NIDUS II Measurement Core:

Optimizing Delirium Assessment in Research Proposals

Richard N. Jones, Sc.D. Brown University rich_jones@brown.edu

NIDUS Mentoring Session

October 12, 2022

Zoom

Overview

I. **What** reviewers are looking for in research proposals, with regard to choice in measurement instruments

II. **How** NIDUS resources can help giving reviewers what they want and strengthen the design of your research

III. **Strategies** for choosing an instrument for your proposed research

What reviewers are looking for

Part I

What reviewers want

- **Excellent science**
  - Strong designs that answer well-formed questions (approach, rigor)
  - Questions & answers that advance the field (significance, innovation)
  - Research designs that are ethical and feasible (approach, env., inv.)
  - Clarity and efficiency in presentation

How does this relate to delirium assessment?

- **Delirium assessments should**
  - Match with the goals of measurement
  - Match with the population being assessed
  - Match with the assessor
  - Have some validity evidence for research context

Match with goals of measurement

- Delirium case identification
- Delirium severity
  - An episode of delirium, or severity of delirium during a stay?
  - Symptom severity (peak of symptom count/sum; sum over all days)
  - Duration of delirium during stay

Match with the population being assessed

- Type of patient
  - Capacity to participate in assessment

Match with the assessor

- Physician?
- Nurse?
- Other caregiver?
- Family?
- Lay interviewer?

Validity evidence for research context

- Has the instrument been used in patients similar to the planned population previously?
- Is there any validity evidence for the use of the chosen instrument in the planned research context?

Please remember reliability and validity **statistics** are sample-dependent and context-dependent results and do not describe immutable properties of a test.

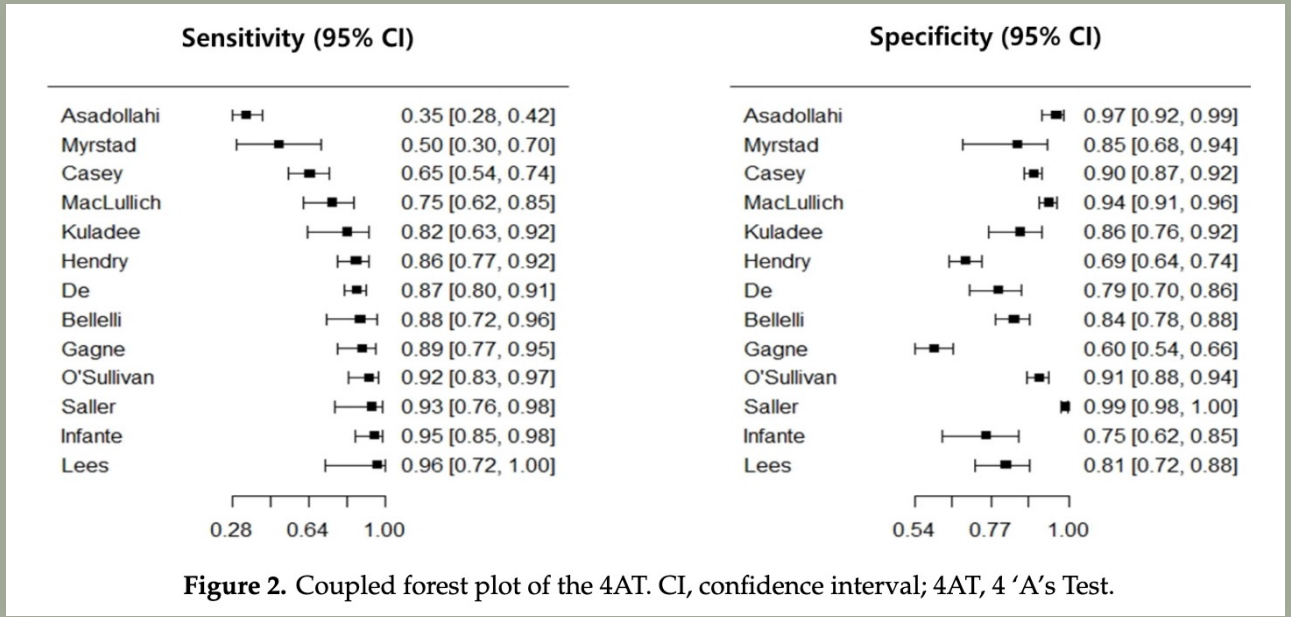
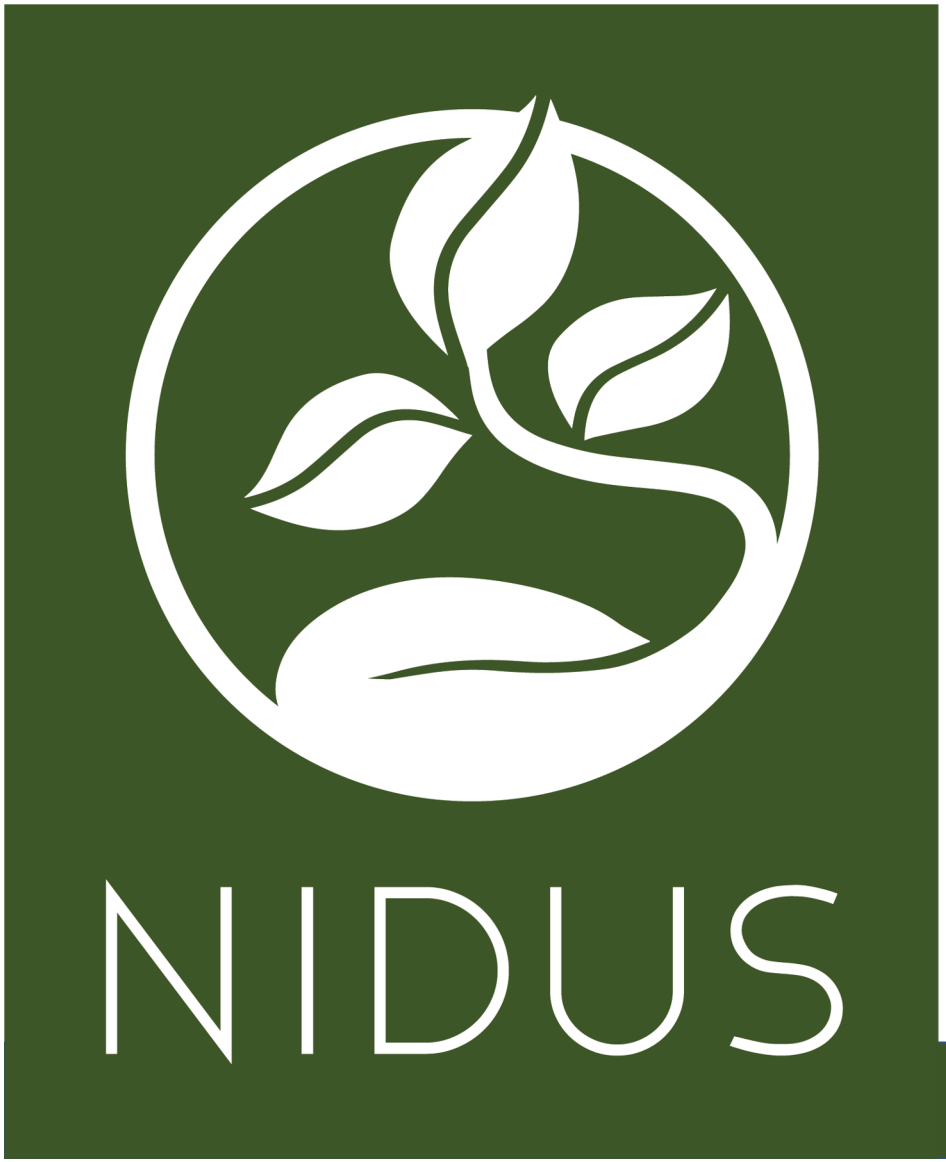


Figure 2. Coupled forest plot of the 4AT. CI, confidence interval; 4AT, 4 'A's Test.

