Age and Ageing 2018; **0:** 1–10
doi: 10.1093/ageing/afy058

SYSTEMATIC REVIEW

Attention, arousal and other rapid bedside screening instruments for delirium in older patients: a systematic review of test accuracy studies

D. W. P. Quispel-Aggenbach[1,2], G. A. Holtman[1], H. A. H. T. Zwartjes[1], S. U. Zuidema[1], H. J. Luijendijk[1]

[1]Department of General Practice and Elderly Care Medicine, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[2]Department of Geriatric Psychiatry, Parnassia BAVO Groep, Rotterdam, the Netherlands

Address correspondence to: H. J. Luijendijk, Department of General Practice and Elderly Care Medicine, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. Tel: +31503616161; Fax: +31503615034.
Email: h.j.luijendijk@umcg.nl

Abstract

Objective: delirium occurs frequently in frail patients but is easily missed. Screening with a rapid, easy-to-use and highly sensitive instrument might help improve recognition. The aim of this study was to review attention, arousal and other rapid bedside screening instruments for delirium in older patients.
Methods: a literature search was performed in PubMed, PsycINFO and Embase. We scrutinized forward citations in Google Scholar, and references of included articles and prior reviews. We included studies among older patients that investigated the sensitivity and specificity of delirium screening instruments that could be administered in 3 min or less, and did not require surrogate information. We extracted study characteristics, risk of bias, sensitivity and specificity.
Results: we identified 27 studies among 4,766 patients in hospitals and nursing homes. They tested many different single and several combined screening instruments. Prevalence of delirium varied between 4% and 57%. Only one study scored a low risk of bias on all domains. Sensitivity varied between 17% and 100%, and specificity between 38% and 99%. Of the 22 tests with sensitivity ≥90%, seven also had specificity ≥80% in older patients in general. These results were approximately reproduced for the Observational Scale of Level of Arousal (OSLA) and Richmond Agitation and Sedation Scale (RASS): sensitivity and specificity were >80%.
Conclusion: two arousal tests—OSLA and RASS—had reproduced high sensitivity and specificity in older patients. Nurses can administer these tests during daily interaction with patients. Test accuracy studies about rapid screening tools for delirium superimposed on dementia were scarce.

Keywords: *delirium, screening, rapid test, test accuracy study, systematic review, older people*

Introduction

Delirium is a serious neuropsychiatric disorder with potentially severe consequences such as longer hospital stay, poor cognitive and functional recovery, increased risk of nursing home placement and death [1]. It occurs in 10–40% of older patients in hospitals and nursing homes [2, 3]. Frailty, age above 80 years and the presence of dementia increase the risk of delirium [4].

Around one-third of delirium cases go undetected [5]. The overlap with dementia and depression might hinder recognition, as might a history of psychiatric disease [6]. Lack of trained health care professionals can also contribute to failure in identifying delirium [7]. Screening frail older

persons regularly may help detect delirium more quickly and has been advised in many guidelines [8]. The goal of screening for a disease is to identify persons that are at increased risk of having that disease in a large population (triage). If screening tests are applied to detect diseases that are easily missed, as is the case with delirium, they need to be very sensitive [9–12]. Usually, a screening test cannot be used to make a definitive diagnosis, because vital diagnostic information has not been collected [13]. Subsequently, screen-positive patients need to receive a diagnostic work-up to confirm the diagnosis [9, 12, 14]. Diagnostic tests need to be very specific [9–11]. Ideally, relatively untrained personnel can perform a screening test quickly, easily and as part of routine of their clinical practice.

A number of instruments have been developed to screen for delirium such as the DOSS, the CRS and the DSI (see list of abbreviations below). In addition, diagnostic tools for delirium such as the CAM and the DRS-R98 have been used to screen for delirium [15, 16]. These tests cover all diagnostic and many supporting criteria for delirium, including (surrogate) information about acute onset and fluctuation that patients with cognitive disorders cannot provide reliably. All of the above instruments require a lot of time to administer regularly. In addition, some screening tools such as the DOSS have not been validated in patients with dementia [17]. The CAM and DRS-R98 require training, expertize and experience to be administered correctly. It is likely that the lack of an easy-to-use and rapid screening tool for delirium has hampered the implementation of regular screening [18].

In recent years, several screening tools with a test-time of 3 min or less have been developed and validated. Such instruments may allow screening of many patients in relatively little time. The aim of this study was to review the sensitivity and specificity of rapid screening instruments for delirium in older patients.

Methods

Search strategy and selection criteria

Two authors performed an independent literature search (D.W.P.Q. and H.J.L.). First, they searched PubMed, Embase and PsycINFO with the search terms 'delirium, acute confusion, encephalopathy, clouding of consciousness, toxic psychosis', 'tool, test, instrument, assessment, questionnaire, interview, diagnostic, screening'; and 'sensitivity, specificity, accuracy, validity, reliability, predictive value, likelihood-ratio' (see online Appendix A, available in *Age and Ageing* online). Secondly, they scrutinized references of the selected articles and four prior reviews [15, 16, 19, 20]. Thirdly, they performed a forward citation search in Google Scholar for each included article. Finally, they asked the authors of the included studies per email whether they knew unpublished studies. If title or abstract suggested that the study investigated the test accuracy of a rapid screening instrument for delirium, the full (un)published paper—if available—was obtained. Two authors assessed the papers independently (D.W.P.Q. and H.J.L.) for eligibility. The search was finalized in 12 December 2017.

Studies were selected if they met the following 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. Exclusion criteria were: (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 [21]; tests based on symptoms elicited during history taking; tests part of establishing the reference standard diagnosis; and studies performed in patients on mechanical ventilation. No restriction was made with respect to year of publication or language.

