reported in the included or another article • The study reported sensitivity and specificity of a screening tool; and the study was performed in patients aged 60 years or older • No restriction was made with respect to language
Exclusion criteria	• Index tests to diagnose delirium (CAM, DRS-R98) or delirium tremens, or to rate the severity of delirium (MDAS) or the accompanying cognitive impairment (CTD) • Tests based on surrogate information because it generally takes more than 3 min to reach a caregiver and administer the test, and retrieving surrogate information is often unsuccessful • Tests based on symptoms elicited during history taking • Tests part of establishing the reference standard diagnosis • Studies performed in patients on mechanical ventilation
Number of studies included in the systematic review	27
Studies from the systematic review that	• Voyer 2015 • Bilodeau 2016 • Koop 2016

are relevant for use in the current review	- Pelletier 2017
Studies from the systematic review that are not relevant for use in the current review	- Jitapunkul 1992 - Pompei 1995 - Macleod 1997 - O'Keeffe 1997 - Adamis 2006 - Bryson 2011 - Leung 2011 - Chester 2012 - Han 2013 - Emerson 2014 - Han 2015 - Lees 2013 - Tieges 2013 - O'Regan 2014 - Shoaib 2015 - Voyer 2016 - Adamis, 2016 - Hendry 2016 - Leonard 2016 - O'Regan 2016 - Fick 2015 - Lin 2015 - Bedard 2017 - Dyer 2017 - Grossmann 2017 - Richardson 2017
Setting	Acute care hospital and nursing homes
Patient population	Patients aged 60 years or older
Tests included	RADAR
Person delivering the test	- Nurses - Nurse assistants - Research assistants
Reference standard	- DSM-IV-TR (with CAM) - DSM-5 - DSM-5 (with CAM) - CAM
Outcomes	RADAR (Test duration: 7 seconds - <1 min) (cut-off score: >0 item present) Voyer 2015 - RADAR 1–4 × daily - Sensitivity (95% CI): 65% (43–84%); Specificity (95% CI): 71% (64–78%)

	- RADAR 3–4 × daily - Sensitivity (95% CI): 73% (39–94%); Specificity (95% CI): 67% (57–76%) Bilodeau 2016 - in dementia patients only - RADAR - Sensitivity (95% CI): 100% (3–100%); Specificity (95% CI): 77% (58–90%) Koop 2016 - RADAR - Sensitivity (95% CI): 100% [3–100%] (score closest to day of delirium diagnosis); Specificity (95% CI): 69% [39–91%] Pelletier 2017 - RADAR - Sensitivity (95% CI): 100% (16–100%); Specificity (95% CI): 72% (59–86%)
Additional comments	- For outcomes, CI in squared brackets were calculated by the authors with data in the primary article - Test duration derived from inclusion criteria
Sources of funding	The Dutch Ministry of Health supported this work (grant number 325414)

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Rosgen, 2018

Bibliographic Reference	Rosgen, B; Krewulak, K; Demiantschuk, D; Ely, EW; Davidson, JE; Stelfox, HT; Fiest, KM; Validation of Caregiver-Centered Delirium Detection Tools: A Systematic Review.; Journal of the American Geriatrics Society; 2018; vol. 66 (no. 6); 1218-1225

Study Characteristics

Study design	Systematic review
Countries of included studies	Australia
Databases searched	• MEDLINE • EMBASE • PsycINFO • CINAHL • Scopus
Years searched	From database inception to May 15, 2017, with no restrictions
Inclusion criteria	• Original/primary peer-reviewed research • Observational study design (e.g. cohort study, cross-sectional study) • Studies conducted in adult patients (≥ 18 years old) in any hospital setting • Studies that reported on the validity of caregiver-centred delirium detection tools
Exclusion criteria	Not reported
Number of studies included in the systematic review	6
Studies from the systematic review that are relevant for use in the current review	• Sands 2010
Studies from the systematic review that are not relevant for use in the current review	• Buss 2007 • Hendry 2015 • Martins 2014 • Rhodius-Meester 2013 • Schuman 2016
Setting	Hospitals
Patient population	Adult patients (≥ 18 years old) in any hospital setting

Tests included	SQID
Person delivering the test	Not reported
Reference standard	Psychiatric interview conducted by trained physicians using DSM-IV criteria
Outcomes	SQiD (duration not reported) (cut-off score: "yes" = suspected delirium) • Sensitivity (95% CI): 80% (28.4-99.5%) • Specificity (95% CI): 71% (41.9-91.6%) • Positive predictive value (95% CI): 50% (15.7-84.3%) • Negative predictive value (95% CI): 91% (58.7-99.8%)
Additional comments	Table 2 states that reference standard is CAM but this is incorrect (primary study checked)
Sources of funding	Not reported

