BMJ Open

Reliability and accuracy of delirium assessments among investigators at multiple international centres

Hannah R Maybrier,[1] Angela M Mickle,[1] Krisztina E Escallier,[1] Nan Lin,[2,3] Eva M Schmitt,[4] Ravi T Upadhyayula,[1] Troy S Wildes,[1] George A Mashour,[5] Kerry Palihnich,[4] Sharon K Inouye,[4,6] Michael Simon Avidan,[1] on behalf of the PODCAST Research Group

To cite: Maybrier HR, Mickle AM, Escallier KE, *et al*. Reliability and accuracy of delirium assessments among investigators at multiple international centres. *BMJ Open* 2018;**8**:e023137. doi:10.1136/bmjopen-2018-023137

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-023137).

SKI and MSA contributed equally.

Received 15 May 2018
Revised 22 October 2018
Accepted 25 October 2018

For numbered affiliations see end of article.

Correspondence to
Dr Michael Simon Avidan; avidanm@wustl.edu

ABSTRACT

Introduction Delirium is a common, serious postoperative complication. For clinical studies to generate valid findings, delirium assessments must be standardised and administered accurately by independent researchers. The Confusion Assessment Method (CAM) is a widely used delirium assessment tool. The objective was to determine whether implementing a standardised CAM training protocol for researchers at multiple international sites yields reliable inter-rater assessment and accurate delirium diagnosis.

Methods Patients consented to video recordings of CAM delirium assessments for research purposes. Raters underwent structured training in CAM administration. Training entailed didactic education, role-playing with intensive feedback, apprenticeship with experienced researchers and group discussions of complex cases. Raters independently viewed and scored nine video-recorded CAM interviews. Inter-rater reliability was determined using Fleiss kappa. Accuracy was judged by comparing raters' scores with those of an expert delirium researcher.

Results Twenty-seven raters from eight international research centres completed the study and achieved almost perfect agreement for overall delirium diagnosis, kappa=0.88 (95% CI 0.85 to 0.92). Agreement of the four core CAM features ranged from fair to substantial. The sensitivity and specificity for identifying delirium were 72% (95% CI 60% to 81%) and 99% (95% CI 96% to 100%), considering an expert rater's scores as the reference standard (delirious, n=3; non-delirious, n=6). Delirium severity ratings were tightly clustered, with most scores within 5% of the median.

Conclusion Our results demonstrate that, with appropriate training and ongoing scoring discussions, researchers at multiple sites can reliably detect delirium in postsurgical patients. These results support the premise that methodologically rigorous multi-centre studies can yield standardised and accurate determinations of delirium.

Strengths and limitations of this study

- Patients assessed were representative of adults older than 60 recovering from major surgery in the early postoperative period.
- Participation was by a multidisciplinary, international group of raters.
- The determination of delirium severity as well as its binary appraisal was helpful, since all nine patients had some features of delirium.
- Video-recording modality might hinder interpretation of subtle features.
- Sensitivity, specificity and positive and negative predictive value calculations assumed that the expert rater provided a reliable reference standard assessment of delirium.

INTRODUCTION

Delirium is an acute change in cognition, manifesting predominantly as inattention and disorganised thinking. In 2010, the US Census Bureau reported that the proportion of the US population over the age of 65 is 12%,[1] and the population of older adults is anticipated to increase substantially in the near future. It is estimated that 30%–50% of older postsurgical patients experience delirium,[2] which is associated with longer intensive care unit (ICU) and hospital stays, and increased morbidity and mortality.[3] Considering the major impact that delirium is having on patients and healthcare in our rapidly ageing society, it is important to conduct rigorous multicentre, international research focusing on prevention and treatment of delirium.

In order to effectively research delirium, adequate tools for measurement must be available. A systematic review performed in 2015 found that the Confusion Assessment Method (CAM) was the most widely used tool to identify delirium in hospital patients.[4] The CAM uses a structured patient interview including tests for attention, memory, orientation and patient self-report of delirium symptoms. After the patient interview, a