Data-extraction

Two authors (D.W.P.Q., H.J.L. or G.A.H., H.J.L.) independently extracted the following study characteristics: setting, number of participants, prevalence of delirium and of dementia, the index test (screening instrument), the administrator and test-time of the index test, and reference standard (criteria used to diagnose delirium).

They also assessed risk of bias with the QUADAS-2 tool [22]. This tool consists of four domains: patient selection, index test, reference standard and flow and timing of the index test and reference standard. In addition, the tool requires the assessment of the applicability of the patient population, index test and target condition. Risk of bias and applicability concerns were scored as low or high, or unclear if information was missing. We modified the assessment to fit the specifics of our review (see online Appendix A, available in *Age and Ageing* online).

Finally, the test accuracy of the screening instruments in terms of sensitivity and specificity were extracted for all patients and patients diagnosed with dementia as well as inter-rater reliability. Sensitivity and specificity concerned patient level data (not per assessment) and current delirium (if measured during a period, we used the day with highest delirium prevalence) for the tester with the lowest level of training in psychiatric assessment (in case of multiple testers) and the cut-off with highest sensitivity (in case of multiple cut-offs). When information about study characteristics or results was missing in the publication, we requested the author to provide it. Differences in data-extraction and risk of bias assessments were resolved in consensus meetings.

Statistical analysis

We presented the reported sensitivity, specificity, and inter-rater reliability of the tests in all patients, and patients with dementia. We found that confidence intervals around sensitivity and specificity were missing for a number of studies. Therefore, we extracted the raw data of these studies

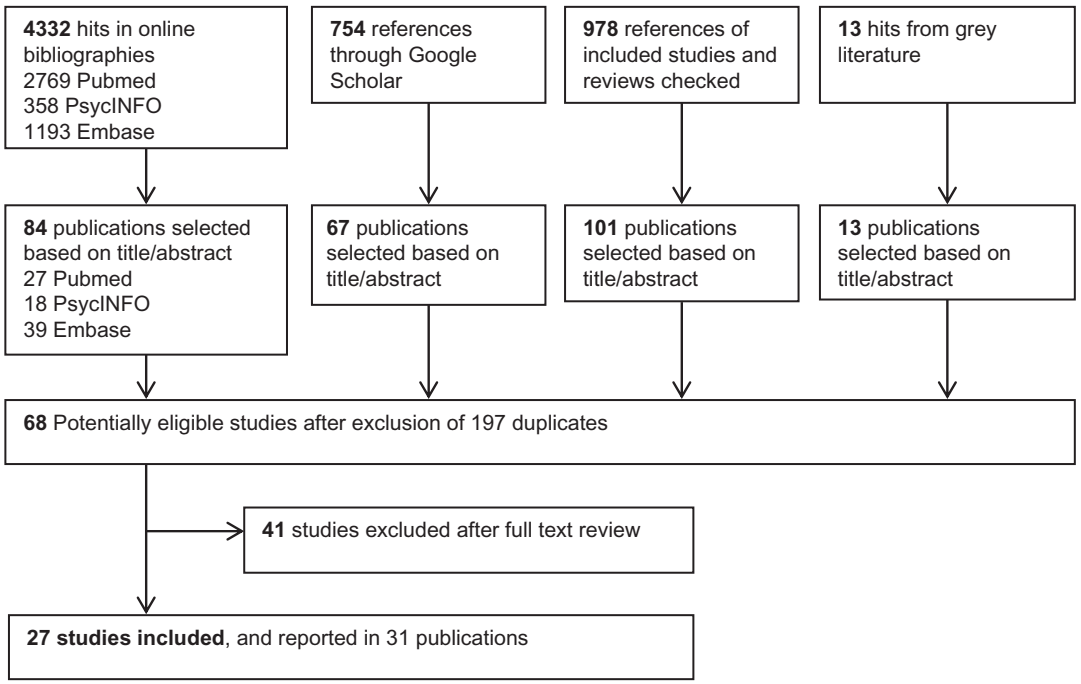

Figure 1. Flow diagram of literature search and selection.

(number of true positives, false positives, true negatives, false negatives) and calculated the 95% confidence intervals with STATA 14.0. Results were not pooled across studies.

Declaration of sources of funding

The Dutch Ministry of Health supported this work (grant number 325414). The sponsor had no role in its design or conduct, interpretation of results, and reporting.

Results

The literature search yielded 6,077 hits. The search in the online bibliographies yielded 84 potentially eligible articles, the forward citation search 67, references of reviews and articles 101, and responses of 18 authors 13. After exclusion of duplicates we assessed 68 full-texts for eligibility. Finally, 27 studies were included that were reported in 31 publications (Figure 1) [9–11, 14, 23–49]. Most excluded studies did not report test accuracy of a rapid test (see online Appendix A for references, available in *Age and Ageing* online).

Study characteristics

The studies investigated 1–20 different single or combined tests. MOTYB was studied most often (seven studies). Table 1 presents the key characteristics of the study designs. The setting was mostly a geriatric, surgical or acute care ward, or emergency department of a hospital. One study was performed in a consultation-liaison psychiatry service, one in a hospice and four studies in a nursing home. The number of participants varied between 14 and 500. The prevalence of delirium varied between 4% and 57%.