Critical appraisal - GDT Crit App - ROBIS checklist

Section	Question	Answer
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Tieges, 2021

Bibliographic Reference	Tieges, Zoe; Maclullich, Alasdair M J; Anand, Atul; Brookes, Claire; Cassarino, Marica; O'connor, Margaret; Ryan, Damien; Saller, Thomas; Arora, Rakesh C; Chang, Yue; Agarwal, Kathryn; Taffet, George; Quinn, Terence; Shenkin, Susan D; Galvin, Rose; Diagnostic accuracy of the 4AT for delirium detection in older adults: systematic review and meta-analysis.; Age and ageing; 2021; vol. 50 (no. 3); 733-743

Study Characteristics

Study design	Systematic review
Countries of included studies	USA, Thailand, Russia, UK, Norway, Ireland, Germany, Iran, Italy, Canada, Australia
Databases searched	• MEDLINE(OVID) • EMBASE (OVID)

	• PsycINFO (EBSCO) • CINAHL (EBSCO) • clinicaltrials.gov • the Cochrane Central Register of Controlled trials
Years searched	2011 (the year the 4AT was published online) to 21 December 2019
Inclusion criteria	• Participants aged ≥65 • Studies that examined the diagnostic accuracy of the 4AT for detection of delirium • Study included a reference standard assessment of delirium made using standardised diagnostic criteria or a validated tool • Study design was cross-sectional, retrospective or prospective cohort If identified studies included adults both younger and older than the threshold age, the study authors were contacted to enquire about the possibility to access data on the older adults only.
Exclusion criteria	Studies in patients with delirium tremens were excluded
Number of studies included in the systematic review	17 studies from 16 papers
Studies from the systematic review that are relevant for use in the current review	All studies
Studies from the systematic review that are not relevant for use in the current review	N/A
Setting	• Hospitals • One study conducted in Nursing homes and daily care centres
Patient population	Participants aged ≥65
Tests included	4AT

Person delivering the test	• Nurse • Psychiatrist • Researcher • Neurologist • Medical student
Reference standard	• DSM-5 • Chart review by 2-3 physicians • CAM • DSM-IV-TR
Outcomes	4AT (duration not reported) (cut-off score: ≥4) • Sensitivity (95% CI): 88% (80–93%) • Specificity (95% CI): 88% (82–92%)
Sources of funding	Supported by the Wellcome Trust-University of Edinburgh Institutional Strategic Support Fund.Grant no. IS3-T06/03.

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

van Velthuijsen, 2016

Bibliographic Reference	van Velthuijsen, EL; Zwakhalen, SM; Warnier, RM; Mulder, WJ; Verhey, FR; Kempen, GI; Psychometric properties and feasibility of instruments for the detection of delirium in older hospitalized patients: a systematic review.; International journal of geriatric psychiatry; 2016; vol. 31 (no. 9); 974-89

Study Characteristics

Study design	Systematic review
Countries of included studies	The Netherlands, Italy, Sweden, Hong Kong, Germany, USA, Australia, Brazil, Spain, Finland, Portugal, Canada, Ireland, Thailand, Czech Republic
Databases searched	• PubMed • MEDLINE • PsycINFO

	• CINAHL
Years searched	From database conception until 14 September 2015
Inclusion criteria	• Studies that reported the psychometric qualities of delirium detection instruments • Studies aimed at the detection of delirium in older hospitalised patients (mean or median age 65+) • Studies where the reference standard was a diagnosis made by a medical doctor based on the criteria of the Diagnostic and Statistical Manual of Mental Disorders (DSM, editions III, IV or V) or the International Classification of Diseases (ICD) • Apart from age, no restrictions were set for study population or hospital wards
Exclusion criteria	Articles not written in English
Number of studies included in the systematic review	43
Studies from the systematic review that are relevant for use in the current review	• Bellelli 2014 • Fabbri 2001 • González 2004 • Hestermann 2009 • Laurila 2002 • Leung 2008 • Lin 2015 • Martins 2015 • Monette 2001 • Pompei 1995 • Ryan 2009 • Thomas 2012 • Wongpakaran 2011 • Marcantonio 2016 • van Gemert 2007 • Koster 2009 • Lin 2015 • Giusti and Piergentili 2012 • Han 2014 • Luetz 2010 • Mitasova 2012 • Pipanmekaporn 2014 • Han 2013 • Neufeld 2013 • Lingehall 2013 • Luetz 2010 • Leung 2008 • Powers 2013

