


Delirium detection methodologies: Implications for outcome measurement in clinical trials in postoperative delirium

Esther S. Oh^{1,2,3,5}  | Paul B. Rosenberg² | Nae-Yuh Wang¹ | Frederick E. Sieber⁴ | Karin J. Neufeld^{2,5}

¹Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

²Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

³Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

⁴Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

⁵Johns Hopkins University School of Nursing, Baltimore, Maryland, USA

Correspondence

Esther S. Oh, The Johns Hopkins University, School of Medicine, Division of Geriatric Medicine and Gerontology, Asthma and Allergy Center, 5501 Hopkins Bayview Circle, Rm 1B.76, Baltimore, MD 21224, USA.
Email: eh9@jhmi.edu

Funding information

National Center for Advancing Translational Sciences; National Institute on Aging; National Institutes of Health

Abstract

Objective: Delirium is a common postoperative complication of hip fracture. Various methods exist to detect delirium as a reference standard. The goal of this study was to characterize the properties of the measures obtained in a randomized controlled trial, to document their relationship to the Diagnostic and Statistical Manual of Mental Disorders:Text Revision based diagnosis of postoperative delirium by a consensus panel, and to describe the method in detail to allow replication by others.

Methods: A secondary analysis of the randomized trial STRIDE (A Strategy to Reduce the Incidence of Postoperative Delirium in Elderly Patients) was conducted. Delirium assessments were performed in 200 consecutive hip fracture repair patients ≥ 65 years old. Assessors underwent extensive training in delirium assessment and the final delirium diagnosis was adjudicated by a consensus panel of three physicians with expertise in delirium assessment.

Results: A total of 680 consensus panel delirium diagnoses were completed. There were only 19 (2.8%, 19/678) evaluations where the delirium adjudication by the consensus panel differed from delirium findings by the Confusion Assessment Method (CAM). In 16 (84%, 16/19) of the cases, CAM was negative but the consensus panel diagnosed the patient as having delirium based on all of the available information including the CAM.

Conclusion: The consensus panel diagnosis was more sensitive compared to CAM alone, however the magnitude of the difference was not large. When assessors are well trained and delirium assessments are closely supervised throughout the study, CAM may be adequate for delirium diagnosis in a clinical trial. Future studies are needed to test this hypothesis.

KEYWORDS

clinical trials, confusion assessment method (CAM), consensus panel diagnosis, delirium, hip fracture

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. International Journal of Geriatric Psychiatry published by John Wiley & Sons Ltd.

Key points

- The consensus panel diagnosis was more sensitive compared to the Confusion Assessment Method (CAM) alone.
- The magnitude of the differences in the proportion of delirium cases detected by CAM alone compared to the consensus panel diagnosis was not large.
- Delirium assessors need to undergo rigorous standardized training in cognitive evaluation, clinical observation and data gathering from informants and care providers.

1 | INTRODUCTION

Delirium is an abrupt and fluctuating disturbance in attention and awareness, developing over a short period of time and representing a change from baseline.¹ It is a condition that takes a toll on a large number of individuals and has a significant economic impact. For example, in the United States alone, delirium affects more than 2.6 million adults age 65 and older each year, and costs over 164 billion in annual healthcare expenditures.² In addition to high health care costs, delirium is also associated with long term consequences including poor functional recovery³ and institutionalization.⁴

A condition associated with high incidence of delirium is hip fracture repair; incidence ranges from 13% to 56%.⁵ It is one of the more common immediate complications after surgery⁶ and is also associated with a multitude of longer term complications including prolonged rehabilitation, loss of independence, and high mortality.⁷

Although there has been an effort to develop physiological measures of delirium such as biomarkers for diagnostic purposes,² delirium still remains a clinical diagnosis, based upon examination of the patient, interviewing informants (e.g. family, health care providers), and reviewing the medical record. Clinician based diagnosis as the reference standard is still the most common method of validating delirium screening methods in both clinical and research settings.² Challenges to accurate and rigorous diagnosis of delirium are frequently inherent to the syndrome and include fluctuation of symptoms during the day,¹ delirium types that are more difficult to recognize such as hypoactive delirium which may be characterized by drowsiness and inactivity as well as a milder degree of symptoms, as in subsyndromal delirium.⁸ In addition, presence of other conditions that either mimic delirium or frequently exist as comorbid conditions such as depression or dementia may also result in misclassification of the condition.² All of these factors may result in under-detection of delirium, especially when there is lack of training in delirium screening and assessment.⁹

The employment of a panel of clinical experts to review the evidence and come to consensus in the diagnosis of delirium (consensus panel) has been used in research settings to overcome aforementioned factors in delirium underdetection. For example, in a recent study that examined previously published delirium diagnosis methodologies, 70% of the authors of these studies reported using consensus panels to review delirium diagnoses.¹⁰ However, there are few studies that describe the methods in significant detail including the assessment training process,^{9,11,12} and it still remains unclear

whether the consensus diagnostic process significantly improves the sensitivity and specificity of delirium detection. In addition, although consensus panel diagnosis is often used as a reference standard for validation of delirium screening tools, the properties of this methodology compared to commonly used delirium detection instruments also remain unclear. Therefore, the goal of this study was to characterize the properties of a DSM-IV based diagnosis of postoperative delirium by a consensus panel process in a large clinical trial, and to describe the method in detail so that the process can be replicated in future studies.