Table 2 shows the results of the risk of bias assessment. One of the 27 included studies had a low risk of bias on all items [30]. Twelve studies scored reasonably well with only one or two domains with a high or unclear risk of bias. Fourteen studies scored a high risk of bias for selection of patients due to exclusion criteria that we deemed inappropriate such as previous diagnosis of dementia [25, 34, 44] or psychiatric illness [37, 39, 44, 48], expected hospital stay of ≤2 days [11, 24], patients in rehabilitation, respite care [23], ophthalmological, or gynecological wards [46], and being too unwell or cognitively incapable to consent to participation [10, 26, 35, 45]. One study enrolled patients in office hours only [24] and another excluded patients older than 80 years [25]. In addition, almost all studies requested patients to provide informed consent before inclusion, which might have led to exclusion of relatively severe cases of delirium. Significant heterogeneity existed in the professional background of the individuals performing the index tests. Applicability concerns were low for most populations, screening tests and target conditions.

Test accuracy

Table 3 presents the sensitivity and specificity of the rapid screening instruments for delirium. Most were attention or level of arousal assessment tests. The test-time varied from 7 sec for RADAR to 3 min for combinations of tests per assessment. All tests were described as easy and requiring minimal training (up to 45 min) and minimal clinical experience. The articles described how the tests needed to be rated and which cut-offs to use (see online appendix for content of tests, available in *Age and Ageing* online).

Twenty-six studies reported results for mixed groups of patients with and without dementia. The sensitivity of single

Table 1. Characteristics of included studies

Study	Index test	Tester	Reference standard	Study population			
				N	Setting (ward, type patients)	Delirium prevalence, %	Dementia prevalence, %
Jitapunkul 1992 [23]	AMT-10	Researcher	DMS-III	184	Acute geriatric	22	18
Pompei 1995 [24]	DSF, Vigilance A, DSF + Vigilance A	Research assistants	DSM-III-R	432	Medical and surgical	15	NT
Macleod 1997 [25]	Writing name and address	Speech-language therapist	DSM-IV and DRS	20	Terminal cancer, hospice	NA (case-control)[c]	0
O'Keeffe 1997 [26]	GAR, DSF, DSB, Vigilance A, DCT-1, DCT-2	Geriatrician	DSM-III	90	Acute geriatric	21	24
Adamis 2006 [27]	Signature MMSE sentence	Neuropsychologist	CAM with DRS	94	Elderly Care Unit, hospital	32	NR
Bryson 2011 [28]	CDT	Nurses trained in psychometric testing	CAM	88	Abdominal aorta surgery	26	NA
Leung 2011 [29]	DSF, DSB, DSF + DSB	Nurses	DSM-IV	144	Acute medical and geriatric	18	22
Chester 2012 [30]	mRASS	Nurse	DSM-IV	95	Tertiary VA hospital	11	NR
Han 2013 [9]	RASS + Lunch BW (DTS)	Research assistant	DSM-IV-TR	406	Emergency department	12	6
Emerson 2014 [31]	CDT	Emergency physician					
Han 2015 [32]	RASS	Research assistant					
Lees 2013 [33]	AMT-4, AMT-10, CDT, Cog-4, GCS	Medical student	CAM	111	Acute stroke	11	41
Tieges 2013 [34]	OSLA, RASS	Graduate psychologist	CAM	30	Hip fracture	33	NA
O'Regan 2014 [10]	MOTYB, SSF5, SSF5 then MOTYB	Junior medical staff	DSM-IV	133[a]	One hospital[b]	NR[a]	NR[a]
Fick 2015 [11]	20 single items of 3D-CAM and pairs of items, ALOC	Research assistants	DSM-IV	201	General and geriatric medicine	21	28
Lin 2015 [35]	SQeeC	Geriatrician in training	DSM-IV	100	General medicine	12	30
Shoaib 2015 [36]	Pictorial Facial Scale	Nurses/nurse aids	DSM-IV	55	Acute geriatric	26	NT
Voyer 2015 [37]	RADAR	Nurse or research assistant	DSM-IV-TR (with CAM)	142	Acute care hospital and nursing home	15	4
Voyer 2016 [38]	10 items from HDS	Research assistant	DSM-V with CAM	51		4	71
Adamis, 2016 [14]	DST (DSF+DSB), Vigilance A, Serial 7s, MOTYB	Medical students (fifth year)	CAM	200	Geriatric unit	17	63
Bilodeau 2016 [39]	RADAR	Nurse-assistants	DSM-V	31	Nursing home	3	100
Hendry 2016 [40]	AMT-10, AMT-4, MOTYB	Nurse	DSM-V	500	Geriatric unit	19	32
Koop 2016 [41]	RADAR	Nurse-assistants	CAM	14	Rehabilitation ward of nursing home	7	7
Leonard 2016 [42]	World BW, MOTYB, SSF, SSB, Vigilance A, Vigilance B, CDT, IPT and combinations	Trained raters/ psychiatrists	DRS-R98-severity ≥15 or DSM-IV	193	Consultation-liaison psychiatry service	57	51
O'Regan 2016 [43]	CDT, SSF, MOTYB, IPT	Medical expert	DRS-R98	470	Emergency department	39	25
Bedard 2017 [44]	O3DY	Research assistant	CAM	305	Emergency department of four hospitals	NR	NR
Dyer 2017 [45]	AMT-4	Research assistant	CAM-ICU	220	Emergency department	13	24
Grossmann 2017 [46, 47]	mRASS, MOTYB	Nurses	DSM-IV-TR	298	Emergency department	7	14
Pelletier 2017 [48]	RADAR	Nurse-assistants	DSM-V (with CAM)	45	Nursing home	4	93
Richardson 2017 [49]	OSLA, SAVEAHAART, OSLA +SAVEAHAART	Delirium experts	DSM-V	114	Acute and rehabilitation hospitals	46	52

NR stands for not reported; NT for not tested; [a]subgroup of patients aged 69 or older; [b]all wards except ED, ICU and isolation rooms; [c]determined in 10 patients with delirium and a random sample of 10 patients without delirium; NA not applicable (all or most patients with dementia excluded at entry).

