Cognitive screening instruments for dementia: comparing metrics of test limitation

Andrew J Larner¹

1 Cognitive Function Clinic, Walton Centre for Neurology and Neurosurgery,
Liverpool, United Kingdom

Correspondence: AJ Larner, Cognitive Function Clinic, Walton Centre for Neurology
and Neurosurgery, Lower Lane, Fazakerley, Liverpool, L9 7LJ, United Kingdom
e-mail: a.larner@thewaltoncentre.nhs.uk

Abstract

Cognitive screening instruments (CSIs) for dementia and mild cognitive impairment are usually characterised in terms of measures of discrimination such as sensitivity, specificity, and likelihood ratios. However, CSIs also have limitations. Several metrics exist which may be used to denote test limitations but they are seldom examined. Data from several pragmatic test accuracy studies of CSIs were interrogated to calculate various measures of limitation, namely: misclassification rate; net harm to net benefit ratio; and the likelihood to be diagnosed or misdiagnosed. Intra- and inter-test performance for measures of discrimination and limitation were compared. The study found that some tests with very high sensitivity but low specificity for dementia fared poorly on measures of limitation, with high misclassification rates, low net harm to net benefit ratios, and low likelihoods to be diagnosed or misdiagnosed; some had likelihoods favouring misdiagnosis over diagnosis. Tests with a better balance of sensitivity and specificity fared better on measures of limitation. When choosing which CSIs to administer, measures of test limitation should be considered as well as measures of test discrimination. Although high test sensitivity may be desirable to avoid false negatives, false positives also have a cost. Identification of tests having high misclassification rate, low net harm to net benefit ratio, and low likelihood to be diagnosed or misdiagnosed, may have implications for their use in clinical practice.

Keywords: cognitive screening; dementia; diagnosis; limitations; memory clinic

Introduction

Like all screening and diagnostic tests, cognitive screening instruments (CSIs) are usually characterised in terms of the conditional probabilities of sensitivity (Sens) and specificity (Spec), where Sens (or true positive rate, TPR) is the correct identification of those with dementia or cognitive impairment, and Spec (or true negative rate, TNR) is the correct exclusion of those without disease (see Box 1 for definitions of metrics discussed in this paper, their formulae and score ranges). Information from both Sens and Spec may be combined in metrics such as the Youden index (Y) and positive and negative likelihood ratios (LR+, LR-), of which the latter may be qualitatively classified as causing slight, moderate, large, or very large change in probability of disease or its absence.¹ Sens and Spec are suggested keywords for reports of diagnostic test accuracy studies in dementia (STARDdem)² and LRs were used as the basis for recommendations made by the UK National Institute for Health and Care Excellence for tests suitable for dementia.³ Systematic reviews and meta-analyses of CSIs, for example those produced by the Cochrane Dementia and Cognitive Improvement Group,⁴ typically quote summary test Sens, Spec, and LRs.

Like all screening and diagnostic tests, CSIs are not perfect. They have shortcomings, inadequacies, or failures, which may be termed “limitations”. Tests have potential harms (misdiagnosis) as well as benefits (correct diagnosis). The limitations comprise failure to identify dementia or cognitive impairment when it is in fact present and identifying these states when they are in fact absent. These rates, respectively false negative (FNR) and false positive (FPR), are implicit in the measures of Sens and Spec since, by the principle of summation, they are their complements or negations ($FNR = 1 - \text{Sens}$; $FPR = 1 - \text{Spec}$). Other metrics of test limitation include inaccuracy (Inacc; also sometimes known as fraction incorrect, or error rate) and error odds ratio (EOR), although these measures are seldom used in clinical practice.

Other metrics of test limitation, which, like all those already mentioned, may be derived from the 2x2 contingency table of diagnostic test accuracy studies, form the subject of the current study. These are: the misclassification rate; the net harm to net benefit ratio; and the likelihood to be diagnosed or misdiagnosed.

The sum of FNR and FPR is used here to define the misclassification rate (MR), following the usage of Perkins and Schisterman.⁵ (Confusingly this term has also been sometimes used interchangeably with Inacc.) Minimization of MR is used in some of the methods for setting a test threshold from inspection of the receiver operating characteristic (ROC) curve of a test accuracy study.