2 | METHODS

2.1 | Study population

Our study is a secondary analysis of a randomized trial STRIDE (A Strategy to Reduce the Incidence of Postoperative Delirium in Elderly Patients) which comprised of 200 consecutive hip fracture repair patients ≥ 65 years who met the enrollment criteria from 2011 to 2016 (ClinicalTrials.gov; NCT00590707). The STRIDE trial assessed delirium incidence in older patients undergoing hip fracture repair following administration of lighter versus heavier propofol sedation during surgery.¹³ The details of the inclusion and exclusion criteria were published previously.¹⁴ The study was approved by the Johns Hopkins Institutional Review Board, and all participants gave written informed consent.

2.2 | Delirium and cognitive assessment methodology

2.2.1 | Delirium assessment training

Three research assistants (RAs) were trained by a board-certified psychiatrist and director of the inpatient psychiatry consultation service with ≥ 20 years of clinical experience including delirium assessment (KN). The initial training included a 3 h didactic course on clinical features of delirium and dementia as well as 2.5 h of training videos, featuring actors with symptoms of delirium. The RAs practiced rating Delirium Rating Scale-Revised 98 (DRS-R-98)¹⁵ and Confusion Assessment Method (CAM)¹⁶ assessment forms with the training video, and with volunteer patients

who were not involved in the STRIDE study. The RAs independently completed the ratings with the volunteer patients and compared the agreement with the study psychiatrist who observed the assessment in person. The RAs were permitted to start assessing study patients in the STRIDE study after the following requirements were met: (1) a minimum of 10 patients (50% with delirium) had been evaluated and (2) once the assessments were in 100% agreement with the study psychiatrist for five consecutive patients. The average training duration for each RA ranged from 10 to 20 h total. During the study, we measured agreement among raters and kappa statistics were between 0.7 and 0.8, which is consistent with substantial agreement.

2.2.2 | Cognitive baseline assessment training

Each of the RA's completed the online training course for the Clinical Dementia Rating (CDR).¹⁷ They also observed the study geriatrician (EO) at an outpatient Memory Clinic to learn clinical features of dementia.

2.2.3 | Delirium and cognitive baseline data collection

1. Direct patient assessment.

The RAs first asked the patient open ended interview questions about (1) the patient's sleep, (2) the care that the patient was receiving in the hospital, (3) the presence of any odd or unusual experiences, beliefs or perceptions during the past day and (4) the level of pain on a 10-point rating scale. The trained RAs used the following tools to screen cognitive function prior to surgery and at each subsequent visit to assess for delirium including the Mini-Mental State Examination (MMSE),¹⁸ Abbreviated Digit Span Test,¹⁹ and test of attention including reciting backwards the months of the year from December to January (MOTYB),²⁰ or Days Of the Week Backwards (DOWB) from Sunday to Monday.²¹ Patients were given credit if they were able to state months of the year backwards starting at December and correctly state at least 7 months or more.²² All 7 days of the week must have been performed correctly in backwards order in order to get credit for DOWB. The RAs also completed the CAM.¹⁶ Patients were assessed for the two core features of CAM including acute change in mental status with fluctuating course and inattention (features 1 and 2). They were also assessed for two other features including disorganized thinking and altered level of consciousness (features 3 and 4). The diagnosis of delirium by CAM required the presence of features 1 and 2 and either 3 or 4.¹⁶ Each interview was conducted once a day at approximately the same time each day. All patient outcome measures were purchased or used with appropriate permission from copyright holders.

2. Informant interviews.

The RAs interviewed informants (family, friends) about the patient's function prior to surgery using the Short Form of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE)²³ at study entry during the baseline assessment prior to surgery.