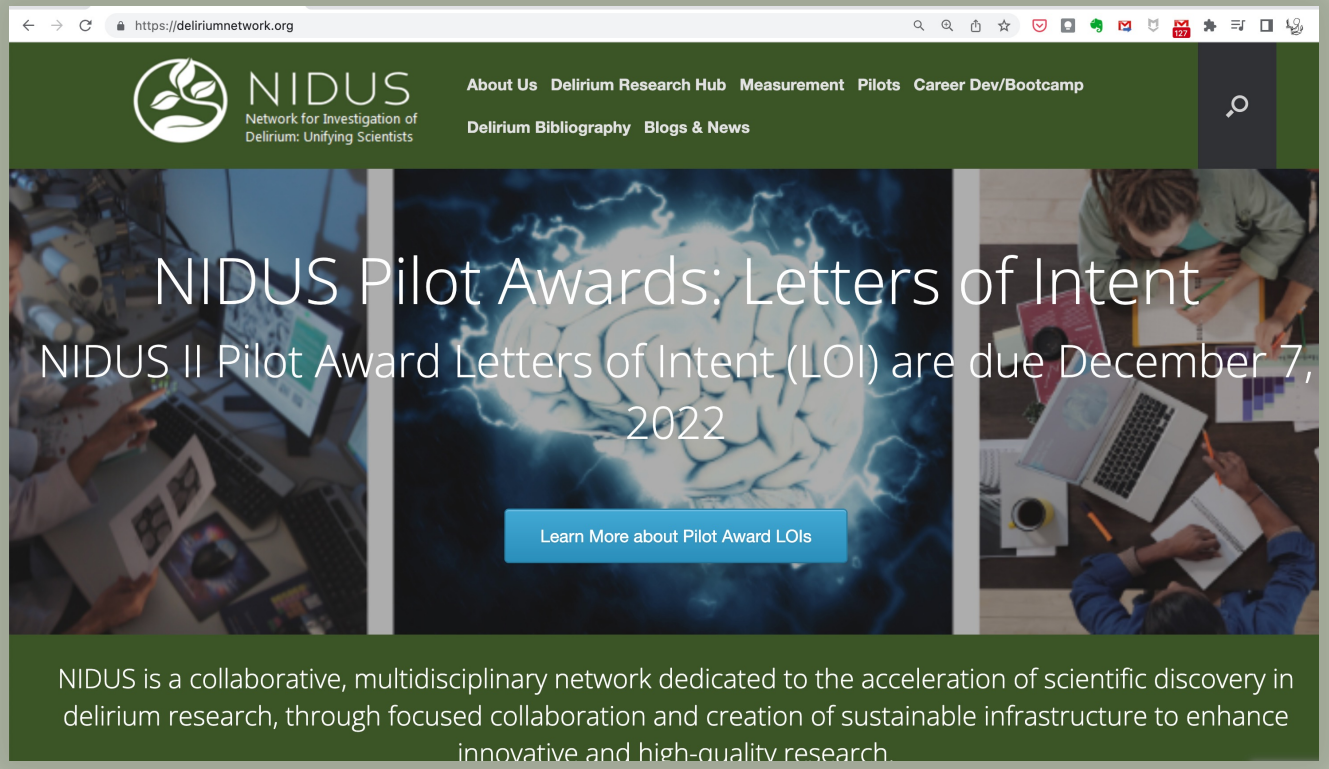
Jeong, E., Park, J., & Lee, J.. (2020). Diagnostic Test Accuracy of the 4AT for Delirium Detection: A Systematic Review and Meta-Analysis. *International Journal of Environmental Research and Public Health*, *17*(20), 7515. https://doi.org/10.3390/ijerph17207515