Studies from the systematic review that are not relevant for use in the current review	• Albert 1992 • Andrew 2009 • Chester 2012 • De Jonghe 2005 • De Negreiros 2008 • De Rooij 2006 • Grossmann 2014 • Han 2015 • Martins 2014 • Neelon 1996 • Ni Chonchubhair 1995 • O'Regan 2014 • Otter 2005 • Pompei 1995 • Rhodius-Meester 2013 • Rockwood 1996 • Salih 2012 • Sorensen Duppils and Johansson 2011 • Schuurmans 2003 Studies did not concern an index test of interest or did not report outcomes of interest
Setting	Hospitals
Patient population	Older hospitalised patients (mean or median age 65+)
Tests included	4AT CAM 3D-CAM DOS SQID ICDSC CAM-ICU B-CAM NuDESC
Person delivering the test	• Nurses • Informal carers • Doctors • Psychiatrists • Psychologists • Researchers / research assistants

Reference standard	• DSM-III • DSM-III-R • DSM-IV • DSM-IV-TR • ICD-10
Outcomes	4AT (Time taken: <2 min) (cut-off score: 4) • Sensitivity: 90% • Specificity: 84% CAM (Time taken: 5min - <15 min) (cut-off score: NR) Fabbri 2001 • Sensitivity: 94% • Specificity: 96% González 2004 • Sensitivity: 90% • Specificity: 100% Hestermann 2009 • Sensitivity: 77% • Specificity: 96-100% Laurila 2002 • Sensitivity: 80-85% • Specificity: 63-84% Leung 2008 • Sensitivity: 76% • Specificity: 100%

	Martins 2015 • Sensitivity: 79% • Specificity: 99% Monette 2001 • Sensitivity: 64% • Specificity: 93% Pompei 1995 • Sensitivity: 46% • Specificity: 92% Ryan 2009 • Sensitivity: 88% • Specificity: 100% Thomas 2012 • Sensitivity: 74-82% • Specificity: 91-100% Wongpakaran 2011 • Sensitivity: 92% • Specificity: 100% 3D-CAM (Time taken: 3 min) (cut-off score: NR) • Sensitivity: 95% • Specificity: 94% b-CAM (<1 min) (cut-off score: >1) • Sensitivity: 78-84%

	• Specificity: 96-97% CAM-ICU (Time taken: 1 - 2 min) (cut-off score: NR) Han 2014 • Sensitivity: 69-72% • Specificity: 99% Luetz 2010 • Sensitivity: 79% • Specificity: 97% Mitasova 2012 • Sensitivity: 76% • Specificity: 98% Neufeld 2013 • Sensitivity: 28% • Specificity: 98% Pipanmekaporn 2014 • Sensitivity: 92% • Specificity: 95% Powers 2013 • Sensitivity: 45% • Specificity: 89% DOS (Time taken: 5 min) Koster 2009 (cut-off score ≥2) • Sensitivity: 100%

	• Specificity: 97% Van Gemert 2007 (cut-off score = 3) • Sensitivity: 89% • Specificity: 87% ICDSC (Time taken: NR) (cut-off score ≥4) • Sensitivity: 69% • Specificity: 100% NuDESC (Time taken: <2 min) Lingehall 2013 (cut-off score ≥2) • Sensitivity: 72% • Specificity: 81% Leung 2008 (cut-off score >0) • Sensitivity: 96% • Specificity: 79% Luetz 2010 (cut-off score 2 and 1) • Sensitivity: 82% • Specificity: 83% Neufeld 2013 (cut-off score ≥2 ≥1) • Sensitivity: 32-80% • Specificity: 69-92% SQiD (Time taken: NR) (cut-off score: NR) • Sensitivity: 77% • Specificity: 51%

Additional comments	One study reported the time taken to administer ICDSC as 'fast', but did not report a time in minutes
Sources of funding	Funding for this study was provided by Maastricht University Medical Center and Maastricht University

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	Low
Overall study ratings	Applicability as a source of data	Fully applicable

Watt, 2021

Bibliographic Reference	Watt, Christine L; Scott, Mary; Webber, Colleen; Sikora, Lindsey; Bush, Shirley H; Kabir, Monisha; Boland, Jason W; Woodhouse, Rebecca; Sands, Megan B; Lawlor, Peter G; Delirium screening tools validated in the context of palliative care: A systematic review.; Palliative medicine; 2021; vol. 35 (no. 4); 683-696