Prior to surgery and daily up to 5 days after surgery, the RAs interviewed the patient's family, friends and medical staff taking care of the patient about patient's ability to pay attention, their thinking, whether they were too drowsy, or too restless, presence of confusion, disorientation, hallucinations, agitation and pulling at IV lines or trying to get out of bed, slowed movements and sleep-wake cycle. Due to the fact that family members were almost always involved in the consent process, family members were available for questions preoperatively. If there were no family member available during the hospitalization, nursing and other staff (physical and occupational therapists) were interviewed each day for evidence of change from baseline. Finally, all notes in the medical record from the previous 24 h were reviewed by the RA's for any symptoms suggestive of delirium. These descriptions were noted in the research notes.

3. RA narrative.

The RAs completed a brief written narrative paragraph each day, and an assessment was completed that included a description of the patient's appearance, behavior, attention, level of arousal, and the environment when the patient was being interviewed. The patient's performance on the cognitive screens was summarized and included any additional information of importance provided by the informants and medical record entries from the past 24 h. Postoperative day 1 also included information from the preceding 24 h. This written synthesis formed the basis for the presentation to the consensus panel members. The process intentionally incorporated all available sources of information whether from the medical record, direct observation, or interview of clinical staff.

2.2.4 | Consensus panel for delirium diagnosis

The consensus panel consisted of two psychiatrists (KN, PR) and a geriatrician (EO) with expertise in delirium assessment. All of the members of the consensus panel were masked to the treatment assignment. The members of the panel reviewed all of the information gathered by the RAs including narrative of the patient and informant interviews, results of the assessments including CAM and DRS-R-98, and information gathered from chart review for the preceding 24 h. Each member of the panel independently rated the patient according to the DSM-IV.²⁴ The criteria included (i) disturbance in consciousness with reduced ability to focus, sustain or shift attention; (ii) a change in cognition or the development of a perceptual disturbance i.e. not better accounted for by a pre-existing, established, or evolving dementia; (iii) the disturbance develops over

a short period of time and tends to fluctuate during the course of the day.

There were three possible ratings for the panelists including the following (i) no delirium, no criteria met; (ii) no delirium, but one or two criteria met; (iii) delirium, all criteria met. The ratings of the three panelists were compared afterwards, and the panel discussed each case until consensus was achieved. All of the consensus ratings were completed prospectively, and ratings were not changed after the information about the subsequent days were revealed.

2.3 | Analyses

Patient characteristics were described using mean and standard deviation (SD) for continuous variables, and with frequency and percentage (%) for categorical variables. Between group differences of these characteristics were compared using 2-sample *t*-test, analysis of variance *F*-test, or their nonparametric equivalent (e.g., Wilcoxon rank sum test and Kruskal–Wallis test) as appropriate, for continuous variables, and using chi-square test or Fisher's exact test for categorical variables. We conducted secondary data analysis of non-independently obtained measures including the consensus panel diagnosis, the CAM and the DRS-R-98. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for CAM test results and DRS-R-98 severity score (>15) classifications against post-operative delirium diagnosis made by the expert consensus panel. Sensitivity analysis was performed with DRS-R-98 severity score >10. Estimates were derived based on results from post-op day 1, and all observations from post-op day 1 through hospital discharge or up to post-op day 5 if length of stay was more than 5 days. Exact 95% confidence intervals (CIs) were derived for the estimates, except for the estimates based on multiple day observations from the same patients, where the 95% CIs were derived based on generalized estimating equations (GEE) results using robust inferences. A *p*-value of ≤ 0.05 was considered to be statistically significant, and SAS 9.4 software was used for data analyses.

3 | RESULTS

A total of 538 hip fracture patients were screened from 2011 to 2016, and 200 patients were randomized to lighter or heavier propofol sedation. Of the 538 patients, 42.8% (225/538) met eligibility. Among those that met eligibility, 4.4% (10/538) declined participation and 6.6% (15/538) became ineligible during the time between consent and surgery. A total of 680 consensus panel delirium assessments were completed for the 200 patients from post-surgery day one to day five or hospital discharge, including 677 CAM and 674 DRS-R-98 assessments that were reviewed by the consensus panel. Among all the 680 patient-days with consensus panel evaluations, there were three patient-days without the CAM evaluation, and additional three (six in total) patient-days without the DRS-98 evaluation. On average, each patient had 3.39 (SD 1.18, range 1–5)

CAM and 3.38 (SD 1.18, range 1–5) DRS-R-98 assessments. The number of assessments were influenced by each patient's hospital length of stay. Baseline patient characteristics by delirium consensus panel diagnosis are listed in Table 1. In addition, most patients were retired (78.0%, 156/200), living in their own home (72.5%; 145/200) either by themselves or with their spouses (66%, 132/200). A total of 36.5% (73/200) of the patients were diagnosed with delirium on at least 1 day during hospitalization by consensus panel diagnosis. In comparison, 30% (60/200) of the patients developed delirium on at least 1 day during their hospital stay by CAM alone and 13.5% (27/200) by DRS-R-98 Severity Scale score of >15.^{12,15} Patients who met the delirium criteria in one of the three methods had significantly lower MMSE scores,¹⁸ lower abbreviated digit span (backwards) scores,²⁵ higher (worse) IQCODE,²³ and higher (worse) global CDR as well as CDR sum of boxes scores compared to those without delirium (Table 2). There were no statistically significant differences between abbreviated digit span (forward) scores across delirium diagnosis methods.