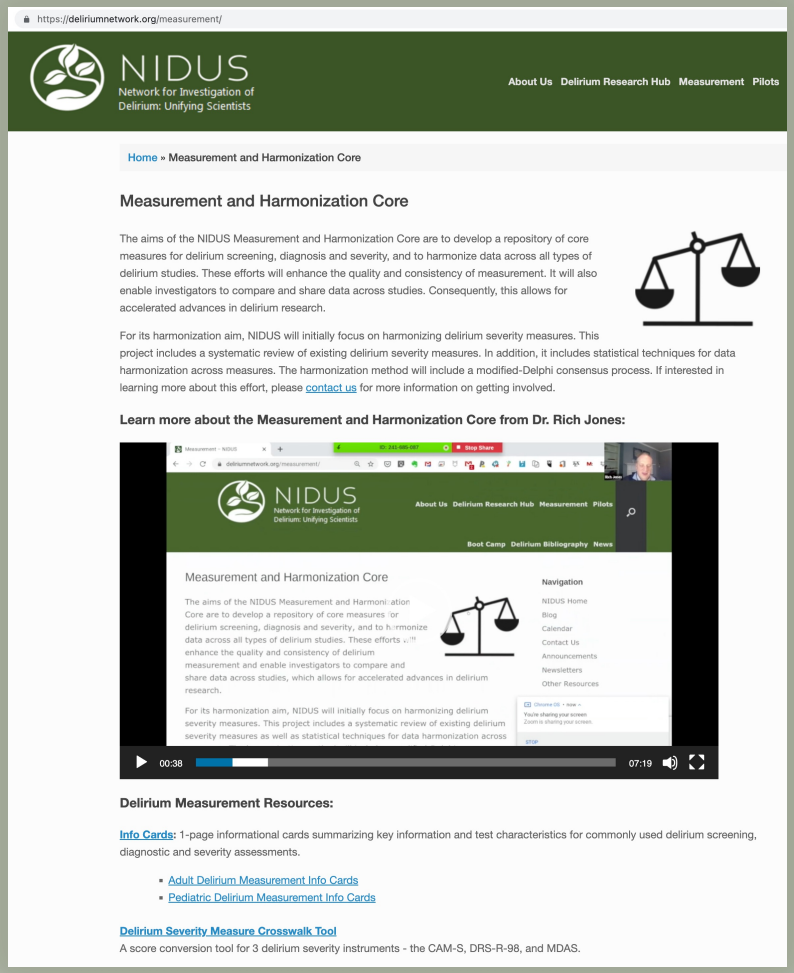
NIDUS resources that might be helpful

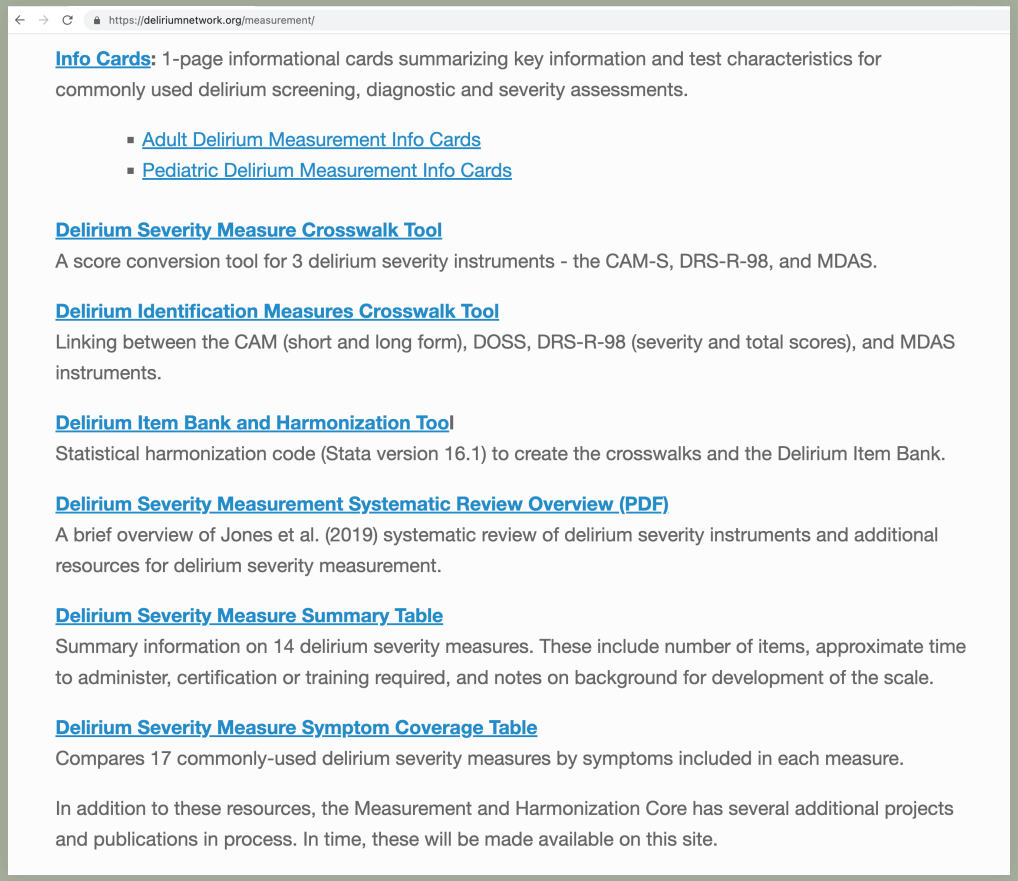
Part II

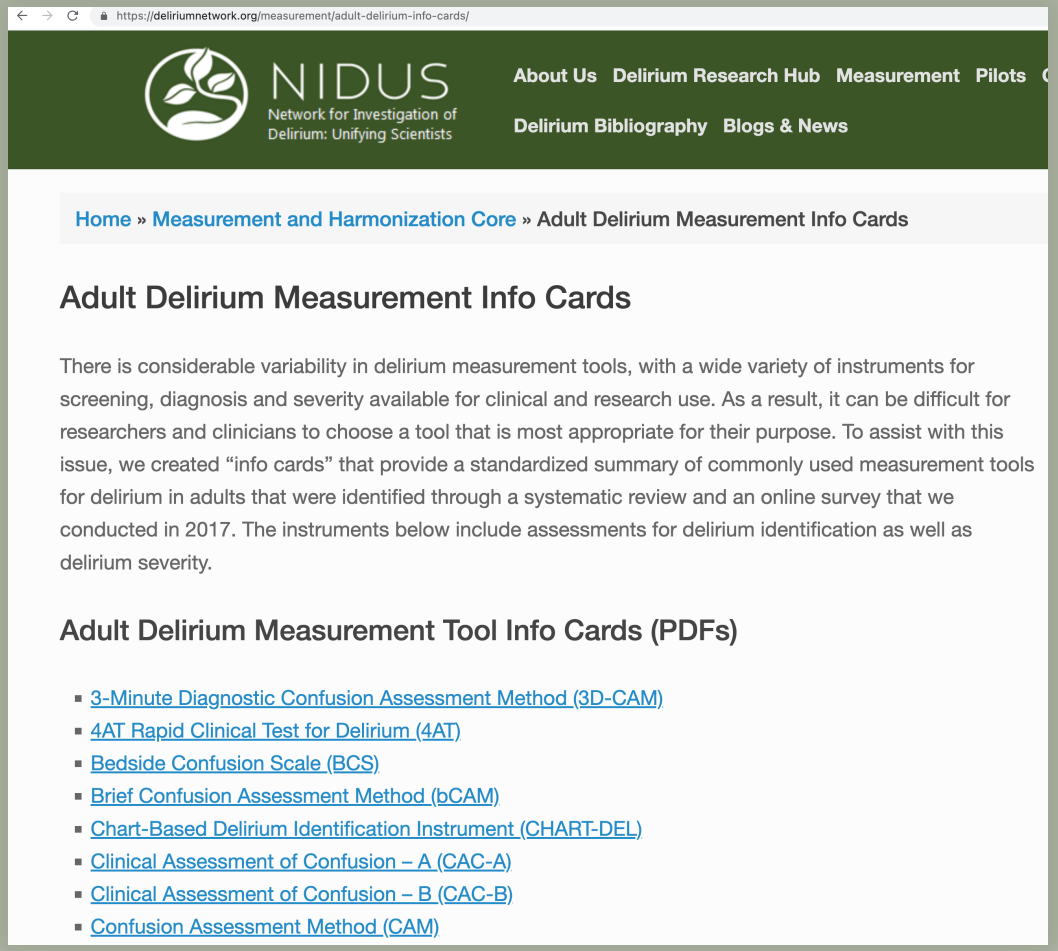


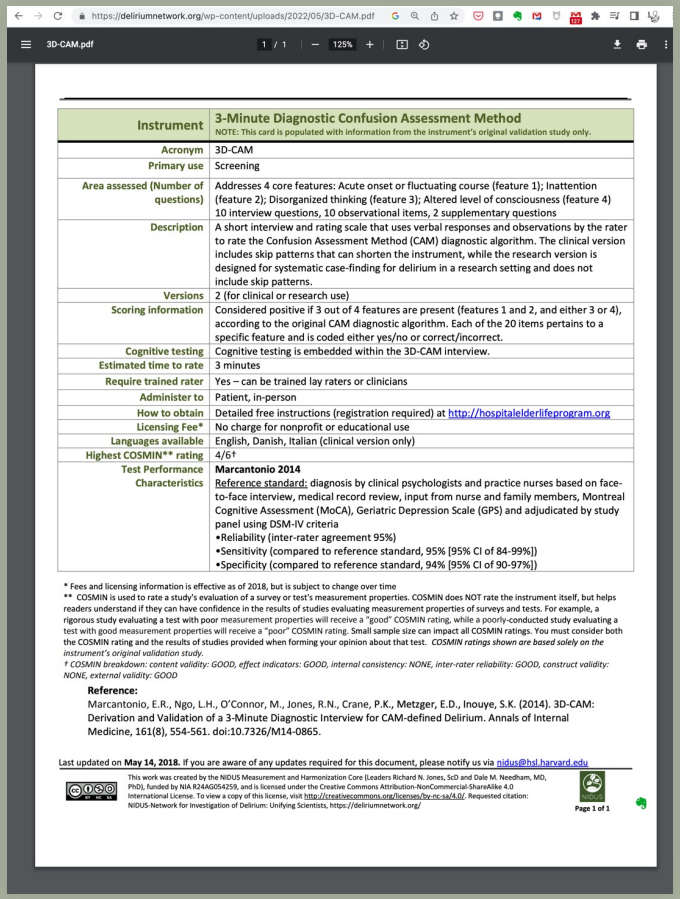
Measurement and Harmonization Core


https://deliriumnetwork.org
NIDUS
Network for Investigation of Delirium: Unifying Scientists
About Us
Delirium Research Hub
Measurement
Pilots
Career Dev/Bootcamp
Delirium Bibliography
Blogs & News
NIDUS Pilot Awards: Letters of Intent
NIDUS II Pilot Award Letters of Intent (LOI) are due December 7, 2022
Learn More about Pilot Award LOIs
NIDUS is a collaborative, multidisciplinary network dedicated to the acceleration of scientific discovery in delirium research, through focused collaboration and creation of sustainable infrastructure to enhance innovative and high-quality research.


https://deliriumnetwork.org/measurement/
NIDUS
Network for Investigation of Delirium: Unifying Scientists
About Us Delirium Research Hub Measurement Pilots
Home » Measurement and Harmonization Core
Measurement and Harmonization Core
The aims of the NIDUS Measurement and Harmonization Core are to develop a repository of core measures for delirium screening, diagnosis and severity, and to harmonize data across all types of delirium studies. These efforts will enhance the quality and consistency of measurement. It will also enable investigators to compare and share data across studies. Consequently, this allows for accelerated advances in delirium research.
For its harmonization aim, NIDUS will initially focus on harmonizing delirium severity measures. This project includes a systematic review of existing delirium severity measures. In addition, it includes statistical techniques for data harmonization across measures. The harmonization method will include a modified-Delphi consensus process. If interested in learning more about this effort, please contact us for more information on getting involved.
Learn more about the Measurement and Harmonization Core from Dr. Rich Jones:
00:38
07:19
Delirium Measurement Resources:
Info Cards: 1-page informational cards summarizing key information and test characteristics for commonly used delirium screening, diagnostic and severity assessments.
Adult Delirium Measurement Info Cards
Pediatric Delirium Measurement Info Cards
Delirium Severity Measure Crosswalk Tool
A score conversion tool for 3 delirium severity instruments - the CAM-S, DRS-R-98, and MDAS.


https://deliriumnetwork.org/measurement/
Info Cards: 1-page informational cards summarizing key information and test characteristics for commonly used delirium screening, diagnostic and severity assessments.
Adult Delirium Measurement Info Cards
Pediatric Delirium Measurement Info Cards
Delirium Severity Measure Crosswalk Tool
A score conversion tool for 3 delirium severity instruments - the CAM-S, DRS-R-98, and MDAS.
Delirium Identification Measures Crosswalk Tool
Linking between the CAM (short and long form), DOSS, DRS-R-98 (severity and total scores), and MDAS instruments.
Delirium Item Bank and Harmonization Tool
Statistical harmonization code (Stata version 16.1) to create the crosswalks and the Delirium Item Bank.
Delirium Severity Measurement Systematic Review Overview (PDF)
A brief overview of Jones et al. (2019) systematic review of delirium severity instruments and additional resources for delirium severity measurement.
Delirium Severity Measure Summary Table
Summary information on 14 delirium severity measures. These include number of items, approximate time to administer, certification or training required, and notes on background for development of the scale.
Delirium Severity Measure Symptom Coverage Table
Compares 17 commonly-used delirium severity measures by symptoms included in each measure.