The ratio of net harm to net benefit (H/B) may be defined as the harm (H) of treating a person without disease (i.e. false positive) to the net benefit (B) of treating a person with disease (i.e. true positive), the latter term equating to the harm of a false negative result.⁶ The H/B ratio may be calculated from the Bayes’ equation as the product of the pre-test odds of disease and the positive likelihood ratio at the specified test cut-off (which is equivalent to the slope of the ROC curve, TPR/FPR, at that point) and hence is equivalent to the post-test odds.⁷ A higher H/B ratio means the test is less likely to miss cases, and hence less likely to incur the harms of false negatives, and hence a higher H/B ratio is deemed better (note that this scoring of H/B ratio may seem counterintuitive if one thinks solely of “harms” and “benefits”, hence the important qualification of “net”).

More recently, another metric attempting to denote test limitation has been introduced: the likelihood to be diagnosed or misdiagnosed (LDM).^{8,9} LDM is based on “number needed” metrics which are generally deemed to be more intuitive and hence applicable for both clinicians and patients than Sens and Spec. One form of LDM is given by the ratio of the number needed to misdiagnose,¹⁰ which is the inverse of Inacc, to the number needed to diagnose, which is the inverse of Youden index. Hence LDM may also be conceptualised as a ratio of harms (misdiagnosis) and benefits (diagnosis). LDM ranges from 0 to infinity but, as for likelihood ratios, has an inflection point at 1 such that $LDM < 1$ indicates a test in which misdiagnosis is overall more likely than diagnosis, and $LDM > 1$ indicates a test in which diagnosis is overall more likely than misdiagnosis, and hence $LDM \gg 1$ is desirable and $LDM = \infty$ is the perfect diagnostic test (where Sens = Spec = Y = 1, and Inacc = 0).

The purpose of this study was to compare these three indices of test limitation (MR, H/B ratio, LDM) for several brief CSIs in common clinical usage for dementia diagnosis, namely the Mini-Mental State Examination (MMSE),¹¹ the Montreal Cognitive Assessment (MoCA),¹² the Mini-Addenbrooke’s Cognitive Examination (MACE),¹³ the Six-item Cognitive Impairment Test (6CIT),¹⁴ the informant Ascertain Dementia 8 (iAD8),¹⁵ and the Test Your Memory (TYM) test,¹⁶ as well as for a more recently described instrument, Free-Cog.

Methods

Data from previously undertaken and reported pragmatic prospective test accuracy studies¹⁷ were used (Table 1), each undertaken using a standardised methodology in a dedicated cognitive disorders clinic located in a regional neuroscience centre. In all studies subjects gave informed consent and study protocol was approved by the institute’s committee on human research (Walton Centre for Neurology and Neurosurgery Approval: N 310).

In each study, criterion diagnosis of dementia followed standard diagnostic criteria (DSM-IV) and was made independent of scores on index CSIs to avoid review bias. Cross classification of criterion diagnosis with index test result, dichotomised by test cut-off, in a standard 2x2 contingency table allowed all cases to be classified as true positive (TP), false positive (FP), false negative (FN) and true negative (TN). Where possible, test cut-offs documented in the respective index studies were used to avoid bias. All studies followed either the STAndards for the Reporting of Diagnostic accuracy studies (STARD)¹⁸ or the derived guidelines specific for dementia studies, STARDdem,² dependent on the exact date at which each test accuracy study was undertaken.

Results

Examining measures of test discrimination (Table 2), many were highly sensitive (MoCA, Free-Cog, MACE, AD8) but had low specificity (MoCA, MACE, AD8). Positive likelihood ratios were qualitatively either slight (MoCA, MACE, AD8, TYM) or moderate (6CIT, Free-Cog, MMSE), none achieving the large or very large classification.

Examining measures of test limitation (Table 3), few achieved a misclassification rate of ≤ 0.5 (Free-Cog, 6CIT, MMSE). Only one test (6CIT) achieved H/B ratio of 1. LDM values of < 1 (likelihood of misdiagnosis greater than correct diagnosis) were recorded for some tests (MoCA, MACE, AD8). Of note, the tests with high sensitivity but low specificity generally fared worse on these metrics examining test limitation, whilst those with a better balance of Sens and Spec (reflected in the higher LR+s) did better. This was also evident in the overall ranking of CSIs by outcome of the examined measures of discrimination and limitation (Table 4).