A total of 19 (2.8%, 19/678) delirium evaluations differed in the adjudication by the consensus panel diagnosis from the delirium findings by the CAM. In most of the evaluations (84%, 16/19), CAM was negative, but the consensus panel diagnosis adjudicated the patient as having delirium based on all of the available information including the CAM. In three of the evaluations (16%, 3/19), CAM was positive, but the consensus panel did not diagnose delirium. However, in all three cases, the consensus panel diagnosis adjudicated the case as "subsyndromal delirium" as it did not reach the DSM-IV criteria.

When sensitivity and specificity of delirium screening tools were determined using consensus panel diagnosis as the reference standard, sensitivity and specificity for CAM was 79.3% and 98.0%, respectively. For DRS-R-98 with cut off score of >15, the sensitivity and specificity were 33.1% and 100%, respectively (Table 3). When the cut off score of DRS-R-98 was decreased to >10, sensitivity was 73.6% and specificity 95.2% (Table S1). The PPV and NPV of the CAM was 93.3% and 94.6% respectively. By comparison, the PPV and NPV of the DRS-R-98 was 94.7% and 85.7% respectively. These estimates were derived from pulling all daily measures during participants' hospital stay up to postop day five after adjusting for correlation of repeated assessments within the participant.

4 | DISCUSSION

The goal of this study was to describe and evaluate the properties of DSM-IV based diagnosis of postoperative delirium in a hip fracture population by a consensus panel. The consensus panel detected more cases of postoperative delirium compared to CAM and DRS-R-98 alone. However, the differences in the number of delirium cases detected by the consensus panel compared to CAM were not large (~3%) when CAM assessments were administered by well-trained RAs and based upon rigorous standardized cognitive evaluation, clinical observation and data gathering from informants and care providers.

Our study results are similar to the study by Zou et al. which compared delirium diagnosis based on one-time assessment by a clinician and diagnosis based on multiple-time point assessment by a trained nurse clinician using CAM with reference rating by a consensus panel.¹¹ In this study, delirium diagnosis based on one-time assessment by a clinician had a low sensitivity and poor

level of agreement (e.g. kappa coefficient) compared to reference rating by the consensus panel. However, delirium diagnosis by a trained nurse clinician who assessed for delirium at multiple time points using CAM had high sensitivity and specificity as well as high level of agreement with reference rating by the consensus panel.¹¹

TABLE 1 Patient characteristics by delirium consensus panel diagnosis

	All patients (n = 200)	Delirium (n = 73)	No delirium (n = 127)	p-value
Demographics				
Age, mean (SD) ^a years	81.8 (7.7)	83.7 (7.2)	80.6 (7.8)	<0.01
Female	146 (73.0)	52 (71.2)	94 (74.0)	0.67
Caucasian	194 (97.0)	72 (98.6)	122 (96.1)	0.42
Education level				0.45
Less than high school	76 (38.0)	32 (43.8)	44 (34.6)	
High school	76 (38.0)	25 (34.2)	51 (40.2)	
Some college	28 (14.0)	11 (15.1)	17 (13.4)	
College graduate or higher	20 (10.0)	5 (6.8)	15 (11.8)	
Status prior to surgery				
Charlson comorbidity index, mean (SD)	1.5 (1.8)	1.8 (2.1)	1.3 (1.5)	0.15
ASA ^b physical status classification score, >3	11 (5.5)	6 (8.2)	5 (3.9)	0.21
Activities of Daily Living, ^c mean (SD)	4.61 (1.52)	4.08 (1.66)	4.91 (1.35)	<0.0001
Instrumental Activities of Daily living, mean (SD) ^d	5.85 (2.22)	5.19 (2.38)	6.23 (2.04)	<0.01
pre-surgery cognitive testing				
IQCODE, ^e mean (SD)	51.87 (7.20)	54.13 (8.46)	50.60 (6.06)	<0.05
CDR global score				<0.01
0	82 (41.4)	21 (29.6)	61 (48.0)	
0.5	94 (47.5)	37 (52.1)	57 (44.9)	
1	16 (8.1)	7 (9.9)	9 (7.1)	
2	6 (3.0)	6 (8.5)	0 (0.0)	
Mean (SD)	0.38 (0.42)	0.53 (0.54)	0.30 (0.31)	<0.01
CDR sum of boxes, mean (SD)	1.47 (2.53)	2.37 (3.52)	0.97 (1.55)	<0.01
Abbreviated digit span (total score) ^f	3.77 (1.01)	3.57 (1.10)	3.88 (0.91)	<0.054
Forward, mean (SD)	2.64 (0.56)	2.60 (0.55)	2.67 (0.56)	0.25
Backward, mean (SD)	1.12 (0.71)	0.97 (0.75)	1.20 (0.67)	<0.05
Mini mental status exam score ^g mean (SD)	24.30 (3.68)	22.99 (3.89)	25.05 (3.34)	<0.001
Surgery characteristics				
Type of hip surgery repair				<0.01
Hemiarthroplasty	69 (34.5)	20 (27.4)	49 (38.6)	
Total hip arthroplasty	11 (5.5)	1 (1.4)	10 (7.9)	
Cannulated screw	9 (4.5)	1 (1.4)	8 (6.3)	
Intramedullary hip screw	110 (55.0)	51 (69.9)	59 (46.5)	
Girdlestone	1 (0.5)	0 (0.0)	1 (0.8)	