In addition to these resources, the Measurement and Harmonization Core has several additional projects and publications in process. In time, these will be made available on this site.


https://deliriumnetwork.org/measurement/adult-delirium-info-cards/
NIDUS
Network for Investigation of Delirium: Unifying Scientists
About Us
Delirium Research Hub
Measurement
Pilots
Delirium Bibliography
Blogs & News
Home » Measurement and Harmonization Core » Adult Delirium Measurement Info Cards
Adult Delirium Measurement Info Cards
There is considerable variability in delirium measurement tools, with a wide variety of instruments for screening, diagnosis and severity available for clinical and research use. As a result, it can be difficult for researchers and clinicians to choose a tool that is most appropriate for their purpose. To assist with this issue, we created “info cards” that provide a standardized summary of commonly used measurement tools for delirium in adults that were identified through a systematic review and an online survey that we conducted in 2017. The instruments below include assessments for delirium identification as well as delirium severity.
Adult Delirium Measurement Tool Info Cards (PDFs)
3-Minute Diagnostic Confusion Assessment Method (3D-CAM)
4AT Rapid Clinical Test for Delirium (4AT)
Bedside Confusion Scale (BCS)
Brief Confusion Assessment Method (bCAM)
Chart-Based Delirium Identification Instrument (CHART-DEL)
Clinical Assessment of Confusion – A (CAC-A)
Clinical Assessment of Confusion – B (CAC-B)
Confusion Assessment Method (CAM)



https://deliriumnetwork.org/wp-content/uploads/2022/05/3D-CAM.pdf

3D-CAM.pdf 1 / 1 125%

Instrument	3-Minute Diagnostic Confusion Assessment Method NOTE: This card is populated with information from the instrument's original validation study only.
Acronym	3D-CAM
Primary use	Screening
Area assessed (Number of questions)	Addresses 4 core features: Acute onset or fluctuating course (feature 1); Inattention (feature 2); Disorganized thinking (feature 3); Altered level of consciousness (feature 4) 10 interview questions, 10 observational items, 2 supplementary questions
Description	A short interview and rating scale that uses verbal responses and observations by the rater to rate the Confusion Assessment Method (CAM) diagnostic algorithm. The clinical version includes skip patterns that can shorten the instrument, while the research version is designed for systematic case-finding for delirium in a research setting and does not include skip patterns.
Versions	2 (for clinical or research use)
Scoring information	Considered positive if 3 out of 4 features are present (features 1 and 2, and either 3 or 4), according to the original CAM diagnostic algorithm. Each of the 20 items pertains to a specific feature and is coded either yes/no or correct/incorrect.
Cognitive testing	Cognitive testing is embedded within the 3D-CAM interview.
Estimated time to rate	3 minutes
Require trained rater	Yes – can be trained lay raters or clinicians
Administer to	Patient, in-person
How to obtain	Detailed free instructions (registration required) at http://hospitalelderlifeprogram.org
Licensing Fee*	No charge for nonprofit or educational use
Languages available	English, Danish, Italian (clinical version only)
Highest COSMIN** rating	4/6†
Test Performance Characteristics	**Marcantonio 2014** Reference standard: diagnosis by clinical psychologists and practice nurses based on face-to-face interview, medical record review, input from nurse and family members, Montreal Cognitive Assessment (MoCA), Geriatric Depression Scale (GPS) and adjudicated by study panel using DSM-IV criteria •Reliability (inter-rater agreement 95%) •Sensitivity (compared to reference standard, 95% [95% CI of 84-99%]) •Specificity (compared to reference standard, 94% [95% CI of 90-97%])