Study Characteristics

Study design	Systematic review
Countries of included studies	USA, Belgium, The Netherlands, Ireland, Australia
Databases searched	• Medline • Embase • PsycINFO • CENTRAL (all via Ovid) • The Cumulative Index of Nursing and Allied Health Literature – CINAHL (via EBSCO Host)
Years searched	• January 1, 1980 to May 3, 2019 • An update was performed from May 1, 2019 to May 3, 2020
Inclusion criteria	Primary quantitative research studies that assessed the validation of delirium screening tools in adult (18+ years old), palliative care eligible populations
Exclusion criteria	• Qualitative studies • Conference abstracts • Editorials • Magazine articles

	• Studies conducted in paediatric, peri-operative, and critical care populations
Number of studies included in the systematic review	17
Studies from the systematic review that are relevant for use in the current review	• Ryan 2009 • Neefjes 2019 • Jorgensen 2017 • Detroyer 2014 • Sands 2010 • Wilson 2019 • de la Cruz 2015
Studies from the systematic review that are not relevant for use in the current review	• Andrew 2009 • Barahona 2018 • Breitbart 1997 • Cacchione 2002 • Grassi 2001 • Hamano 2015 • Kang 2019 • Klankluang 2020 • Lawlor 2000 • Stillman and Rybicki 2000 Studies did not concern a diagnostic test of interest
Setting	• Inpatient oncology • Community hospice • Palliative care units (including hospital and hospice palliative care units)
Patient population	Adults (18+ years) in palliative care
Tests included	CAM DOS SQID B-CAM NuDESC
Person delivering the test	Not reported

Reference standard	- DSM-IV - DRS-R-98 - CAM - DSM-5 - MDAS
Outcomes	CAM (duration not reported) (cut-off score: binary) Ryan 2009 - Sensitivity (95% CI): 88% (62–98) - Specificity (95% CI): 100% (88–100) Nu-DESC (duration not reported) (diagnostic score: ⩾7) de la Cruz 2015 - Sensitivity (95% CI): nurse: 63% (NR); caregiver evening: 35% (NR); caregiver night: 21% (NR) - Specificity (95% CI): nurse: 67% (NR); caregiver evening: 80%; caregiver night: 85% DOS (duration not reported) Detroyer 2014 (optimal cut-off score ⩾3; diagnostic score binary) - Sensitivity (95% CI): 81.8% (52−95%) - Specificity (95% CI): 96.1% (90−98%) Jorgensen 2017 (optimal cut-off score ⩾3; diagnostic score ⩾18) - Sensitivity (95% CI): 97% (81−100%) - Specificity (95% CI): 89% (75−96%) Neefjes 2019 (optimal cut-off score ⩾3; diagnostic score ⩾17.5) - Sensitivity (95% CI): >99.9% (95.8–100%) - Specificity (95% CI): 99.6.% (95.5–100%) SQiD (duration not reported) (cut-off score: binary) Sands 2010

	- Sensitivity (95% CI): 80% (28.4–99.5%) - Specificity (95% CI): 71% (41.9–91.6%) b-CAM (duration not reported) (cut-off score: binary) Wilson 2019 - Sensitivity (95% CI): 80% (40−96%) - Specificity (95% CI): 87% (67−96%)
Sources of funding	The author(s) received no financial support for the research, authorship, and/or publication of this article

Critical appraisal - GDT Crit App - ROBIS checklist

Section	**Question**	**Answer**
Overall study ratings	Overall risk of bias	High
Overall study ratings	Applicability as a source of data	Fully applicable