(Continues)

TABLE 1 (Continued)

	All patients (n = 200)	Delirium (n = 73)	No delirium (n = 127)	p-value
Surgery duration, mean (SD) minutes	129.6 (37.4)	126.3 (39.0)	131.5 (36.4)	0.21
Incision to end of surgery, mean (SD) minutes	91.5 (34.4)	86.3 (35.3)	94.6 (33.6)	<0.05

Abbreviation: IQCODE, Informant Questionnaire on Cognitive Decline in the Elderly.

^aSD: standard deviation.

^bASA: American Society of Anesthesiologists.

^cADL: Activities of Daily Living is based on the Physical Self-Maintenance Scale (PSMS), which is scored from a minimum of 0 to a maximum of 6 which indicates full independence.²⁶

^dIADL: Instrumental Activities of Daily Living is scored from 0 to 8 with 8 indicating full independence.²⁶

^eIQCODE: Short Form of the Informant Questionnaire on Cognitive Decline in the Elderly Informant Questionnaire.²³

^fRaw test scores for verbal fluency and digit span were transformed to T scores based upon population norms standardized for age, sex, education, and race with mean = 50 and SD = 10.

^gMMSE: Mini-mental state examination scores range from 0 to 30 with 30 indicating good cognitive function; p-values are calculated from Fisher's Exact test, t-test or a Wilcoxon rank-sum test.

TABLE 2 Baseline cognition by measures of delirium

	Confusion assessment method (CAM)		Delirium rating Scale-98R severity scale		Consensus panel DSM- IV TR delirium diagnosis	
	Patients with any CAM + in postop period	Patients with no CAM+ in postop period	Patients with any score >15 in postop period	Patients with all scores ≤15 in the postop period	Patients with DSM IV TR delirium diagnosis in postop period	Patients without any DSM IV TR delirium diagnosis in postop period
N (%)	60 (30)	140 (70)	27 (13.5)	173 (86.5)	73 (36.5)	127 (63.5)
MMSE						
Mean (SD)	22.5 (4.01)	25.1 (3.23)**	20.6 (4.28)	24.9 (3.2)**	23.0 (3.89)	25.0 (3.34)**
Abbreviated digit span forward (out of 3)						
Mean (SD)	2.6 (0.53)	2.7 (0.57)	2.7 (0.56)	2.6 (0.56)	2.6 (0.55)	2.7 (0.56)
Backward (out of 2)						
Mean (SD)	0.9 (0.75)	1.2 (0.68)*	1.0 (0.77)	1.1 (0.70)	1.0 (0.75)	1.2 (0.67)*
Total (backwards and forwards)						
Mean (SD)	3.6 (1.07)	3.9 (0.95)	3.6 (1.17)	3.8 (0.97)	3.6 (1.10)	3.9 (0.91)*
IQCODE						
Mean (SD)	55.0 (9.01)	50.6 (5.85)**	58.7 (10.46)	50.8 (5.95)**	54.1 (8.46)	50.6 (6.06)*
Clinical dementia rating						
Global score mean (SD)	0.6 (0.57)	0.3 (0.31)**	0.8 (0.60)	0.3 (0.35)**	0.5 (0.54)	0.3 (0.31)*
Sum of boxes mean (SD)	2.7 (3.77)	1.0 (1.53)*	4.1 (4.42)	1.1 (1.83)*	2.4 (3.52)	1.0 (1.55)*

Note: Data were pulled for all 200 participants for each assessment (CAM, DRS-R-98, DSM-IV TR) which were conducted daily during participants' hospital stay. p values are calculated from t-test or a Wilcoxon rank-sum test.