* Fees and licensing information is effective as of 2018, but is subject to change over time

** COSMIN is used to rate a study's evaluation of a survey or test's measurement properties. COSMIN does NOT rate the instrument itself, but helps readers understand if they can have confidence in the results of studies evaluating measurement properties of surveys and tests. For example, a rigorous study evaluating a test with poor measurement properties will receive a "good" COSMIN rating, while a poorly-conducted study evaluating a test with good measurement properties will receive a "poor" COSMIN rating. Small sample size can impact all COSMIN ratings. You must consider both the COSMIN rating and the results of studies provided when forming your opinion about that test. *COSMIN ratings shown are based solely on the instrument's original validation study.*

† COSMIN breakdown: content validity: GOOD, effect indicators: GOOD, internal consistency: NONE, inter-rater reliability: GOOD, construct validity: NONE, external validity: GOOD

Reference:

Marcantonio, E.R., Ngo, L.H., O'Connor, M., Jones, R.N., Crane, P.K., Metzger, E.D., Inouye, S.K. (2014). 3D-CAM: Derivation and Validation of a 3-Minute Diagnostic Interview for CAM-defined Delirium. Annals of Internal Medicine, 161(8), 554-561. doi:10.7326/M14-0865.

Last updated on **May 14, 2018.** If you are aware of any updates required for this document, please notify us via nidus@hsl.harvard.edu

Instrument	3-Minute Diagnostic Confusion Assessment Method NOTE: This card is populated with information from the instrument's original validation study only.
Acronym	3D-CAM
Primary use	Screening
Area assessed (Number of questions)	Addresses 4 core features: Acute onset or fluctuating course (feature 1); Inattention (feature 2); Disorganized thinking (feature 3); Altered level of consciousness (feature 4) 10 interview questions, 10 observational items, 2 supplementary questions
Description	A short interview and rating scale that uses verbal responses and observations by the rater to rate the Confusion Assessment Method (CAM) diagnostic algorithm. The clinical version includes skip patterns that can shorten the instrument, while the research version is designed for systematic case-finding for delirium in a research setting and does not include skip patterns.

Versions	2 (for clinical or research use)
Scoring information	Considered positive if 3 out of 4 features are present (features 1 and 2, and either 3 or 4), according to the original CAM diagnostic algorithm. Each of the 20 items pertains to a specific feature and is coded either yes/no or correct/incorrect.
Cognitive testing	Cognitive testing is embedded within the 3D-CAM interview.
Estimated time to rate	3 minutes
Require trained rater	Yes – can be trained lay raters or clinicians
Administer to	Patient, in-person
How to obtain	Detailed free instructions (registration required) at http://hospitalelderlifeprogram.org
Licensing Fee*	No charge for nonprofit or educational use
Languages available	English, Danish, Italian (clinical version only)
Highest COSMIN rating**	4/6†

Languages available	English, Danish, Italian (clinical version only)
Highest COSMIN rating**	4/6†
Test Performance Characteristics	**Marcantonio 2014** Reference standard: diagnosis by clinical psychologists and practice nurses based on face-to-face interview, medical record review, input from nurse and family members, Montreal Cognitive Assessment (MoCA), Geriatric Depression Scale (GPS) and adjudicated by study panel using DSM-IV criteria •Reliability (inter-rater agreement 95%) •Sensitivity (compared to reference standard, 95% [95% CI of 84-99%]) •Specificity (compared to reference standard, 94% [95% CI of 90-97%])

* Fees and licensing information is effective as of 2018, but is subject to change over time

** COSMIN is used to rate a study's evaluation of a survey or test's measurement properties. COSMIN does NOT rate the instrument itself, but helps readers understand if they can have confidence in the results of studies evaluating measurement properties of surveys and tests. For example, a rigorous study evaluating a test with poor measurement properties will receive a "good" COSMIN rating, while a poorly-conducted study evaluating a test with good measurement properties will receive a "poor" COSMIN rating. Small sample size can impact all COSMIN ratings. You must consider both the COSMIN rating and the results of studies provided when forming your opinion about that test. *COSMIN ratings shown are based solely on the instrument's original validation study.*

† COSMIN breakdown: content validity: GOOD, effect indicators: GOOD, internal consistency: NONE, inter-rater reliability: GOOD, construct validity: NONE, external validity: GOOD

Reference:

Marcantonio, E.R., Ngo, L.H., O'Connor, M., Jones, R.N., Crane, P.K., Metzger, E.D., Inouye, S.K. (2014). 3D-CAM: Derivation and Validation of a 3-Minute Diagnostic Interview for CAM-defined Delirium. Annals of Internal Medicine, 161(8), 554-561. doi:10.7326/M14-0865.

COSMIN

Consensus-based Standards for the selection of health Measurement Instruments

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737-745.

ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 737–745

Journal of Clinical Epidemiology

The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes

Lidwine B. Mokkink[a,*], Caroline B. Terwee[a], Donald L. Patrick[b], Jordi Alonso[c,d], Paul W. Stratford[e,f], Dirk L. Knol[a], Lex M. Bouter[a,g], Henrica C.W. de Vet[a]

[a]*Department of Epidemiology and Biostatistics, The EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands*
[b]*Department of Health Services, University of Washington, Seattle, WA, USA*
[c]*Health Services Research Unit, Institut Municipal d'Investigacio Medica (IMIM-Hospital del Mar), Barcelona, Spain*
[d]*CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain*
[e]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada*
[f]*School of Rehabilitation Science, McMaster University, Hamilton, Canada*
[g]*Executive Board of VU University Amsterdam, Amsterdam, The Netherlands*

Accepted 5 February 2010

Reliability
Validity
Responsiveness†

Fig. 2. COSMIN taxonomy of relationships of measurement properties. *Abbreviations:* COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; HR-PRO, health related-patient reported outcome.

† Interpretability was thought to be important but not a measurement property, *per se*

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737-745.