rater uses qualitative and quantitative scales to record whether 12 delirium features are present, their severity and if the features fluctuate during the interview. The 12 evaluated features are (1) acute change, (2) inattention, (3) memory impairment, (4) disorganised thinking, (5) altered level of consciousness, (6) disorientation, (7) perceptual disturbances, (8) delusions, (9) psychomotor agitation, (10) psychomotor retardation, (11) sleep–wake cycle disturbance and (12) inappropriate behaviour. Delirium is detected using an algorithm based on four of these features; CAM diagnostic criteria is fulfilled by the presence of (1) either acute change or fluctuation, (2) inattention and (3) either disorganised thinking or (4) an altered level of consciousness. Of note, severity of delirium is calculated using the CAM-S, which considers all features of the CAM except delusions and inappropriate behaviour.[5] When applied to the perioperative setting, acute change would include any change after surgery that is new or worse when compared with the presurgical assessment. Fluctuation is any change in the presence or severity of a feature during the interview. Four of the CAM features are eligible for the determination of fluctuation: inattention, disorganised thinking, altered level of consciousness, psychomotor agitation and psychomotor retardation. Inattention is tested directly with widely used, brief screening evaluations, including days of the week backwards, months of the year backwards and repetition of digits in forward and reversed sequences (digits forwards and backwards). Evidence of inattention is also recorded throughout the interview, such as failure in following instructions, perseverating on a previous question or requiring questions to be repeated. Disorganised thinking is assessed via qualitative evidence including but not limited to faulty reasoning, illogical thought patterns, tangential or circumstantial speech, poverty of thought, non-sensical speech and evidence of severe disorientation. Altered level of consciousness is defined as an increased or decreased responsiveness to stimuli. According to the CAM scoring guidelines, somnolence or falling asleep during the interview is a manifestation of a decreased level of consciousness; hypervigilance, by contrast, is a sign of increased level of consciousness. The CAM was validated in 1990 and was estimated to be 94%–100% sensitive and 90%–95% specific when compared with a psychiatric assessment.[6] The objective of this study was to determine whether implementation of a standardised delirium assessment training protocol for researchers at multiple international sites would yield reliable inter-rater and accurate assessment of delirium.

METHODS

Study raters

Raters were members of the Prevention of Delirium and Complications Associated with Surgical Treatment (PODCAST: NCT01690988) trial research team, who were trained to assess for delirium with the CAM. PODCAST was a multicentre, randomised controlled trial that tested the hypothesis that a subanaesthetic dose of ketamine before surgery would decrease postoperative delirium incidence and pain severity.[7] A published protocol is available for review.[8]

Written, informed consent was obtained from both patients who participated in video recordings and researchers who participated in this study. Raters included researchers from Washington University in St Louis, Missouri, Memorial Sloan Kettering in New York City, New York, Weill Cornell in New York City, New York, Hartford Hospital in Hartford, Connecticut, University of Michigan in Ann Arbour, Michigan, Harvard University in Boston, Massachusetts, University of Manitoba in Winnipeg, Manitoba, Canada and Asan Medical Centre in Seoul, South Korea.

Patient and public involvement

Delirium assessment reliability is linked to the well-being of patients and their families. By reducing measurement error, delirium research is likely to be more rigorous and impactful. This could result in accurate delirium detection, which could in turn promote early and appropriate management, as well as prevention of negative consequences. Patients were included in the trial via video-recorded interviews. On consent, patients were informed that their involvement would help educate future delirium researchers. Results of this study will be disseminated to patients via public forums.

CAM training

The raters completed a rigorous training regimen in preparation for the PODCAST trial. Training began with a 3-hour didactic session on the conduct and scoring of the CAM. This included the independent scoring of a video-recorded patient interview by raters, followed by evaluation of scoring accuracy and additional training focusing on areas with deficits. The trainee then shadowed a trained rater while interviewing patients, independently scoring each interview and comparing CAM results with the trainer. Trained raters were researchers who successfully completed the training protocol or attended a comprehensive training session developed by the Hospital Elder Life Program, a division of The Center of Excellence for Delirium in Aging: Research, Training, and Educational Enhancement (https://www.hospitalelderlifeprogram.org/). To successfully complete training, the trainee's independent ratings on the presence or absence of all 12 features of the CAM, including the presence of fluctuation, were required to be in alignment with the trained rater for two delirious and two non-delirious patients. Importantly, this detailed training approach mitigates gestalt-driven learning, which could be present if only agreement on the binary outcome were required. The trainee also had to satisfactorily complete two proctored interviews independently. In general, training took several weeks (or approximately 20 cumulative hours) to complete.

All international research sites used a validated version of the CAM instrument. Versions of the CAM instrument were forward translated and back translated from English to Korean. The final validated Korean translation of the CAM was approved by the Hospital Elder Life Program.