Table 2. Risk of bias and applicability of included study

Study	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient population	Index test	Target condition
Jitapunkul 1992	H	L	L	U	L	L	L
Pompei 1995	H	L	H	H	L	L	L
Macleod 1997	H	H	U	L	H	L	L
O'Keeffe 1997	L	H	L	H	L	L	L
Adamis 2006	L	L[a]	L	H	L	L	L
Bryson 2011	L	H	H	L	H	L	L
Leung 2011	H	H	L	L	L	L	L
Chester 2012	L	L	L	L	L	L	L
Han 2013	H	L	L	H	L	L	L
Lees 2013	L	L	U	L	H	L	L
Tieges 2013	H	H	H	U	H	L	L
O'Regan 2014	L	U	H	H	L	L	L
Fick 2015	H	L	L	L	L	L	H
Lin 2015	L	H	L	L	L	L	L
Shoaib 2015	U	L	L	U	L	L	L
Voyer 2015	H	L[a]	U[a]	L	L	H[b,c]	L
Adamis 2016	H	H	H	U	L	L	L
Bilodeau 2016	H	U	U	U	L	H[b]	L
Hendry 2016	L	H	L	H	H	L	L
Koop 2016	L	L	U	U	L	H[b]	L
Leonard 2016	L	H	U	H	L	L	L
O'Regan 2016	L	H	H	H	L	L	L
Bedard 2017	H	H	H	U	L	L	L
Dyer 2017	U	U	H	L	L	L	L
Grossmann 2017	H	L	L	H	L	L	L
Pelletier 2017	H	U	U	U	L	H[b]	L
Richardson 2017	U	H	H	U	L	L	L

H stands for high, L for low and U for unclear; [a]high for tests other than GAR (Adamis 2006) or RADAR (Voyer 2016); [b]RADAR required two or more medication administrations per day; [c]low for HDS.

and combinations of tests varied between 17% and 100%. Twenty-two instruments had a sensitivity of 90% or higher. Of these tests, only the RASS + Lunch BW (DTS) had a lower confidence interval limit above 90%. The specificity of the tests varied between 14% and 100%. Of the tests with sensitivity of 90% or more only the AMT-4, DCT-2, GAR, MOTYB, OSLA, RASS and 'writing name and address' had specificity of 80% or more. Sensitivity results were reproduced for AMT-4, OSLA and RASS, but sensitivity and specificity results (approximately) only for OSLA and RASS (see online Appendix A, available in *Age and Ageing* online).

Nine studies reported test accuracy of screening tools in patients with dementia. Sensitivity varied from 21% to 100%, and specificity from 15% to 96%. Eight tests had a sensitivity of 90% or higher, but only the OSLA + SAVEAHAART showed specificity of 80% or higher. None of the findings in patients with dementia have been reproduced consistently. In both groups 'older patients in general' and 'patients with dementia', six tests had high sensitivity of 90% or higher, but none had specificity of 80% or higher.

In general, confidence intervals around sensitivity and specificity were wide in most studies, indicating insufficiently large study populations. Most studies did not report interrater reliability, but if reported, it was generally high.

Discussion

We performed a systematic review of rapid and easy-to-administer screening instruments for delirium in older patients. The tools took 3 min or less to administer. The AMT-4, DCT-2, GAR, OSLA, RASS and 'writing name and address' had sensitivity above 90% and specificity above 80% in older patients in general. The OSLA + SAVEAHAART performed well in those with dementia.

Promising tests

Successful implementation of a screening delirium tool is affected by the administration time, the training required, the burden posed to the patient, and its appropriateness in the clinical setting it is used [12, 16, 21]. To minimize the burden of screening on professionals, patients and resources, and maximize the number of cases found, we and other authors propose a two-step approach [12, 30, 35, 47, 50]. A highly sensitive tool is needed in the first step to detect as many possible cases of delirium as possible (few false-negative cases), and a highly specific tool in the second step to make definitive diagnoses (few false-positive cases).

Most tests with sensitivity of 90% or more and specificity of 80% or more either require observation of level of arousal (GAR, RASS, OSLA), a combination of such a test