Abbreviations: DSM- IV TR, Diagnostic and Statistical Manual of Mental Disorders: Text Revision; IQCODE, Informant Questionnaire on Cognitive Decline in the Elderly; MMSE, Mini-Mental State Examination.

*p < 0.05.

**p < 0.001.

In general, increased sensitivity of delirium detection by the consensus panel is most likely due to the fact that more information is incorporated into delirium adjudication including objective cognitive and delirium testing, informant (family, nursing staff)

interviews, chart review and a narrative account of the encounter with the patient (appearance, behavior, environment, etc.). In addition, the consensus panel consisted of three physicians (two psychiatrists and one geriatrician) who had extensive experience in

TABLE 3 Diagnostic characteristics of screening tools on the initial day and on multiple days of observations compared to the consensus panel reference standard diagnosis

	Confusion assessment method (CAM)			Delirium rating Scale-98R severity scale (DRS-R-98) ^a		
	First day of observation (n = 200 patients) % (95% CI)	Multiple days of observation crude (n = 678 patient-days) % (95% CI)	Multiple days of observation adjusted (n = 678 patient-days) ^b % (95% CI)	First day of observation (n = 200 patients) % (95% CI)	Multiple days of observation crude (n = 675 patient-days) % (95% CI)	Multiple days of observation adjusted (n = 675 patient-days) ^b % (95% CI)
Delirium (consensus panel) ^c	26.5 (20.4, 32.6)	19.3 (16.5, 22.5)	19.1 (15.5, 23.6)	26.5 (20.4, 32.6)	19.3 (16.5, 22.5)	19.1 (15.5, 23.6)
Prevalence of delirium per neuropsychiatric examination, %	22.5 (16.7, 28.3)	15.9 (13.4, 18.9)	15.5 (12.2, 19.7)	8.5 (4.6, 12.4)	6.4 (4.8, 8.5)	5.9 (4.0, 8.7)
Sensitivity	79.3 (68.3, 90.2)	77.9 (71.1, 85.3)	76.6 (72.8, 80.4)	33.1 (25.9–42.2)	29.9 (22.0, 40.7)	31.9 (28.1, 35.7)
Specificity	98.0 (95.7, 100.0)	98.9 (98.0, 99.8)	98.9 (98.0, 99.8)	100 (100–100)	100 (91.6–100)	100 (98.0, 100)
Positive predictive value (PPV)	93.3 (86.1, 100.0)	94.4 (90.1, 98.8)	93.3 (88.8, 97.8)	100 (100–100)	100 (100–100)	94.7 (84.3, 100)
Negative predictive value (NPV)	92.9 (88.9, 97.0)	94.9 (93.1, 96.7)	94.7 (92.6, 96.8)	80.3 (74.6, 86.1)	86.2 (83.6, 88.9)	85.7 (82.3, 89.2)

Note: Among cases where both screening tests and neuropsychiatric exam were completed on the same day.

^aDRS-98R Severity Scale >15 (i.e. 16 and up).

^bAdjustment was done for the repeated assessments of individual patients using generalized estimating equations models with robust estimates.

^cA total of 73 (36.5%) of 200 unique patients were diagnosed with delirium, per neuropsychiatric evaluation, on at least 1 day during their hospital stay.

delirium diagnosis. In our study, the RAs underwent rigorous delirium assessment training, and were trained to incorporate objective assessment results (e.g. MMSE, digit span) as well as informant and chart review based clinical history in scoring the daily CAM scores. In such instance, we found that the rate of delirium diagnosis by CAM alone performed by well trained RAs was very close to the reference standard by the consensus panel. Our RAs also completed the online training course for the CDR¹⁷ and observed a study physician at an outpatient memory clinic for the purpose of rating CDR as part of the study. However, these types of training are probably beyond the scope of a focused delirium training. Although some of the elements of this study are implementable at a systems level, extensive effort that was given in this study for getting data on patients on each day and longitudinal follow-up takes a tremendous amount of resource and may be difficult to implement as part of a routine care.