Data collection

Researchers at Washington University School of Medicine acquired Institutional Review Board permission to video-record patients after surgery, and patients provided written informed consent for the video to be used for education and research activities. All interviewed patients were 60 or older and were within 3 days of major surgery requiring general anaesthesia at Barnes Jewish Hospital in St Louis, Missouri. Patients were selected sequentially by surgery date. The first nine consenting patients with good quality video recordings were included in this study. One interview per patient was used in the videos. Three of the nine patients met CAM criteria for delirium (33%) according to the expert assessor, who served as the reference standard for this study.

Study raters were instructed via email to independently view videos of nine patient interviews through a password-protected Vimeo (New York, New York) account. Raters independently scored patients with the CAM instrument and recorded their scores in a REDCap (Research Electronic Data Capture) database. REDCap is a secure, web-based application designed to support data capture for research studies using an intuitive interface for validated data entry.[9] The presence, absence and fluctuation of the 12 features of the CAM as well as overall delirium diagnosis and severity were collected. Although discussion between raters was encouraged during training, it was strictly prohibited while scoring these videos. Therefore, our results are conservative. All data were entered anonymously. Raters were also asked to complete a brief questionnaire indicating years of experience with the CAM, clinical background, primary language and highest level of education.

Reference standard for CAM analysis

The reference CAM scores for this analysis were determined by an expert rater (KP) with more than 20 years of experience conducting and scoring CAM assessments. This expert was the sole rater at her site, which served as a consultative rather than an enrolment site for the PODCAST study. The rater who served as the reference standard followed the same CAM scoring guidelines as all other raters, detailed above. She was blinded to all other rater's interpretations.

Statistical analyses

Inter-rater reliability among all 27 raters was calculated with the Fleiss kappa. The Landis and Koch benchmark scale was used to interpret the strength of agreement for Fleiss kappa values, according to the following: ≤0 poor; 0 to 0.2 slight; 0.21 to 0.4 fair; 0.41 to 0.6 moderate; 0.61 to 0.8 moderate; and 0.81 to 1 almost perfect.[10] The R package 'raster' was used to calculate the Fleiss kappa for overall diagnosis and four features of CAM algorithm with 95% CI. The reference scores were used to determine sensitivity and specificity. The R package 'epiR' was used to calculate the pooled sensitivity and specificity. The CAM severity ratings across researchers for all nine patients were presented descriptively, as medians, IQR and full ranges.

Table 1 Characteristics of raters

	Total (n=27)	Primary role		
		Non-nurse research (n=16)	Clinician (n=10)	Reference standard (n=1)
Highest level of education				
Bachelor's degree	9 (33%)	8	0	1
Master's degree	7 (26%)	3	4	0
Medical degree	11 (41%)	5	6	0
English as primary language	22 (81%)	12	9	1
Prior delirium experience*				
Clinical setting	14 (52%)	5	8	1
Research setting	6 (22%)	1	4	1
CAM instrument	6 (22%)	1	4	1

*Categories not mutually exclusive.
CAM, Confusion Assessment Method.

RESULTS

Twenty-seven raters submitted complete scores for nine patient videos. Most raters were native English speaking, non-nurse research staff and held a graduate or professional degree. Characteristics of raters are listed in table 1. Characteristics of the nine interviewed patients are detailed in table 2.

Inter-rater reliability

Agreement of overall delirium diagnosis among raters was almost perfect, with kappa=0.88 (0.85–0.92). Agreement in relation to the key features of the CAM diagnostic algorithm varied from fair to substantial (table 3). The lowest agreement was for fluctuation, kappa=0.40; the highest agreement was for disorganised thinking, kappa=0.79. Intrasite agreements were also determined for sites with at least two raters (table 3). Agreement on overall delirium diagnosis was substantial or almost perfect for all locations; however, there was varying agreement for individual features.

Table 2 Characteristics of patients interviewed

	Subjects (n=9)
Age (median (IQR))	66 (66–70)
Race	
White	7 (78%)
Black	2 (22%)
Sex: female	3 (33%)
Ethnicity: non-Hispanic	9 (100%)
Level of education*	
Less than high school graduate	1 (11%)
High school graduate	2 (22%)
Some college, no degree	4 (44%)
Bachelor's degree	1 (11%)
Prior history of delirium	2 (22%)
Alcohol drinks per week	
Less than one	6 (67%)
Three to four	1 (11%)
Five to ten	1 (11%)
Twenty-one to thirty	1 (11%)
Short Blessed Score	
Normal cognition (0–4)	8 (89%)
Questionable impairment (5-9)	1 (11%)
Lawton iADL	
High function (score of 8)	9 (100%)
Depression (PHQ-8 ≥10)	0 (0%)
Surgery type	
Cardiac	3 (33%)
Gynaecological	1 (11%)
Hepatobiliary-Pancreatic	2 (22%)
Urological	2 (22%)
Vascular	1 (11%)

*One patient with missing data.
iADL, independent activities of daily living; PHQ-8, eight-item Patient Health Questionnaire depression scale.