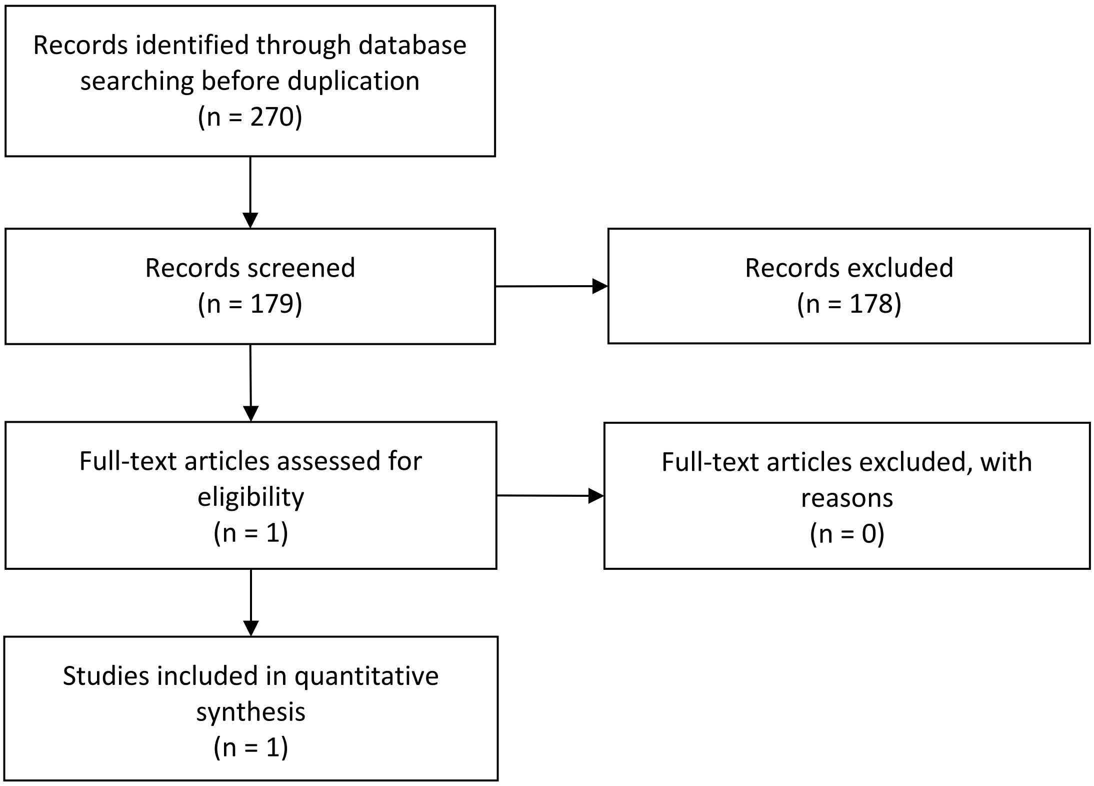
Appendix E – Forest plots

No forest plots were generated for this review because the data were drawn directly from existing systematic reviews and no further meta-analysis was done. The systematic reviews used in this review that provided pooled estimates contain forest plots for the studies they included (see 1.1.13).

Appendix F – GRADE tables

It was not possible to undertake GRADE assessment for the evidence contained in the systematic reviews included in this review.

Appendix G – Economic evidence study selection


Records identified through database searching before duplication (n = 270)
Records screened (n = 179)
Records excluded (n = 178)
Full-text articles assessed for eligibility (n = 1)
Full-text articles excluded, with reasons (n = 0)
Studies included in quantitative synthesis (n = 1)

Appendix H – Economic evidence tables

Table 18: Economic evidence

Study	Study type	Setting	Interventions	Population	Methods of analysis	Base-case results	Sensitivity analyses	Additional comments
MacLullich et al (2019)	Cost-utility analysis Decision tree Cost analysis linked with RCT over 12 weeks	UK emergency department or acute general medical wards NHS & PSS perspective	4AT vs CAM	Patients aged ≥70 years in emergency departments or acute general medical wards (excluded if acute life-threatening illness requiring time-critical intervention, or in a coma, or unable to communicate in English)	Sensitivity, specificity, and costs of true positive and true negative cases were informed by the trial analysis. All other values were collected in a survey of clinical experts during the study. Quality of life values were estimated using the clinical expert survey Modelled states are true or false positives or negatives on either the 4AT or the CAM 12-week time horizon No discounting applied due to the short time horizon	Base-case analysis (Scottish scenario): Costs: 4AT: £4,680 CAM: £4,770 Incr: -£90.35 QALYs: 4AT: 0.14050 CAM: 0.14103 Incr: -0.00053 ICER: £170,553 (SWQ)	In one-way sensitivity analysis the model was most sensitive to the specificity of both tests, cost of false and true negatives, and QALY estimates. The results of the probabilistic sensitivity analysis indicated that there is considerable uncertainty in the estimated cost-effectiveness due to clustering of incremental costs and QALYs around zero. The scenario analysis using English costs had the same QALYs, and had costs as follows: 4AT: £4,416 CAM: £4,478 Incr: -£61.52 ICER: £116,133 (SWQ)	Funded by the National Institute for Health Research (NIHR) No mention of health inequalities. The authors noted some limitations including limited participants (so insufficient power to analyse the economic outcomes) and relying on expert opinion for most model parameters. The analysis demonstrated the potential of the 4AT as a screening test, and improved detection as a potential means of producing large cost savings.