In our study, DRS-R-98 was used to measure delirium severity, which is a widely used tool that can be used to diagnose both delirium and delirium severity.^{15,27} Using the consensus panel diagnosis as the reference, DRS-R-98 alone had high specificity for diagnosing delirium, but had low sensitivity when cut-off score of 15.25 was used. DRS-R-98 therefore appears to be more useful for documenting delirium severity than for diagnosing delirium. However, measuring delirium severity remains important. We previously demonstrated that delirium severity measured by DRS-R-98 was found to be an independent risk factor for 1 year mortality in the STRIDE study,²⁸ similar to other studies that have reported the association of delirium severity with important long-term clinical outcomes.^{29,30}

It is also important to recognize that in addition to the type of delirium detection tools, other factors that may influence delirium detection rate, including the frequency of assessments⁵ as well as different diagnostic criteria that have been revised over time in the Diagnostic and Statistical Manual (DSM).¹² One study comparing DSM-5 and DSM-IV in pooled data from six prospective studies showed that delirium identification rate varied significantly depending on the interpretation of different DSM criteria.³¹

The strength of this study includes a detailed description of how to operationalize training of delirium assessors and a consensus panel, and also provides an estimate of how delirium assessments by well-trained RAs can perform compared to a reference standard. We demonstrate that although the consensus panel detected more cases of delirium, delirium by CAM alone was also very close to the reference standard. However, it is important to point out that CAM was rated by well-trained RAs, and that the same information gathered to rate the CAM was also reviewed during the consensus diagnosis process, and therefore we would expect a high sensitivity/specificity as well as good PPV and NPV. The sample size for this secondary data analysis was constrained to the total sample size of 200 from the original STRIDE trial, which limited the level of estimation precision as demonstrated by the width of the 95% CI for the estimates, especially among the sensitivity estimates which are further constrained by the delirium event rate.

The information presented in this paper is a secondary data analysis of non-independently obtained measures including the consensus panel diagnosis, the CAM and the DRS-R-98. True prediction of the psychometric properties of the screening tools versus that of the consensus panel should be derived from independently

derived measures. While it might be tempting to conclude that the CAM screening tool alone would have performed as well as the consensus panel diagnostic adjudications, we are unable to conclude this. For example, the lead RA (who was responsible for supervision) presented to the consensus panel during each meeting. After she was questioned about the case, she listened to and documented the various arguments by panel members in coming to a decision about diagnosis. While this served as a quality check on the reported data, it may also have shaped the RA's administration during future assessments in the study, although we did not find evidence that the duration elapsed in the study affected the discrepancies between the CAM assessments and the consensus panel evaluations. Thus, the rating processes in this study are not entirely independent; consensus panel discussions may have also served an unplanned training function throughout the study.

5 | CONCLUSION

The consensus panel diagnosis detected more cases of delirium compared to CAM alone, however the magnitude of the difference was not large. When assessors are well trained and receive ongoing supervision, CAM may be adequate for delirium diagnosis in a clinical trial.

ACKNOWLEDGMENT

We would like to acknowledge the support by Johns Hopkins Institute for Clinical and Translational Research (ICTR) which is funded in part by Grant Number UL1 TR 001079 from the National Center for Advancing Translational Sciences (NCATS) a component of the National Institutes of Health (NIH) (EO, NYW), K23AG043504 (NIA/NIH) (EO), R01 AG057725 (NIA/NIH) (EO), R01 AG033615 (NIA/NIH) (FS, KN, NYW), and the Roberts Gift Fund (EO). The funding sources had no role in the design, conduct, or reporting of this study.

CONFLICT OF INTEREST

Authors do not have any conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

All de-identified data will be available to other investigators upon reasonable request.