Table 3. Test characteristics of rapid screening instruments for delirium

Study	Test (cut-off)	Test-time, min	Sensitivity, % (95% CI)		Specificity, % (95% CI)		Inter-rater Reliability, % agreement (kappa)
			All patients	In dementia	All patients	In dementia	
Jitapunkul 1992	AMT-10 (<8)	<2	92 [78–98]	NR	65 [56–73]	NR	NT
Pompei 1995	DSF (<5)	<2	34 [22–48][a]	NT	90 [87–93]	NT	NT
	Vigilance A (>2 errors)		61 [47–74]	NT	77 [73–81]	NT	NT
	DSF + Vigilance A (both failed)		26 [15–40]	NT	97 [95–99]	NT	NT
Macleod 1997	Writing name and address	<1	100 [69–100]	NA	100 [69–100]	NA	NT
O'Keeffe 1997	GAR (<7)	Each <2	94 [73–100]	NR	99 [92–100]	NR	ICC = 0.83
	DSF (NR)		NR	NR	NR	NR	NR
	DSB (< 4)		83 [59–96]	NR	96 [88–99]	NR	NR
	Vigilance A (>2 errors)		83 [59–96]	NR	83 [72–91]	NR	NR
	DCT-1 (NR)		NR	NR	NR	NR	NR
	DCT-2 (<9 in two trials)		94 [73–100]	NR	87 [77–94]	NR	NR
Adamis 2006	Signature (abnormal)	<0.5	54 [32–76]	NR	88 [76–96]	NR	NT
	MMSE sentence (abnormal)	<0.5	NR	NR	NR	NR	
Bryson 2011	CDT (≤18)	2	66 (38–85)	NA	64 (52–76)	NA	NT
Leung 2011	DSF (<8)	<2	58 [37–77]	NR	72 [63–80]	NR	NR
	DSB (<3)	<2	81 [61–93]	NR	63 [53–71]	NR	NR
	DSF + DSB (NR)	3	NR	NR	NR	NR	NR
Chester 2012	mRASS (≠ 0)	<0.5	64 (52–76)	NR	93 (90–96)	NR	98 (0.48)
Han 2013	RASS (≠ 0) + Lunch BW (DTS) (>1 error)[e]	<1	98 (90–100)	NR	56 (51–61)	NR	89 (0.79)
Emerson 2014	CDT (abnormal)	<2	94 [84–99]	NR	44 [39–49]	NR	NR (0.84)
Han 2015	RASS (≠ 0)	<0.3	84 (74–94)	NR	88 (84–91)	NR	NR (0.63)
Lees 2013	AMT-4 (<4)	Each	83 (52–98)	NR	61 (51–71)	NR	NT
	AMT-10 (<8)	<2	75 (43–95)	NR	61 (51–71)	NR	NT
	CDT (<3 on 0–3 scale)		67 (22–96)	NR	38 (28–49)	NR	NT
	Cog-4 (>0)		70 (35–93)	NR	44 (35–55)	NR	NT
	GCS (eye open <4 & verbal response <5)		17 (02–48)	NR	81 (71–88)	NR	NT
Tieges 2013	OSLA (>3)	<1	90 [56–100]	NA	90 [68–99]	NA	NT
	RASS (≠ 0)	<1	90 [56–100]	NA	85 [62–97]	NA	NT
O'Regan 2014	MOTYB (1 error up to July)	<2	84 (68–94)	88 (68–97)	90 (82–95)	71 (29–96)	NR
	SSF (<5)	1–1.5	95 (82–99)	NR	58 (48–68)	NR	NR
	SSF (<5), then MOTYB (1 error up to July)	<3	81 (65–92)	NR	91 (83–96)	NR	NR
Fick 2015	MOTYB (1 error)[b]	<2	83 (69–93)	89 (72–98)	69 (61–76)	61 (41–78)	NR
	DSB (<4)	<2	83 (69–93)	86 (67–96)	52 (44–60)	54 (34–72)	
	Day of week (an inadequate answer)	<1	71 (55–84)	75 (55–89)	92 (87–96)	75 (55–89)	
	Day of week + MOTYB (1 fail)[c]	<3	93 (81–99)	96 (82–100)	48 (40–56)	43 (24–63)	
	Day of week + DSB (1 fail)	<3	93 (81–99)	93 (76–99)	48 (40–56)	39 (22–59)	
	DSB + MOTYB (1 fail)	<3	93 (81–99)	93 (76–99)	42 (34–50)	39 (22–59)	
	ALOC	<2	19 (9–34)	21 (8–41)	99 (96–100)	NR	
Lin 2015	SQeeC (an inadequate answer)	1	83 (52–98)	83 (36–99)	81 (72–89)	59 (36–79)	NT
Shoaib 2015	Pictorial Facial Scale (≠ 0)	<1	86 [57–98]	NT	67 [49–80]	NT	76 (0.41)
Voyer 2015	RADAR 1–4 × daily (>0 item present)	0.5	65 (43–84)	NR	71 (64–78)	NR	82–98 (0.34–0.79)
	RADAR 3–4 × daily (>0 item present)	0.5	73 (39–94)	71 (29–96)	67 (57–76)	43 (26–61)	72–100 (0.30–1.00)
Voyer 2016	Serial 7s (failure)[d]	<1	96 (78–100)	NR	14 (9–20)	NR	
	Serial 3s (failure)	<1	87 (66–97)	NR	47 (39–55)	NR	
	MOTYB (failure)	<1	83 (61–95)	63 (24–91)	63 (55–70)	79 (71–86)	