Incr: Incremental; SWQ: South West Quadrant

Table 19: Economic evaluation quality checklist

Study identification MacLullich et al. (2019) The 4 'A's test for detecting delirium in acute medical patients: a diagnostic accuracy study		
Category	**Rating**	**Comments**
Applicability		
1.1 Is the study population appropriate for the review question?	Yes	Patients aged ≥70 in emergency departments or acute general medical wards
1.2 Are the interventions appropriate for the review question?	Yes	4AT vs CAM

Study identification **MacLullich et al. (2019) The 4 'A's test for detecting delirium in acute medical patients: a diagnostic accuracy study**		
Category	**Rating**	**Comments**
1.3 Is the system in which the study was conducted sufficiently similar to the current UK context?	Yes	
1.4 Is the perspective for costs appropriate for the review question?	Yes	NHS and PSS perspective
1.5 Is the perspective for outcomes appropriate for the review question?	Yes	
1.6 Are all future costs and outcomes discounted appropriately?	NA	The analysis was over 12 weeks only, so discounting was not necessary in this scenario.
1.7 Are QALYs, derived using NICE's preferred methods, or an appropriate social care-related equivalent used as an outcome? If not, describe rationale and outcomes used in line with analytical perspectives taken (item 1.5 above).	Partly	QALYs were used, but due to lack of reported values in the literature, expert elicitation of quality of life (and other parameter values) was conducted using a survey of clinical experts and experienced health professionals. The questionnaire asked experts to provide a best guess utility between 0-1, as well as a higher and lower estimate.
1.8 OVERALL JUDGEMENT	**DIRECTLY APPLICABLE**	There is no need to use section 2 of the checklist if the study is considered 'not applicable'.
Limitations		
2.1 Does the model structure adequately reflect the nature of the topic under evaluation?	Yes	Simple decision tree with costs and outcomes assigned to the four outcomes of the test (true and false positives and negatives).
2.2 Is the time horizon sufficiently long to reflect all important differences in costs and outcomes?	Yes	Delirium is a fairly short-term condition, and the 12-week time horizon is long enough to capture differences
2.3 Are all important and relevant outcomes included?	Yes	Sensitivity, specificity, mortality, quality of life
2.4 Are the estimates of baseline outcomes from the best available source?	Yes	Within-trial analysis
2.5 Are the estimates of relative intervention effects from the best available source?	Yes	Within-trial analysis
2.6 Are all important and relevant costs included?	Yes	Hospital costs, cost of tests in hospital, community health service costs
2.7 Are the estimates of resource use from the best available source?	Yes	Collected as part of the trial
2.8 Are the unit costs of resources from the best available source?	Yes	
2.9 Is an appropriate incremental analysis presented or can it be calculated from the data?	Yes	
2.10 Are all important parameters whose values are uncertain subjected to appropriate sensitivity analysis?	Yes	Scenario analysis of Scottish//English cost estimates, one-way sensitivity analysis varying parameters over a range of +/- 25%, probabilistic analysis.
2.11 Has no potential financial conflict of interest been declared?	Yes	
2.12 OVERALL ASSESSMENT	**MINOR LIMITATIONS**	

Appendix I – Health economic model

No economic modelling was done for this review.

Appendix J – Excluded studies

Study	Reason for exclusion
Gélinas, Céline (2018) Delirium Assessment Tools for Use in Critically Ill Adults: A Psychometric Analysis and Systematic Review. Critical Care Nurse 38(1): 38-54	- Review article but not a systematic review
Helfand, Benjamin K I, D'Aquila, Madeline L, Tabloski, Patricia et al. (2021) Detecting Delirium: A Systematic Review of Identification Instruments for Non-ICU Settings. Journal of the American Geriatrics Society 69(2): 547-555	- Study does not contain report relevant outcomes
Jones, RN, Cizginer, S, Pavlech, L et al. (2019) Assessment of Instruments for Measurement of Delirium Severity: A Systematic Review. JAMA internal medicine 179(2): 231-239	- Study does not contain report relevant outcomes

Appendix K – Research recommendations – full details

K.1.1 Research recommendation

What is the diagnostic accuracy, and ease of implementation, of different delirium assessment tools:

- for people with pre-existing cognitive impairment, for example dementia, learning disability or severe depression
- for people who do not speak English as a first language
- in different settings, for example emergency departments, residential care homes or virtual consultations

when delivered by different types of healthcare practitioners, for example healthcare assistants or allied health professionals such as paramedics?