ORCID

Esther S. Oh  <https://orcid.org/0000-0002-9638-5763>

REFERENCES

- American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. 2013. <https://doi.org/10.1176/appi.books.9780890425596>
- Oh ES, Fong TG, Hsieh TT, Inouye SK. Delirium in older persons: advances in diagnosis and treatment. *JAMA*. 2017;318:1161-1174.
- Marcantonio ER, Flacker JM, Michaels M, Resnick NM. Delirium is independently associated with poor functional recovery after hip fracture. *J Am Geriatr Soc*. 2000;48:618-624.
- Krogseth M, Wyller TB, Engedal K, Juliebø V. Delirium is a risk factor for institutionalization and functional decline in older hip fracture patients. *J Psychosom Res*. 2014;76:68-74.
- Oh ES, Li M, Fafowora TM, et al. Preoperative risk factors for postoperative delirium following hip fracture repair: a systematic review. *Int J Geriatr Psychiatry*. 2015;30:900-910.
- Oh ES, Sieber FE, Leoutsakos J, Inouye SK, Lee HB. Sex differences in hip fracture surgery: preoperative risk factors for delirium and postoperative outcomes. *J Am Geriatr Soc*. 2016;64:1616-1621.
- Bentler SE, Liu L, Obrizan M, et al. The aftermath of hip fracture: discharge placement, functional status change, and mortality. *Am J Epidemiol*. 2009;170:1290-1299.
- Hosker C, Ward D. Hypoactive delirium. *BMJ*. 2017;357:j2047.
- Numan T, van den Boogaard M, Kamper AM, Rood PJ, Peelen LM, Slooter AJ, Dutch Delirium Detection Study Group. Recognition of delirium in postoperative elderly patients: a multicenter study. *J Am Geriatr Soc*. 2017;65:1932-1938.
- Neufeld KJ, Nelliot A, Inouye SK, et al. Delirium diagnosis methodology used in research: a survey-based study. *Am J Geriatr Psychiatry*. 2014;22:1513-1521.
- Zou Y, Cole MG, Primeau FJ, McCusker J, Bellavance F, Laplante J. Detection and diagnosis of delirium in the elderly: psychiatrist diagnosis, confusion assessment method, or consensus diagnosis? *Int Psychogeriatr*. 1998;10:303-308.
- Kuhn E, Du X, McGrath K, et al. Validation of a consensus method for identifying delirium from hospital records. *PLoS One*. 2014;9:e111823.
- Sieber FE, Neufeld KJ, Gottschalk A, et al. Effect of depth of sedation in older patients undergoing hip fracture repair on postoperative delirium: the STRIDE randomized clinical trial. *JAMA Surg*. 2018;153:987-995.
- Li T, Wieland LS, Oh E, et al. Design considerations of a randomized controlled trial of sedation level during hip fracture repair surgery: a strategy to reduce the incidence of postoperative delirium in elderly patients. *Clin Trials*. 2017;14:299-307.
- Trzepacz PT, Mittal D, Torres R, Canary K, Norton J, Jimerson N. Validation of the Delirium Rating Scale-revised-98: comparison with the delirium rating scale and the cognitive test for delirium. *J Neuropsychiatry Clin Neurosci*. 2001;13:229-242.
- Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegel AP, Horwitz RI. Clarifying confusion: the confusion assessment method: a new method for detection of delirium. *Ann Intern Med*. 1990;113:941-948.
- Morris J. Current vision and scoring rules the clinical dementia rating (CDR). *Neurol*. 1993;43:2412-2414.
- Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975;12:189-198.
- Strub R, Black F. *The Mental Status Examination in Neurology*. F.A. Davis Company: Philadelphia; 2000.
- O'Regan NA, Ryan DJ, Boland E, et al. Attention! A good bedside test for delirium? *J Neurol Neurosurg Psychiatr*. 2014;85:1122-1131. <https://doi.org/10.1136/jnnp-2013-307053>
- Marra A, Jackson JC, Ely EW, et al. Focusing on inattention: the diagnostic accuracy of brief measures of inattention for detecting delirium. *J Hosp Med*. 2018;13:551-557. <https://doi.org/10.12788/jhm.2943>
- Bellelli G, Morandi A, Davis DH, et al. Validation of the 4AT, a new instrument for rapid delirium screening: a study in 234 hospitalised older people. *Age Ageing*. 2014;43:496-502.
- Jorm A. A short form of the informant Questionnaire on cognitive decline in the elderly (IQCODE): development and cross-validation. *Psychol Med*. 1994;24:145-153.
- American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. 2000.

25. Wechsler D. *WMS-R: Wechsler Memory Scale-Revised Psychological Corporation*; 1987.
26. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. *Gerontologist*. 1969;9: 179-186.
27. Jones RN, Cizginer S, Pavlech L, et al. Assessment of instruments for measurement of delirium severity: a systematic review. *JAMA Intern Med*. 2019;179:231-239.
28. Sieber F, Neufeld KJ, Gottschalk A, et al. Depth of sedation as an interventional target to reduce postoperative delirium: mortality and functional outcomes of the Strategy to Reduce the Incidence of Postoperative Delirium in Elderly Patients randomised clinical trial. *Br J Anaesth*. 2019;122:480-489.
29. Vasunilashorn SM, Marcantonio ER, Gou Y, et al. Quantifying the severity of a delirium episode throughout hospitalization: the combined importance of intensity and duration. *J General Intern Med*. 2016;31:1164-1171.
30. Rosgen BK, Krewulak KD, Stelfox HT, Ely EW, Davidson JE, Fiest KM. The association of delirium severity with patient and health system outcomes in hospitalised patients: a systematic review. *Age Ageing*. 2020;49:549-557.
31. Meagher DJ, Morandi A, Inouye SK, et al. Concordance between DSM-IV and DSM-5 criteria for delirium diagnosis in a pooled database of 768 prospectively evaluated patients using the delirium rating scale-revised-98. *BMC Med*. 2014;12:1-10.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Oh ES, Rosenberg PB, Wang N-Y, Sieber FE, Neufeld KJ. Delirium detection methodologies: implications for outcome measurement in clinical trials in postoperative delirium. *Int J Geriatr Psychiatry*. 2022;37(3): 1-9. <https://doi.org/10.1002/gps.5695>