Sensitivity and specificity

Sensitivity and specificity for the determination of delirium using the CAM were assessed with the assumption that the expert rater provided a reference standard. The analysis resulted in an overall sensitivity of 72% and specificity of 99% with a disease (delirium) prevalence of 33%. Sensitivities and specificities for individual features of the CAM are listed in table 4.

Descriptive statistics for the delirium severity ratings are presented in figure 1 legend. The distribution of delirium severity ratings for scored by the 27 researchers across the nine patient videos is shown. Delirium severity ratings were tightly clustered for each patient video, with most raters scoring within one point (ie, ±5%) of the median severity score of all raters.

Video 1 median score 0 (IQR 0.0–0.0), reference standard score=0; video 2 median score 11 (IQR 11.0–12.1), reference standard score=10; video 3 median score 0 (IQR 0.0–0.0), reference standard score=0; video 4 median score 4 (IQR 3.0–4.0), reference standard score=3; video 5 median score 5 (IQR 5.0–6.0), reference standard score=6; video 6 median score 2 (IQR 2.0–3.0), reference standard score=2; video 7 median score 4 (IQR 4.0–5.0), reference standard score=4; video 8 median score 2 (IQR 2.0–2.0), reference standard score=1; video 9 median score 9 (IQR 8.0–10.0), reference standard score=11. The patients shown in videos 2, 5 and 9 were determined to be CAM positive by the expert rater. Whiskers represent 10th and 90th percentiles, circles represent reference standard scores. Per reference standard: 33% (3/9) observed cases with delirium.

DISCUSSION

Overall, we found almost perfect inter-rater reliability in overall delirium determination following our standardised training protocol. However, results for certain features of delirium were more varied. When considering individual features of delirium, there was substantial agreement for disorganised thinking, whereas there was only fair agreement in determination of fluctuation. Compared with the expert reference assessor, sensitivity of the researchers was good, and specificity was excellent. For all nine patients, the delirium severity ratings of the researchers were tightly clustered, with the majority scoring within ±5% of the median severity score. The tight clustering of delirium severity ratings is of particular importance, since severity rather than presence is now considered to be more important as a primary outcome of delirium studies.

The concordance in assessment of disorganised thinking was surprising, since this is often anecdotally considered the most subjective feature by our group and other CAM experts. It is particularly important to detect disorganised thinking appropriately, as it is often the tie-breaking criterion for delirium determination. Presence of fluctuation can be subtle and might have been difficult to appreciate from a video recording. Also, since fluctuation can be ascertained from one of several features (inattention, disorganised thinking, altered level of consciousness, psychomotor retardation and psychomotor agitation), discrepancies among raters can easily arise. Our result of 72% sensitivity for delirium diagnosis suggests that assessors trained with the described methodology might misdiagnose patients with delirium as not having delirium 28% of the time. Taking a closer look at the discrepancies in overall diagnosis, there were 22 false negatives and one false positive. All 22 false negatives were attributed to one particular patient video; raters correctly identified acute change and inattention but incorrectly determined disorganised thinking as not present. The one false positive determination was due to an incorrect conclusion that disorganised thinking and inattention were

Table 3 Fleiss kappa calculations for overall diagnosis and five features of CAM algorithm with 95% CIs