Reliability	The degree to which the measurement is free from measurement error
Validity	The degree to which [the] instrument measures the construct(s) it purports to measure
Responsive-ness	The ability of [the] instrument to detect change over time in the construct to be measured
Interpretability	The degree to which one can assign qualitative meaning -- that is, clinical or commonly understood connotations -- to an instrument's quantitative scores or change in scores.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737-745., Table 2 (page 743)

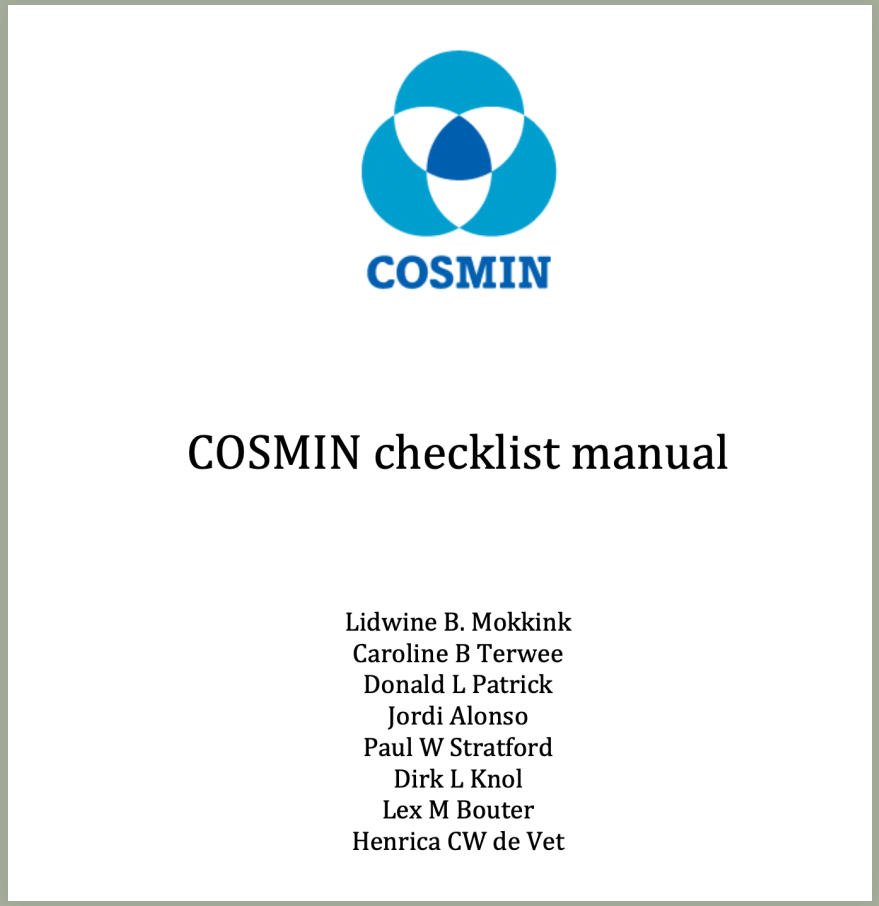
Validity

Content validity	The degree to which the content of [the] instrument is an adequate reflection of the construct to be measured
Hypothesis testing	The degree to which the scores of [the] instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the instrument validly measures the construct to be measured

Reliability

Internal consistency Reliability	The degree of the interrelatedness among the items

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737-745., Table 2 (page 743)



Mokkink, L., Terwee , C., Patrick, D., Alonso , J., Stratford, P., Knol, D., Bouter, L., & de Vet, H. (2012). *COSMIN checklist manual*. COSMIN network. www.emgo.nl

Four step procedure for completing the checklist

When completing the COSMIN checklist, four steps should be taken (Figure 2), which will be further explained in the next paragraphs.

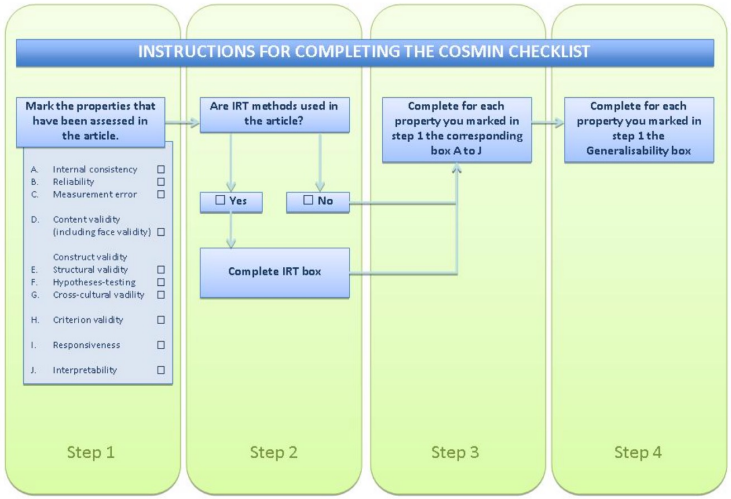


Figure 2. Four-step procedure for completing the COSMIN checklist.

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)

Box A. Internal consistency

		yes	no	?
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	☐	☐	☐
Design requirements		**yes**	**no**	**?**
2	Was the percentage of missing items given?	☐	☐	
3	Was there a description of how missing items were handled?	☐	☐	
4	Was the sample size included in the internal consistency analysis adequate?	☐	☐	☐
5	Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	☐	☐	
6	Was the sample size included in the unidimensionality analysis adequate?	☐	☐	☐
7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	☐	☐	☐
8	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**NA**
9	for Classical Test Theory (CTT): Was Cronbach's alpha calculated?	☐	☐	☐
10	for dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	☐	☐	☐
11	for IRT: Was a goodness of fit statistic at a global level calculated? e.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	☐	☐	☐

13	for ordinal scores: Was a weighted kappa calculated?	☐	☐	☐	☐
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	☐	☐	☐	

Effect indicators

A “COSMIN-guided” review of measurement properties

Does the scale consist ENTIRELY of effect indicators?

Effect indicators are CAUSED by delirium.

Effect indicators are appropriate for use in a reflective measurement model.

Cause or formative indicators are factors that might be risk factors for, or otherwise determine levels of, delirium or delirium severity.

Acknowledging that the pathophysiology of delirium is imperfectly understood, please use your best judgement.

Content validity

A “COSMIN-guided” review of measurement properties

Do the authors describe procedures for ensuring that all items refer to relevant aspects of delirium or delirium severity?

Content validity describes the extent to which the items included in a scale sample from the potential universe of possible questions that could be used to assess the target construct.

In the COSMIN framework (Mokkink et al 2010 BMC Med Res Meth https://doi.org/10.1186/1471-2288-10-22) the assessment of content validity is a judgement of the (a) relevance and (b) comprehensiveness of the items. Relevance refers to the match of questions to the target population. Comprehensiveness refers to the extent that the items included address the breadth of the domain or construct being assessed, and the clarity with which those constructs and or domains are defined. Both aspects of content validity are to be defended and adjudged with content expertise.

Internal consistency

A “COSMIN-guided” review of measurement properties

If internal consistency reliability was reported (e.g., Cronbach's alpha) what was the estimate?

If an internal consistency statistic was reported, please put check the other box and type of statistic in the other bullet. (e.g., "0.91 Cronbach's alpha").

What was the sample size used for internal consistency estimation?

Inter-rater reliability

A “COSMIN-guided” review of measurement properties

Inter-rater reliability refers to assessments of the agreement of two or more raters when making ratings on a single patient or research participant.

Was inter-rater reliability assessed?

What was the sample size for the assessment of inter-rater reliability?

Convergent validity

A “COSMIN-guided” review of measurement properties

Convergent validity describes the extent to which measure of a construct correlates with other measures of the same construct.

Sometimes this is called "Construct validity". However, nowadays we take all aspects of validity as evidence of construct validity.

What was the sample size for the assessment of convergent validity?