	Days of week B (failure)	<1	48 (27–69)	NR	85 (79–90)	NR	
	Counting 93 > 85 (failure)	<1	48 (27–69)	NR	85 (79–90)	NR	
	Counting 10 > 1 (failure)	<1	30 (13–53)	NR	87 (81–92)	NR	
	MOTYF (failure)	<1	26 (10–48)	NR	89 (84–94)	NR	
	Days of week F (failure)	<1	26 (10–48)	NR	90 (84–94)	NR	
	Counting 1 > 10 (failure)	<1	17 (5–39)	NR	91 (86–95)	NR	
	Counting 10 objects (failure)	<1	17 (5–39)	NR	94 (89–97)	NR	
Adamis 2016	DST (>1)	<2	74 (55–87)	NT	62 (54–69)	NT	90% during pre-study training
	Vigilance A (>2 errors)	<2	82 (65–93)	NT	60 (52–68)	NT	
	Serial 7s (>2 errors)	<2	91 (75–98)	NT	46 (38–54)	NT	
	MOTYB (up to July >0 error)	<2	82 (65–93)	NT	66 (58–73)	NT	
Bilodeau 2016	RADAR (>0 item present)	7 sec	NA	100 (3–100)	NA	77 (58–90)	94–99 (0.76–1.00)
Hendry 2016	AMT-10 (<5)	<2	87 (77–93)	NR	64 (58–69)	NR	NT
	AMT-4 (<4)	<2	93 (85–97)	NR	54 (48–59)	NR	
	MOTYB (<6)	2	91 (83–96)	NR	50 (44–55)	NR	
Koop 2016	RADAR (>0 item present)	7 sec	100 [3–100][a]	NR	69 [39–91]	NR	90 (0.08)
Leonard 2016	World BW (>0 error)	< 1	90 (83–95)	NR	41 (30–52)	NR	90% during pre-study training
	MOTYB (>1 error up to July)	< 2	75 (67–83)	NR	70 (59–80)	NR	
	SSF (<5)	1–1.5	75 (66–83)	NR	56 (45–68)	NR	
	SSB (<3)	1–1.5	77 (68–84)	NR	58 (47–69)	NR	
	Vigilance A (<27)	1–1.5	75 (66–83)	NR	73 (72–90)	NR	
	Vigilance B (<18)	1.5	94 (88–98)	NR	56 (45–67)	NR	
	CDT (<6 on Sunderland rating)	2 min	72 (63–81)	NR	64 (52–74)	NR	
	IPT (<4)	2	71 (61–79)	NR	73 (61–82)	NR	
	MOTYB + Vigilance A (1 fail)	<3	91 (84–96)	NR	59 (48–70)	NR	
O'Regan 2016	CDT (<10 on 15-point scale)	Each	81 (72–88)	NR	63 (57–69)	NR	NR
	SSF (<5)	<2	90 (84–94)	90 (75–97)	41 (35–47)	25 (3–43)	
	MOTYB (>0 error up to July)		85 (78–90)	83 (68–93)	58 (52–64)	33 (19–51)	
	IPT (>0 error)		93 (86–96)	87 (69–96)	40 (34–46)	15 (6–32)	
	20 > 1 (NR)		70 (62–77)	66 (49–79)	69 (63–74)	44 (27–62)	
Bedard 2017	O3DY (<4)	<1	85 (62–97)	NR	58 (52–64)	NR	NR
Dyer 2017	AMT-4 (<4)	≤1	92 [75–99]	100 [82–100]	82 [75–87]	52 [37–67]	NT
Grossmann 2017	mRASS (≠0)	0.5	70 (48–85)	55 (28–79)	93 (90–96)	83 (66–93)	NT
	MOTYB in 30s (>2 errors or >1 error & >30s)	0.5	95 (76–99)	100 (74–100)	86 (81–90)	63 (46–78)	
Pelletier 2017	RADAR (>0 item present)	<1	100 (16–100)	NR	72 (59–86)	NR	94–97 (0.44–0.70)
Richardson 2017	OSLA (>3)	<1	85 [72–93]	74 [55–88]	82 [71–91]	96 [82–100]	NT
	SAVEAHAART (>3 errors)	<1	90 [79–97]	84 [66–95]	64 [51–76]	73 [51–87]	
	OSLA + SAVEAHAART (>9)	<2	84 [72–93]	94 [79–99]	97 [89–100]	92 [77–99]	

[ ] CI in squared brackets were calculated with data in article; ICC for intraclass correlation coefficient, NA for not applicable, NR for not reported.

[a]Score closest to day of delirium diagnosis.

[b]Top three single items.

[c]Top three pairs of items.

[d]Because the items are arranged in descending order of difficulty, researchers had to assume that participants had succeeded in the easier items and failed the more difficult ones, even if the research assistant had not necessarily administered them.

[e]Sensitivity and specificity of DTS, CDT and RASS administered by research assistant (almost similar when administered by physician).

with an attention test (OSLA + SAVEAHAART), or multiple cognitive tests (AMT-4, DCT-2). Sensitivity results were reproduced for AMT-4, OSLA and RASS, but specificity results only for OSLA and RASS in general older populations. Remarkably, both latter tests are level of arousal tests. The OSLA + SAVEHAART might perform well in terms of sensitivity and specificity in patients with dementia. Hence, level of arousal also seems to distinguish delirium from dementia. However, these study results have not always been reproduced and study populations were sometimes small.

There was no apparent relationship of test accuracies with risk of bias, delirium criteria used, and prevalence of delirium. Naturally, reported test accuracies need to be interpreted with caution because most studies reported those that would correctly classify most patients, delirious or not. In other words, high sensitivity was not always the aim, and would have been achieved if lower specificity had been accepted. With high applicability of patient populations, index tests and target condition, indirectness is not a serious concern.

Our review complements the findings of a prior study about single-item screening questions for delirium [12]. Such questions are short too but probe (subjective) symptoms of delirium such as confusion and hallucinations with the patient, surrogate or a health professional [11, 33, 51]. Sensitivity was mostly poor, but specificity sometimes very high. Other reviews about delirium instruments did not focus on short tests and did not capture the recently published tests [7, 15, 16, 19].

Methodological challenges

Performing a diagnostic test accuracy study in patients with delirium might be challenging. All studies required patients or their legal representatives to provide informed consent. It is likely that patients with delirium and their families will not give permission due to lack of cognitive and decisional capacity as easily as patients without delirium and their families [52]. As a result, patients with (severe) delirium may not have been represented sufficiently in the study populations. The use of exclusion criteria such as 'included only in office hours' [9], 'an expected hospital stay of ≤2 days' [11], 'dementia with MMSE 10 or less' [26] and 'not able to speak English' [9, 35] might have negatively influenced the number and diversity of included delirium cases too [20]. In addition, exclusion of patients with dementia might have led to overestimation of specificity, because symptoms of severe dementia overlap considerably with symptoms of delirium [26, 33, 35, 37, 43].