	Overall diagnosis*	CAM feature				
		Acute change	Fluctuation	Inattention	Disorganised thinking	Altered LOC
All sites (n=27)	0.88 (0.85 to 0.92)	0.62 (0.59 to 0.67)	0.40 (0.37 to 0.43)	0.60 (0.56 to 0.63)	0.79 (0.76 to 0.83)	0.58 (0.55 to 0.62)
Site 1 (n=10)	0.94 (0.11 to 1.0)	0.46 (0.36 to 0.56)	0.46 (0.37 to 0.56)	0.60 (0.56 to 0.64)	0.74 (0.64 to 0.84)	0.62 (0.53 to 0.72)
Site 2 (n=4)	0.72 (0.45 to 0.98)	0.23 (–0.04 to 0.5)	0.26 (–0.01 to 0.53)	0.45 (0.35 to 0.55)	0.72 (0.46 to 0.99)	0.44 (0.17 to 0.70)
Site 3 (n=4)	0.85 (0.33 to 1.0)	1.0 (0.33 to 1.0)	0.67 (0.40 to 0.93)	0.60 (0.56 to 0.63)	0.85 (0.33 to 1.0)	0.73 (0.33 to 1.0)
Site 4 (n=3)	1.0 (0.5 to 1.0)	0.41 (0.36 to 0.79)	0.23 (–0.15 to 0.61)	–0.08 (–0.46 to 0.29)	0.81 (0.50 to 1.0)	0.57 (0.19 to 0.95)
Site 5 (n=3)	0.81 (0.5 to 1.0)	1.0 (0.50 to 1.0)	0.40 (0.02 to 0.78)	0.16 (–0.04 to 0.37)	0.81 (0.50 to 1.0)	0.71 (0.5 to 1.0)

*Per reference standard: 33% (3/9) observed cases with delirium.
CAM, Confusion Assessment Method; LOC, level of consciousness.

present during the interview. Even though concordance for disorganised thinking was relatively high compared with other features, these results confirm the notion that improved delirium detection is largely dependent on the rater's ability to identify disorganised thinking. Furthermore, this slightly suboptimal sensitivity might be mitigated by serial delirium assessments and other methods of delirium detection (eg, interview of nursing staff, structured medical chart review) that provide additional opportunities for delirium detection. The positive and negative predictive values (PPV and NPV) of any test depend on the sensitivity and specificity of the test, as well as the prevalence of the disorder.[11] In the context of a delirium prevalence of approximately 33%, our findings of an estimated 72% sensitivity and 99% specificity (when trained assessors in several countries use the CAM) suggest that the CAM would have a PPV of about 98% and an NPV of about 88%. The high specificity coupled with the high PPV suggest that our delirium training methodology would be useful for explanatory research, where false positive diagnoses could be particularly problematic. For example, with a study investigating neuroimaging correlates of delirium, it would be important to be confident that positive diagnoses represent true cases of delirium. However, it is important to note that in a population with a low delirium prevalence (eg, <5%), the PPV of the CAM would probably be <80%, with resulting increased risk of false positive diagnoses.

Comparison with literature

Delirium assessment instruments can be subjective, and previous research has demonstrated that it is often difficult to diagnose in a clinical setting.[11 12] Currently, there is no gold standard or reliable biomarker (eg, MRI or electroencephalogram correlates) for delirium diagnosis. A trial performed in 2014 showed that emergency department nurses and physicians have suboptimal agreement when informally determining patients' delirium status, even after a teaching intervention. Nurses had a sensitivity of 0.27 before the intervention and 0.40 after; physicians' sensitivity was 0.45 before and 0.60 after.[13] However, even with the addition of standardised tests and extensive training and experience, delirium screening is potentially unreliable. For example, a recent study found considerable disagreement between two experts who scored identical video-recorded patient encounters for delirium using both the Delirium Rating Scale-Revised-98 (DRS-R-98) and CAM-ICU.[14] Our overall almost perfect

Table 4 Sensitivity and specificity of CAM instrument with 95% CIs

	Sensitivity	Specificity	PPV	NPV
Acute change	0.97 (0.93 to 0.99)	0.79 (0.65 to 0.89)	0.94 (0.90 to 0.96)	0.87 (0.75 to 0.94)
Fluctuation	0.88 (0.77 to 0.96)	0.77 (0.70 to 0.83)	0.53 (0.46 to 0.60)	0.96 (0.92 to 0.98)
Inattention	0.99 (0.94 to 1.00)	0.65 (0.57 to 0.74)	0.69 (0.57 to 0.69)	0.99 (0.93 to 1.00)
Disorganisation	0.68 (0.56 to 0.78)	0.99 (0.95 to 1.00)	0.96 (0.87 to 0.99)	0.86 (0.82 to 0.89)
Altered LOC	0.58 (0.37 to 0.77)	0.90 (0.85 to 0.94)	0.42 (0.30 to 0.55)	0.94 (0.92 to 0.96)
Overall diagnosis*	0.72 (0.60 to 0.81)	0.99 (0.96 to 1.00)	0.98 (0.89 to 1.00)	0.88 (0.83 to 0.91)

*Per reference standard: 33% (3/9) observed cases with delirium.
LOC, level of consciousness; NPV, negative predictive value; PPV, positive predictive value.