Criterion validity, Predictive validity, or Responsiveness

A “COSMIN-guided” review of measurement properties

This section contains items grouped under the heading "Hypothesis testing" in the COSMIN framework. Evidence for criterion validity would be relationship of a test score to a reference standard (e.g., a Psychiatrists diagnosis). Evidence for Predictive validity would come from the prediction of a clinically relevant outcome (e.g., death, length of stay, costs). Evidence for Responsiveness would be something like the measure is sensitive to change due to treatments or risk factors for the target condition.

Does the manuscript contain a description of the instrument's ability to associate predictably with external criteria, an outcome, or be influenced by a treatment or group with known difference on delirium. This includes a priori hypothesized mean differences and correlations with external variables.

What was the sample size used for describing criterion validity, predictive validity, or responsiveness. If more than one sample size would be appropriate, report the largest.

CAM

Number evaluated	1000
Prevalence of delirium	0.25
Sensitivity	0.82
Specificity	0.99

	Del+	Del-	Total
Test+	205	7	212
Test-	45	743	788
Total	250	750	1000

Positive predictive value	0.97
Negative predictive value	0.94
A' (AUC 1-point ROC)	0.95
Level of test (prop screen +)	0.21

Sensitivity and specificity values from Shi, Q., Warren, L., Saposnik, G., & MacDermid, J. C. (2013). Confusion assessment method: a systematic review and meta-analysis of diagnostic accuracy. Neuropsychiatric disease and treatment, 9, 1359.

4AT

Number evaluated	1000
Prevalence of delirium	0.25
Sensitivity	0.88
Specificity	0.88

	Del+	Del-	Total
Test+	220	90	310
Test-	30	660	690
Total	250	750	1000

Positive predictive value	0.71
Negative predictive value	0.96
A' (AUC 1-point ROC)	0.93
Level of test (prop screen +)	0.31

Sensitivity and specificity values from Tieges, Z., MacLullich, A. M. J., Anand, A., Brookes, C., Cassarino, M., Connor, M., . . . Galvin, R. (2020). Diagnostic Accuracy of the 4AT for delirium detection: systematic review and meta-analysis. medRxiv, 2020.2006.2011.20128280. doi:10.1101/2020.06.11.20128280

3D-CAM

Number evaluated	1000
Prevalence of delirium	0.25
Sensitivity	0.95
Specificity	0.94

	Del+	Del-	Total
Test+	238	45	283
Test-	12	705	717
Total	250	750	1000

Positive predictive value	0.84
Negative predictive value	0.98
A' (AUC 1-point ROC)	0.97
Level of test (prop screen +)	0.28

Sensitivity and specificity values from Marcantonio, E. R., Ngo, L. H., O'Connor, M., Jones, R. N., Crane, P. K., Metzger, E. D., & Inouye, S. K. (2014). 3D-CAM: Derivation and Validation of a 3-Minute Diagnostic Interview for CAM-Defined Delirium A Cross-sectional Diagnostic Test Study. Annals of Internal Medicine, 161(8), 554-561. doi:10.7326/M14-0865

See also: Tieges, Z., Maclullich, A. M., Anand, A., Brookes, C., Cassarino, M., O'connor, M., ... & Galvin, R. (2021). Diagnostic accuracy of the 4AT for delirium detection in older adults: systematic review and meta-analysis. *Age and ageing*, *50*(3), 733-743.

Scoring

A “COSMIN-guided” review of measurement properties

Assign 1 point if each of (1) CONTENT VALIDITY, (2) ALL EFFECT INDICATORS, (3) INTERNAL CONSISTENCY, (4) any aspect of RELIABILITY (other than internal consistency, e.g., inter-rater), (5) CONVERGENT VALIDITY or construct validity, and (6) CRITERION validity or predictive validity or external validity were assessed.

Subtract 1/2 point if INTERNAL CONSISTENCY was based on fewer than 50 observations,

Subtract 1/2 point if if RELIABILITY was based on less than 50 observations

Subtract 1/2 point if CONVERGENT validity was based on less than 50 persons,

Subtract 1/2 point if CRITERION validity was based on less than 50 persons.

NIDUS Measurement core COSMIN rating

- Is a very high-level summary of the *original publication* describing the instrument
- Does not reflect any validation research subsequent to the original publication
- Only partially represents the full COSMIN framework
- Might be unfairly applied to instruments described before the circa 2010 COSMIN framework was described