K.1.2 Why this is important

The committee agreed that in terms of their overall diagnostic accuracy there was little to choose between different tests. However, some tests were designed to be used in specific settings and by specific professional groups and the committee agreed that for future updates it would be useful to be able to focus in on who was delivering the test, where they were delivering it, and how difficult it was to deliver (in terms of time and training needed for example). They were also aware of the difficulties in identifying delirium in people with dementia or the cognitive impairments or affective disorders, and based on the evidence they saw were unable to make recommendations about this.

K.1.3 Rationale for research recommendation

Importance to ‘patients’ or the population	Assessments for delirium are used by a range of healthcare practitioners in a range of settings and the right tool may not be the same in all cases. People with pre-existing dementia and other cognitive impairments may wrongly have their behaviour ascribed to their existing condition rather than being identified as delirium
Relevance to NICE guidance	Having these data will allow NICE to refine its guidance on assessment tools in future updates of this guideline
Relevance to the NHS	Having assessment tools that can be delivered quickly and by a wide range of practitioners will mean that senior nurses and doctors will need to spend less time assessing people for delirium.
National priorities	None
Current evidence base	Covers a limited range of settings, but does not cover settings outside of hospital or peoples homes.
Equality considerations	As noted above, people with pre-existing cognitive impairment will fare better with better assessment because it will be less likely that their delirium is attributed to their pre-existing condition.

K.1.4 Modified PICO table

Population	People with suspected delirium Subgroups: • people with and without existing cognitive impairment such as dementia • people who do not speak English well.
Assessment tool	Tools for identifying delirium
Reference standard	DSM-5 diagnosis by a healthcare professional trained to do so.
Outcome	• Specificity, sensitivity • Likelihood ratios • Area under curve Subgrouped by: • Healthcare practitioner delivering the test • Setting in which the test was delivered
Study design	Diagnostic cross sectional study
Additional information	Time taken to administer the test would be a useful additional parameter.

Appendix L – Methods

This guideline was developed using the methods described in the 2022 NICE guidelines manual.

Declarations of interest were recorded according to the NICE conflicts of interest policy.

Developing the review questions and outcomes

The review question developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NICE guideline development team B and refined and validated by the guideline committee.

The review questions was based on the population, index test(s), reference standard and outcome framework for reviews of diagnostic accuracy

Reviewing research evidence

Review protocols

A review protocol was developed with the guideline committee to outline the inclusion and exclusion criteria used to select studies for the evidence review. The review protocol was prospectively registered in the PROSPERO register of systematic reviews.

Searching for evidence

Evidence was searched for each review question using the methods specified in the review protocol.. A search was undertaken in accordance with NICE methods for systematic reviews of diagnostic accuracy published between 2016 and 2022 (see appendix B for details). The short search date was chosen to ensure that any systematic reviews that were included were up to date and included recent primary studies.

Selecting studies for inclusion

All references identified by the literature searches and from other sources (for example, previous versions of the guideline or studies identified by committee members) were uploaded into EPPI reviewer software (version 5) and de-duplicated. Titles and abstracts were assessed for possible inclusion using the criteria specified in the review protocol. 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.

The decision not to use priority screening was taken by the reviewing team based on the size of the database, heterogeneity of studies included in the review and predicted number of includes and the full database was screened.

The full text of potentially eligible studies was retrieved and assessed according to the criteria specified in the review protocol. A standardised form was used to extract data from included studies.

Incorporating published evidence syntheses

The searches aimed to identify published systematic reviews as the main form of evidence for this review.

Methods of combining evidence

In this guideline, systematic reviews of diagnostic test accuracy were included. Diagnostic test accuracy (DTA) data are classified as any data in which a feature – be it a symptom, a risk factor, a test result or the output of some algorithm that combines many such features – is observed in some people who have the condition of interest at the time of the test and some people who do not. Such data either explicitly provide, or can be manipulated to generate, a 2x2 classification of true positives and false negatives (in people who, according to the reference standard, truly have the condition) and false positives and true negatives (in people who, according to the reference standard, do not).

The 'raw' 2x2 data can be summarised in a variety of ways. Those that were used for decision making in this guideline were as follows:

- **Positive likelihood ratios** describe how many times more likely positive features are in people with the condition compared to people without the condition. Values greater than 1 indicate that a positive result makes the condition more likely.
    - LR^+ = (TP/[TP+FN])/(FP/[FP+TN])
- **Negative likelihood ratios** describe how many times less likely negative features are in people with the condition compared to people without the condition. Values less than 1 indicate that a negative result makes the condition less likely.
    - LR^- = (FN/[TP+FN])/(TN/[FP+TN])
- **Sensitivity** is the probability that the feature will be positive in a person with the condition.
    - sensitivity = TP/(TP+FN)
- **Specificity** is the probability that the feature will be negative in a person without the condition.
    - specificity = TN/(FP+TN)
- **Positive predictive values** describe the probability that a person with a positive feature has the disease.
    - PPV = TP/ (TP+FP)
- **Negative predictive values** describe the probability that a person with a negative feature does not have the disease.
    - NPV = TN/(TN+FN)

Systematic reviews that incorporated these data were eligible for inclusion and reporting.