All studies were performed in hospitalized patients, except four studies that tested RADAR in nursing home patients and one study in a hospice [25, 37, 39, 41, 48]. In one hospital and the nursing home studies, delirium prevalence was 4–7%, lower than estimates from prior observational studies [2]. Cases may have been missed [53]. Additional studies are needed in nursing homes and hospices. Due to the overlap between delirium and (neuropsychiatric symptoms of) dementia, diagnostic expertise is needed to ensure valid reference diagnoses.

When performing a diagnostic test study in delirium, researchers need to consider how to score untestable patients. In some studies, patients were excluded if they were considered too ill or too drowsy. Many of these patients might have had a delirium. Twelve studies reported that untestable patients were considered screen-positive [9–11, 14, 25, 27, 29, 36, 39–41, 46]. We agree with this approach. Delirium will probably be missed less often if untestable patients are scored as screen-positive.

Strengths and limitations

Strength of our study was that we performed a broad search with no restriction related to publication year or language. We used the internationally accepted QUADAS-2 tool to assess risk of bias in diagnostic test accuracy studies. There were pairs of independent data extractors and they used a consensus procedure for disagreements. Our review meets the PRISMA criteria for reporting a review.

As we chose to exclude tests based on surrogate information, some relatively quick (diagnostic) tools were excluded, such as the bCAM [50], 3D-CAM [54], Nu-DESC [55] and 4AT [56]. They seemed to perform (very) well in general older patient populations and patients with dementia. Serious games and mobile computerized tests present interesting options too [57, 58]. Another limitation of our study is that our results are not generalizable to non-older or ICU patients, because we did not include studies in such patients.

Finally, test accuracy studies do not measure outcomes of implementing screening tools. Professionals have reported that they do not always believe that screening will lead to better treatment [59]. This is conceivable in younger patients with a clear underlying disease. Delirium in frail older patients, who often have multiple modifiable predisposing and precipitating conditions, would probably remain undetected and inadequately treated if it is not diagnosed [60, 61]. A diagnosis is also important for adequate psycho-education of patients, relatives and caregivers.

Conclusion

We identified 27 studies that investigated test accuracy of rapid and easy-to-administer bedside delirium screening instruments in older patients. All except one study had at least one source of potential bias. Two tests had high sensitivity and high specificity in more than one study among older hospitalized patients: the OSLA and RASS. Tests of arousal seemed to perform well in patients with dementia too, but results need to be reproduced in larger populations and long-term care settings. The advantage of rapid and frequent screening by non-specialized personnel will be that only screen-positive patients need an extensive diagnostic work-up by a medical specialist.

Key points

- Delirium is often missed, and screening with a rapid and easy-to-use instrument might improve recognition.
- Our review identified two arousal tests—Observational Scale of Level of Arousal (OSLA) and Richmond Agitation and Sedation Scale (RASS)—that identified most delirium cases in older patients.
- Attention and orientation tests had high sensitivity too, but were generally less specific.
- Information about rapid screening tools for delirium in patients with dementia was scarce.

Supplementary data

Supplementary data mentioned in the text are available to subscribers in *Age and Ageing* online.

Acknowledgements

We would like to thank the authors of the included studies who provided additional information our request, and R. Lees and T.J. Quinn for sharing data.

Conflict of interest

The authors do not have any conflicts of interest to disclose.

Abbreviations

4AT: 4 Attention Tests
ALOC: Altered level of consciousness
AMT: Abbreviated Mental Test
AMT-4: 4-point Abbreviated Mental Test
AMT-10: 10-point Abbreviated Mental Test
CAM: Confusion Assessment Method
CDT: Clock Drawing Test
Cog-4: Cognitive examination derived from NIH Stroke Scale (NIHSS)
CRS: Confusion Rating Scale
DCT: Digit Cancellation Test
DCT-1: Digit Cancellation Test with a 1-digit matrix
DCT-2: Digit Cancellation Test with a 2-digit matrix
DOSS: Delirium Observation Screening Scale
DRS-R98: Delirium Rating Scale Revised 1998
DSB: Digit Span Backwards
DSI: Delirium Symptom Interview
DSF: Digit Span Forward
DST: Digit Span Test (DSF + DSB)
DTS: Delirium Triage Screen
DW: Day of the week
GAR: Global Attentiveness Rating
GCS: Glasgow Coma Scale
HDS: Hierarchic Dementia Scale
IPT: Intersecting Pentagons Test
Lunch BW: Lunch spelled backwards
MDAS: Memorial Delirium Assessment Scale
MOTYB: Months of the year recited backwards
MOTYF: Month of the year recited forwards
mRASS: Modified Richmond Agitation and Sedation Scale
Nu-DESC: Nursing Delirium Screening Scale
OSLA: Observational Scale of Level of Arousal
O3DY: Ottowa Day, Date, WORLD BW and Year
RADAR: Recognizing Acute Delirium As part of your Routine
RASS: Richmond Agitation and Sedation Scale
SAVEAHAART: S-A-V-E-A-H-A-A-R-T
Serial 3s: Serial threes subtraction test
Serial 7s: Serial sevens subtraction test
SSB: Spatial span backwards test
SSF: Spatial span forward test
SQeeC: The Simple Query for Easy Evaluation of Consciousness
World BW: World spelled backwards

References

Please note a full list of references is available in the supplementary data.

9. Han JH, Wilson A, Vasilevskis EE *et al*. Diagnosing delirium in older emergency department patients: validity and reliability of the delirium triage screen and the brief confusion assessment method. Ann Emerg Med 2013; 62: 457–65.

10. O'Regan NA, Ryan DJ, Boland E *et al*. Attention! A good bedside test for delirium? J Neurol Neurosurg Psychiatry 2014; 85: 1122–31.