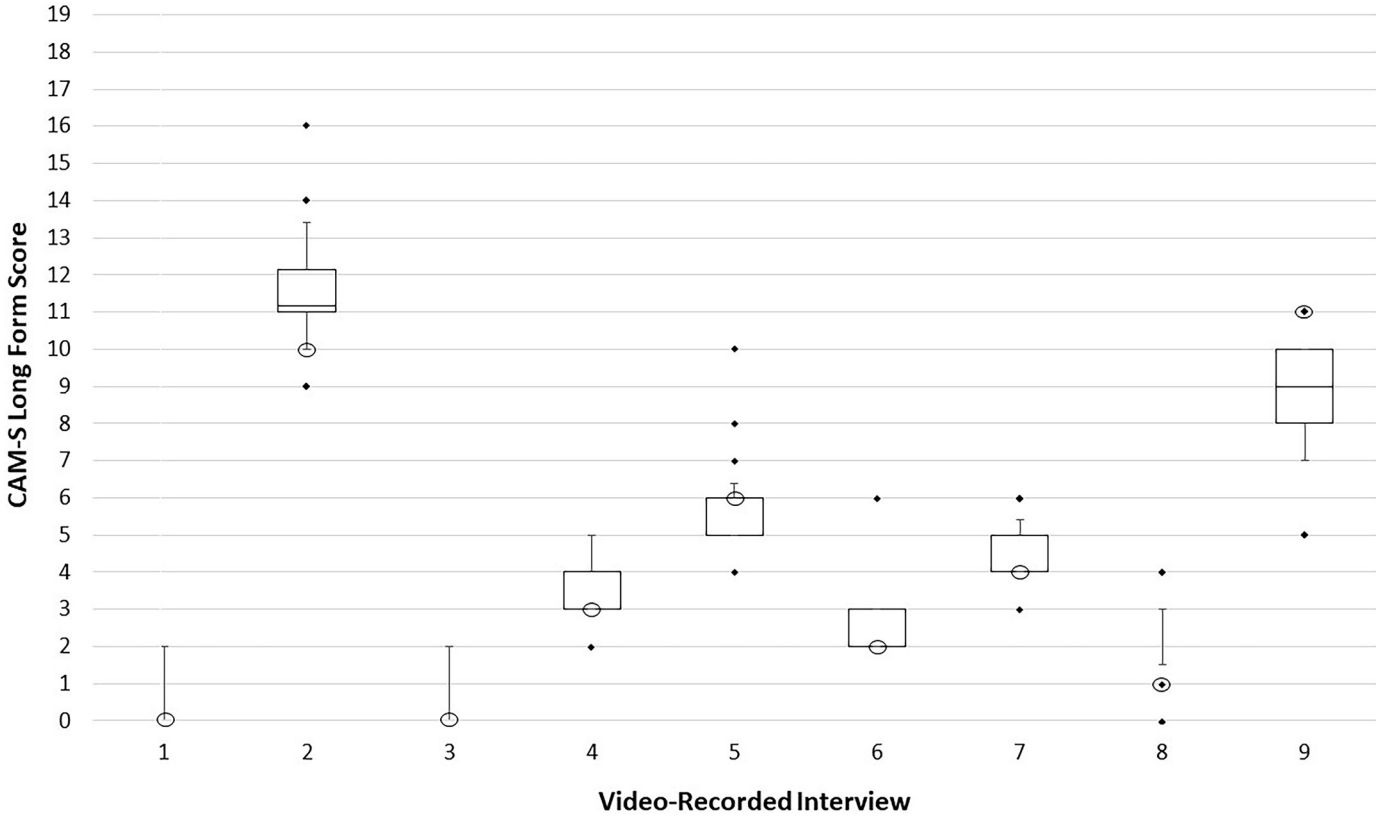

Figure 1 Confusion Assessment Method (CAM)-S severity scores for each video-recorded interview.

agreement using the CAM is encouraging in comparison with previous trials. This may be credited to the strength of the CAM algorithm and the rigorous training and continued education of individual assessors.