Appraising the quality of evidence

Systematic reviews of diagnostic accuracy studies

Included systematic reviews were quality assessed using the ROBIS checklist to assess their methodological quality.

Each published systematic review was classified into one of the following three groups:

- **High quality** – It is unlikely that additional relevant and important data would be identified from primary studies compared to that reported in the review, and unlikely that any relevant and important studies have been missed by the review.
- **Moderate quality** – It is possible that additional relevant and important data would be identified from primary studies compared to that reported in the review, but unlikely that any relevant and important studies have been missed by the review.
- **Low quality** – It is possible that relevant and important studies have been missed by the review.

Each published evidence synthesis was also classified into one of three groups for its applicability as a source of data, based on how closely the review matches the specified review protocol in the guideline. Studies were rated as follows:

- **Fully applicable** – The identified review fully covers the review protocol in the guideline.
- **Partially applicable** – The identified review fully covers a discrete subsection of the review protocol in the guideline (for example, some of the factors in the protocol only).
- **Not applicable** – The identified review, despite including studies relevant to the review question, does not fully cover any discrete subsection of the review protocol in the guideline.

Diagnostic accuracy studies

Where included systematic reviews reported the methodological quality of the included studies, this was captured in the narrative of this review.

GRADE for diagnostic accuracy evidence

Where included systematic reviews reported GRADE assessments of their outcomes, these were reported in the narrative of this review

Reviewing economic evidence

Inclusion and exclusion of economic studies

Literature reviews seeking to identify published cost–utility analyses of relevance to the issues under consideration were conducted for all questions. In each case, the search undertaken for the clinical review was modified, retaining population and intervention descriptors, but removing any study-design filter and adding a filter designed to identify relevant health economic analyses. In assessing studies for inclusion, population, intervention and comparator, criteria were always identical to those used in the parallel clinical search; only cost–utility analyses were included. Economic evidence profiles, including critical appraisal according to the Guidelines manual, were completed for included studies.

Appraising the quality of economic evidence

Economic studies identified through a systematic search of the literature were appraised using a methodology checklist designed for economic evaluations (NICE guidelines manual; 2014). This checklist is not intended to judge the quality of a study per se, but to determine whether an existing economic evaluation is useful to inform the decision-making of the committee for a specific topic within the guideline.

There are 2 parts of the appraisal process. The first step is to assess applicability (that is, the relevance of the study to the specific guideline topic and the NICE reference case); evaluations are categorised according to the criteria in Table 20.

Table 20 Applicability criteria

Level	Explanation
Directly applicable	The study meets all applicability criteria, or fails to meet one or more applicability criteria but this is unlikely to change the conclusions about cost effectiveness
Partially applicable	The study fails to meet one or more applicability criteria, and this could change the conclusions about cost effectiveness

Level	Explanation
Not applicable	The study fails to meet one or more applicability criteria, and this is likely to change the conclusions about cost effectiveness. These studies are excluded from further consideration

In the second step, only those studies deemed directly or partially applicable are further assessed for limitations (that is, methodological quality); see categorisation criteria in Table 21.

Table 21 Methodological criteria

Level	Explanation
Minor limitations	Meets all quality criteria, or fails to meet one or more quality criteria but this is unlikely to change the conclusions about cost effectiveness
Potentially serious limitations	Fails to meet one or more quality criteria and this could change the conclusions about cost effectiveness
Very serious limitations	Fails to meet one or more quality criteria and this is highly likely to change the conclusions about cost effectiveness. Such studies should usually be excluded from further consideration

Where relevant, a summary of the main findings from the systematic search, review and appraisal of economic evidence is presented in an economic evidence profile alongside the clinical evidence.

References

Follmann D, Elliott P, Suh I, Cutler J (1992) Variance imputation for overviews of clinical trials with continuous response. Journal of Clinical Epidemiology 45:769–73

Fu R, Vandermeer BW, Shamliyan TA, et al. (2013) Handling Continuous Outcomes in Quantitative Synthesis In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK154408/

Norman G., Sloan JA., Wyrwich KW. (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 41(5):582-92.