11. Fick DM, Inouye SK, Guess J *et al*. Preliminary development of an ultrabrief two-item bedside test for delirium. J Hosp Med 2015; 10: 645–50.

14. Adamis D, Meagher D, Murray O *et al*. Evaluating attention in delirium: a comparison of bedside tests of attention. Geriatr Gerontol Int 2016; 16: 1028–35.

23. Jitapunkel S, Pillay I, Ebrahim S. Delirium in new admitted elderly patients: a prospective study. Quaterly J Med New Ser 83 1992; 300: 307–14.

24. Pompei P, Foreman M, Cassel CK *et al*. Detecting delirium among hospitalized older patients. Arch Intern Med 1995; 155: 301–7.

25. Macleod A, Whitehead L. Dysgraphia and terminal delirium. Palliat Med 1997; 11: 127–32.

26. O'Keeffe ST, Gosney MA. Assessing attentiveness in older hospital patients: global assessment versus tests of attention. JAGS 1997; 45: 470–3.

27. Adamis D, Reich S, Treloar A *et al*. Dysgraphia in elderly delirious medical inpatients. Aging Clin Exp Res 2006; 18: 334–9.

28. Bryson GL, Wyand A, Wozny D *et al*. The clock drawing test is a poor screening tool for postoperative delirium and cognitive dysfunction after aortic repair. Can J Anesth/J Can Anesth 2011; 58: 267–74.

29. Leung JLM, Lee GTH, Lam YH *et al*. The use of the Digit Span Test in screening for cognitive impairment in acute medical inpatients. Int Psychogeriatr 2011; 23: 1569–74.

30. Chester JG, Harrington MB, Rudolph JL. Serial administration of a modified richmond agitation and sedation scale for delirium screening. J Hosp Med 2012; 7: 450–3.

31. Emerson G, Carlson R, Nicolson S, Han J. The clinical utility of the clock drawing test in detecting delirium in older emergency department patients. Ann Emerg Med 2014; 64: S5.

32. Han JH, Vasilevskis EE, Schnelle JF *et al*. The diagnostic performance of the Richmond. Acad Emerg Med 2015; 22: 878–82.

33. Lees R, Corbet S, Johnston C *et al*. Test accuracy of short screening tests for diagnosis of delirium or cognitive impairment in an acute stroke unit setting. Stroke 2013; 44: 3078–83.

34. Tieges Z, McGrath A, Hall RJ *et al*. Abnormal level of arousal as a predictor of delirium and inattention: an exploratory study. Am J Geriatr Psychiatry 2013; 21: 1244–53.

35. Lin HS, Eeles E, Pandy S *et al*. Screening in delirium: a pilot study of two screening tools, the simple query for easy evaluation of consciousness and simple question in delirium. Australas J Ageing 2015; 34: 259–64.

36. Shoaib A, Kaehr E, Malmstrom T *et al*. Mental status (MS) vital sign: a pictorial facial scale as a screening tool for delirium. JAGS 2015; s116.

37. Voyer P, Champoux N, Desrosiers J *et al*. Recognizing acute delirium as part of your routine [RADAR]: a validation study. BMC Nurs 2015; 14: 19.

38. Voyer P, Champoux N, Desrosiers J, Landreville P *et al*. Assessment of inattention in the context of delirium screening: one size does not fit all! Int Psychogeriatrics 2016; 28: 1293–301.

39. Bilodeau C, Voyer P. Radar: un outil valide pour le repérage du syndrome confusionnel aigu (delirium) en résidences intermédiaires. NPG Neurol - Psychiatr - Geriatr 2016; 17: 144–51.

40. Hendry K, Quinn TJ, Evans J *et al*. Evaluation of delirium screening tools in geriatric medical inpatients: a diagnostic test accuracy study. Age Ageing 2016; 45: 832–7.

41. Koop R, Notter J. Delirium in nursing homes. Res Diss Master Sci Nurs / Adv Heal Care 2016.

42. Leonard M, Exton C, Cullen W *et al*. Attention, vigilance and visuospatial function in hospitalized elderly medical patients: Relationship to neurocognitive diagnosis. J Psychosom Res 2016; 90: 84–90.

43. O'Regan NA, Maughan K, Liddy N *et al*. Five short screening tests in the detection of prevalent delirium: diagnostic accuracy and performance in different neurocognitive subgroups. Int J Geriatr Psychiatry 2017; 32: 1440–9.

44. Bédard C, Voyer P, Eagles D *et al*. LO57: validation of the Ottawa 3DY in community seniors in the ED. CJEM 2017; 19: S47.

45. Dyer AH, Briggs R, Nabeel S *et al*. The Abbreviated Mental Test 4 for cognitive screening of older adults presenting to the Emergency Department. Eur J Emerg Med 2017; 24: 417–22.

46. Grossmann FF, Hasemann W, Kressig RW *et al*. Performance of the modified Richmond Agitation Sedation Scale in identifying delirium in older ED patients. Am J Emerg Med 2017; 35: 1324–6.

47. Hasemann W, Grossmann FF, Stadler R *et al*. Screening and detection of delirium in older ED patients: performance of the modified Confusion Assessment Method for the Emergency Department (mCAM-ED). A two-step tool. Intern Emerg Med 2017; 123456789: 1–8.

48. Pelletier I, Voyer P. L'utilisation du RADAR en centre d'hébergement et de soins de longue durée (CHSLD) pendant sept journées consécutives. 2017.

49. Richardson SJ, Davis DHJ, Bellelli G *et al*. Detecting delirium superimposed on dementia: diagnostic accuracy of a simple combined arousal and attention testing procedure. Int Psychogeriatrics 2017; 29: 1585–93.

Received 28 June 2017; editorial decision 28 February 2018