Although significant time to train researchers is required when using the CAM instrument, it appears to be worth the effort. A 2010 review evaluated 11 bedside delirium instruments based on sensitivity, specificity and likelihood ratios.[15] The CAM instrument had the second highest pooled likelihood ratios (positive, 9.6, 95% CI 5.8 to 16.0; negative, 0.16, 95% CI 0.09 to 0.29), while taking less time to administer than other high-performing delirium screening tools. In addition, the CAM allows for severity rating, often infeasible with brief screening. Previous literature indicates that the sensitivity (46%–100%) and specificity (63%–100%) of the CAM instrument is varied and largely influenced by the quality of training.[16] In spite of the additional challenge of assessor training, the CAM has been perceived as an optimal tool.

Strengths

One advantage of the approach taken in this study was that postsurgical older adults were video recorded in a real-world setting. Second, we included a multidisciplinary and international group of raters with varying backgrounds and levels of clinical experience, which reflects the composition of researchers in many multisite trials. Third, for the reference calibration, we used an unbiased expert rater who was not involved in recruitment or evaluation for the PODCAST trial. Thus, an important strength of this study was the finding that the raters did agree among themselves, as well as having excellent concordance with an impartial, expert and external rater. Finally, the determination of delirium severity as well as its binary appraisal was helpful, since all nine patients had some features of delirium.

Limitations

There were limitations of this study that should be considered. First, the video modality might hinder the ability of raters to interpret certain features of delirium, such as agitation or psychomotor slowing. Second, this analysis did not consider varying interviewing styles. Because each rater observed the same interview conducted by one person, it is possible that different conclusions would have been reached depending on whether the raters themselves were to interview the patient. For example, if a feature is unclear, we encourage raters to ask additional probing questions. This is subject to the judgement of the individual conducting the survey. Follow-up studies could compare separate interviews conducted by two different individuals. This presents a paradoxical issue for testing the CAM. Although sequential interviews by different raters would test the agreement of interviewing styles, delirium is a fluctuating disorder. Features that are present in one moment might not be observable in the next. Additionally, sequential interviews might be hindered by patient comfort level, as several questions conducted in close succession are often not appreciated. A third limitation is that our sample of patients interviewed was small, only three patients fulfilled CAM criteria for delirium,

and the full spectrum of delirium severity was likely not exemplified. When calculating severity scores using the CAM-S Long Form (scores range from 0 to 19),[5] we found that the true CAM-negative patients had an average median CAM-S score of 2.0, and true CAM-positive patients had an average median score of 8.4. It is possible that if a broader range of severity was included, overall agreement would have either declined or improved. A final limitation is that the sensitivity, specificity and PPV and NPV calculations were done under the assumption that an expert rater can provide a reliable 'reference standard' assessment of delirium. However, since no objective measure or biomarker of delirium exists, this or some other assumption is warranted. Also, as noted, PPV and NPV calculations are affected by the prevalence of a disorder in the population of interest.[12] This study does not address item selection, content validity or clinical relevance, which were beyond the scope of this work.

In conclusion, this substudy of the PODCAST trial found that with appropriate and structured training, a group of international researchers with diverse clinical experience and training can achieve good concordance and accuracy in delirium assessment using the CAM instrument. Importantly, this agreement appeared to pertain both to delirium diagnosis and to determination of delirium severity. We attribute this good agreement to a rigorous training protocol with regular quality assessments and discussions regarding patients who are deemed borderline on meeting thresholds within the CAM instrument. Overall, this encouraging finding suggests that the CAM can be a reliable tool for use in multicentre, international clinical trials focusing on delirium or delirium severity as the primary outcome.

Author affiliations
[1]Department of Anesthesiology, Washington University in Saint Louis School of Medicine, Saint Louis, Missouri, USA
[2]Department of Mathematics, Washington University in Saint Louis, St. Louis, Missouri, USA
[3]Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri, USA
[4]Aging Brain Center, Institute for Aging Research, Hebrew SeniorLife, Boston, Massachusetts, USA
[5]Department of Anesthesiology, University of Michigan, Ann Arbor, Michigan, USA
[6]Department of Medicine, Beth Israel Deaconess Medical Center, Hebrew Senior Life, Harvard Medical School, Boston, Massachusetts, USA

Acknowledgements Authors would also like to acknowledge the patient advisers for the support of this research.