Strategies for choosing an instrument

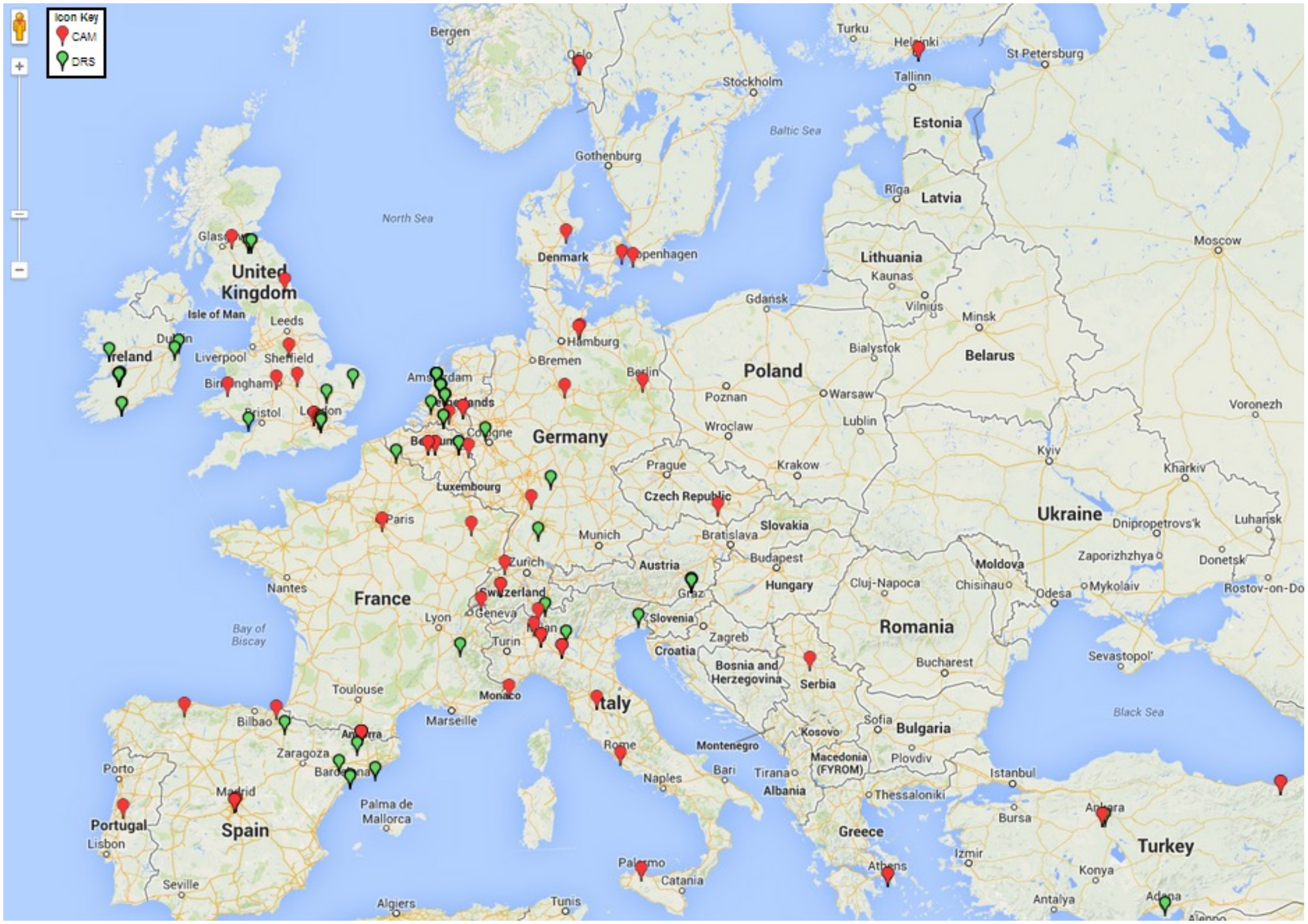
Part III

Feasibility

- What instrument(s) is/are used in your lab/hospital/city by mentors/collaborators?
- Do you have access to training or other resources to make effective use of the instruments?

Reliability & Validity

- Are the instruments suitable for the target population?
- Do you have the right assessors?
- Has the instrument be used in your target population previously?
- With success?
- Do instruments maximize sensitivity and specificity in a way most beneficial to your question?



2014


Icon Key
CAM
DRS
North Dakota
Minnesota
South Dakota
Wyoming
Nebraska
Iowa
Wisconsin
Michigan
Illinois
Indiana
Ohio
Pennsylvania
New York
Maine
Vermont
New Hampshire
Massachusetts
Connecticut
Rhode Island
New Jersey
Maryland
Delaware
District of Columbia
West Virginia
Virginia
Kentucky
Tennessee
North Carolina
South Carolina
Georgia
Florida
Alabama
Mississippi
Louisiana
Arkansas
Oklahoma
Texas
Kansas
Missouri
Colorado
New Mexico
United States
Kansas City
St. Louis
Columbus
Indianapolis
Chicago
Detroit
Toronto
Ottawa
Montreal
Quebec City
New Brunswick
Bermuda
Albuquerque
Ciudad Juarez
Fort Worth
Dallas
Austin
Houston
San Antonio
Chihuahua
Monterrey
Charlotte
Atlanta
Jacksonville
Orlando
Tampa
Miami
Gulf of

2014

Final thought

- If you would like to know which of two or more instruments is the “best” for your target population (sensitivity, specificity, predictive value, reliability)
- The only ***trustworthy*** data to inform this decision would be
  - Head-to-head comparison in same sample (e.g., randomized design)
  - Individual participant data meta-analysis (mega-analysis)
- Individual (but separately conducted) studies **and meta-analyses** are not directly comparable (selection of patients, other design and analysis choices), publication bias, etc.

Measurement and Harmonization Core:

On the web

https://deliriumnetwork.org/measurement/

Questions

Rich Jones
Rich_jones@brown.edu

NIDUS
nidus@hsl.harvard.edu

CAM

Number evaluated	**1000**
Prevalence of delirium	**0.25**
Sensitivity	**0.79**
Specificity	**0.98**

	Del+	Del-	Total
Test+	198	15	213
Test-	52	735	787
Total	250	750	1000

Positive predictive value	**0.93**
Negative predictive value	0.93
A' (AUC 1-point ROC)	0.94
Level of test (prop screen +)	0.21
hit rate (correct decision)	0.93
Efficiency	0.27
Kraemer's k(1,0)	0.73
Kraemer's k(0,0)	0.90
Kraemer's k(.5,0)	0.81

Sensitivity and specificity values from an unpublished update of two meta-analyses of CAM diagnostic test accuracy.

4AT

Number evaluated	1000
Prevalence of delirium	0.25
Sensitivity	**0.88**
Specificity	**0.88**

	Del+	Del-	Total
Test+	220	90	310
Test-	30	660	690
Total	250	750	1000

Positive predictive value	**0.71**
Negative predictive value	0.96
A' (AUC 1-point ROC)	0.93
Level of test (prop screen +)	0.31
hit rate (correct decision)	0.88
Efficiency	0.33
Kraemer's k(1,0)	0.83
Kraemer's k(0,0)	0.61
Kraemer's k(.5,0)	0.70

Sensitivity and specificity values from Tieges, Z., MacLullich, A. M. J., Anand, A., Brookes, C., Cassarino, M., Connor, M., . . . Galvin, R. (2020). Diagnostic Accuracy of the 4AT for delirium detection: systematic review and meta-analysis. medRxiv, 2020.2006.2011.20128280. doi:10.1101/2020.06.11.20128280

3D-CAM

Number evaluated	1000
Prevalence of delirium	0.25
Sensitivity	**0.95**
Specificity	**0.94**

	Del+	Del-	Total
Test+	238	45	283
Test-	12	705	717
Total	250	750	1000

Positive predictive value	**0.84**
Negative predictive value	0.98
A' (AUC 1-point ROC)	0.97
Level of test (prop screen +)	0.28
hit rate (correct decision)	0.94
Efficiency	0.34
Kraemer's k(1,0)	0.93
Kraemer's k(0,0)	0.79
Kraemer's k(.5,0)	0.85

Sensitivity and specificity values from Marcantonio, E. R., Ngo, L. H., O'Connor, M., Jones, R. N., Crane, P. K., Metzger, E. D., & Inouye, S. K. (2014). 3D-CAM: Derivation and Validation of a 3-Minute Diagnostic Interview for CAM-Defined Delirium A Cross-sectional Diagnostic Test Study. Annals of Internal Medicine, 161(8), 554-561. doi:10.7326/M14-0865

Random Forest using EHR (Best algorithm, Corradi 2018)

Number evaluated	1000
Prevalence of delirium	0.25
Sensitivity	**0.78**
Specificity	**0.88**

	Del+	Del-	Total
Test+	195	90	285
Test-	55	660	715
Total	250	750	1000

Positive predictive value	**0.68**
Negative predictive value	0.92
A' (AUC 1-point ROC)	0.90
Level of test (prop screen +)	0.29
hit rate (correct decision)	0.86
Efficiency	0.30
Kraemer's k(1,0)	0.69
Kraemer's k(0,0)	0.59
Kraemer's k(.5,0)	0.63

Sensitivity and specificity values from Corradi, J. P., Thompson, S., Mather, J. F., Waszynski, C. M., & Dicks, R. S. (2018). Prediction of incident delirium using a random forest classifier. Journal of medical systems, 42(12), 1-10; as per Chua, S. J., Wrigley, S., Hair, C., & Sahathevan, R. (2021). Prediction of delirium using data mining: A systematic review. Journal of Clinical Neuroscience, 91, 288-298.

Kraemer's k(1,0) is chance-corrected sensitivity (or chance-corrected positive predictive value); Kraemer's k(0,0) is chance-corrected specificity (or chance-corrected negative predictive value); Kraemer's k(.5,0) is the weighted average of the k(1,0) and k(0,0) and is also known as Cohen's kappa when sensitivity and specificity are equally weighted.