Collaborators The PODCAST Research Group includes: Apakama GP, Aquino K, Arya VK, Avidan MS, Ben Abdallah A, Chen Y, Dicks R, Downey RJ, Emmert DA, Escallier K, Fardous HA, Fritz BA, Funk DJ, Galati J, Gipson KE, Girardi L, Graetz TJ, Grocott H, Gruber AT, Hicks M, Hudetz JA, Inouye SK, Ivascu NS, Jacobsohn E, Jayant A , Kashani HH, Kavosh MS, Kunkler BS, Lee YH, Lenze E, Mashour GA, Maybrier HR, McKinney AS, McKinnon SL, Mickle AM, Monterola M, Muench MR, Murphy MR, Noh GJ, Pagel PS, Pryor KO, Redko M, Richards T, Rogers EM, Schmitt E, Sivanesan L, Steinkamp ML, Tellor B, Thomas S, Torres BA, Upadhyayula R, Veselis RA, Vlisides PE, Waszynski C, Wildes TS, Veltri C, Yulico H.

Contributors HRM contributed by writing and editing the manuscript, managing the electronic database, coordinating delirium assessment training and conducting patient interviews. AMM contributed by editing the manuscript, managing the electronic database and editing patient interviews. KEE contributed by conceptualising study design. NL contributed by performing statistical analyses. EMS, KP and SKI contributed by advising delirium assessment training. EMS and SKI also contributed by editing the manuscript. KP also served as the expert rater. RTU contributed by editing the manuscript and conducting patient interviews. TSW and GAM contributed by conceptualisng the study design and editing the manuscript. MSA contributed by conceptualising the study design, composing and editing the manuscript and overseeing delirium assessment training.

Funding This study was funded by the National Institutes of Health (NIDUS Grant: NIA R24AG054259, and grant T32GM103730) and the NIH/NCI Cancer Center Support Grant (P30 CA008748).

Competing interests None declared.

Patient consent Not required.

Ethics approval Washington University Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data set used for this study can be made available upon request.

REFERENCES

1. United States Census Bureau. *The older population: 2010*, 2011.
2. Whitlock EL, Vannucci A, Avidan MS, *et alMinerva anestesiologica* 2011;77:448–56.
3. Kennedy M, Enander RA, Tadiri SP, *et al*. Delirium risk prediction, healthcare use and mortality of elderly adults in the Emergency Department. *J Am Geriatr Soc* 2014;62:462–9.
4. De J, Wand AP. Delirium screening: a systematic review of delirium screening tools in hospitalized patients. *Gerontologist* 2015;55:1079–99.
5. Inouye SK, Kosar CM, Tommet D, *et al*. The CAM-S: development and validation of a new scoring system for delirium severity in 2 cohorts. *Ann Intern Med* 2014;160:526–33.
6. Inouye SK, van Dyck CH, Alessi CA, *et al*. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. *Ann Intern Med* 1990;113:941–8.
7. Avidan MS, Maybrier HR, Abdallah AB, *et al*. Intraoperative ketamine for prevention of postoperative delirium or pain after major surgery in older adults: an international, multicentre, double-blind, randomised clinical trial. *The Lancet* 2017;390:267–75.
8. Avidan MS, Fritz BA, Maybrier HR, *et al*. The Prevention of Delirium and Complications Associated with Surgical Treatments (PODCAST) study: protocol for an international multicentre randomised controlled trial. *BMJ Open* 2014;4:e005651.
9. Harris PA, Taylor R, Thielke R, *et al*. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377–81.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
11. Selim AA, Ely EW, Wesley Ely E. Delirium the under-recognised syndrome: survey of healthcare professionals' awareness and practice in the intensive care units. *J Clin Nurs* 2017;26:813–24.
12. Trogrlić Z, Ista E, Ponssen HH, *et al*. Attitudes, knowledge and practices concerning delirium: a survey among intensive care unit professionals. *Nurs Crit Care* 2017;22:133–40.
13. Grossmann FF, Hasemann W, Graber A, *et al*. Screening, detection and management of delirium in the emergency department - a pilot study on the feasibility of a new algorithm for use in older emergency department patients: the modified Confusion Assessment Method for the Emergency Department (mCAM-ED). *Scand J Trauma Resusc Emerg Med* 2014;22:19.
14. Numan T, van den Boogaard M, Kamper AM, *et al*. Recognition of delirium in postoperative elderly patients: a multicenter study. *J Am Geriatr Soc* 2017;65:1932–8.
15. Wong CL, Holroyd-Leduc J, Simel DL, *et al*. Does this patient have delirium?: value of bedside instruments. *JAMA* 2010;304:779–86.
16. Grover S, Kate N. Assessment scales for delirium: a review. *World J Psychiatry* 2012;2:58–70.