Discussion

This study has various shortcomings. The findings are of course dependent upon the diagnostic test accuracy studies upon which they are based. These index studies obviously have limitations, for example they were undertaken in different patient populations, albeit all seen in the same clinic and operating the same diagnostic criteria for dementia, and hence may not necessarily be generalizable. Nevertheless the findings suggest significant limitations for many of the CSIs in common usage. The findings might be corroborated by undertaking similar analyses with data reported in systematic reviews of these CSIs where available.

The measures of limitation examined here are seldom used in clinical practice, may be unfamiliar to clinicians, and have no exact ranges. Other methods of assessing test effectiveness and limitation are also available. The metrics examined here do not address utilities⁷ or cost ratios.¹⁹

For MR and the H/B ratio, lower or higher values respectively may be better, but precisely how high or how low is most desirable or optimal has not been defined. LDM values have clearer implications around the inflection point of 1. The influence of disease prevalence on MR is unknown, although as it is based (like Sens, Spec, FPR, and FNR) on strict columnar ratios from the 2x2 contingency table it is notionally uninfluenced by the base rate, although it is well recognised that these measures are affected by the heterogeneity (spectrum bias) of clinical populations.²⁰ Another formulation of LDM, with the denominator based on predictive values, takes account of disease prevalence.^{8,9,17}

Whilst clinicians may be content to use highly sensitive tests, accepting false positives as a reasonable trade-off to ensure no cases are missed (i.e. low false negative rate), metrics of limitation highlight the potential shortcomings of such tests, and emphasize the need to find better tests. Patients undergoing testing may also want to have easily assimilated information on how well the test performs (a false positive diagnosis may have more significance for a patient than for a clinician) as well as its potential risks. Newer biomarker tests of dementia disorders could be subjected to similar analyses of test limitation.

References

1. Jaeschke R, Guyatt G, Sackett, DL. Users' guide to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA* **271**: 703-707, doi: 10.1001/jama.271.9.703 (1994).
2. Noel-Storr AH, McCleery JM, Richard E, Ritchie CW, Flicker L, Cullum SJ, *et al.* Reporting standards for studies of diagnostic test accuracy in dementia: the STARDdem Initiative. *Neurology* **83**: 364-373, doi: 10.1212/WNL.0000000000000621 (2014).
3. National Institute for Health and Care Excellence. *Dementia. Assessment, management and support for people living with dementia and their carers.* NICE Guideline 97. Methods, evidence and recommendations. London: NICE. <https://www.nice.org.uk/guidance/ng97> (2018).
4. Davis DH, Creavin ST, Noel-Storr A, Quinn TJ, Smailagic N, Hyde CH, *et al.* Neuropsychological tests for the diagnosis of Alzheimer's disease dementia and other dementias: a generic protocol for cross-sectional and delayed-verification studies. *Cochrane Database Syst Rev* **3**: CD010460, doi: 10.1002/14651858.CD010460 (2013).
5. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol* **163**: 670-675, doi: 10.1093/aje/kwj063 (2006).
6. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Med (Zagreb)* **26**: 297-307, doi: 10.11613/BM.2016.034 (2016).
7. Hunink MGM, Weinstein MC, Wittenberg E, Drummond MF, Pliskin JS, Wong JB, *et al.* *Decision making in health and medicine. Integrating evidence and values (2nd edition).* Cambridge: Cambridge University Press (2014).
8. Larner AJ. Number needed to diagnose, predict, or misdiagnose: useful metrics for non-canonical signs of cognitive status? *Dement Geriatr Cogn Dis Extra* **8**: 321-327, doi: 10.1159/000492783 (2018).
9. Larner AJ. Evaluating cognitive screening instruments with the "likelihood to be diagnosed or misdiagnosed" measure. *Int J Clin Pract* **73**: e13265, doi: 10.1111/ijcp.13265 (2019).
10. Habibzadeh F, Yadollahie M. Number needed to misdiagnose: a measure of diagnostic test effectiveness. *Epidemiology* **24**: 170, doi: 10.1097/EDE.0b013e31827825f2 (2013).
11. 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* **12**: 189-198, doi: 10.1016/0022-3956(75)90026-6 (1975).
12. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* **53**: 695-699, doi: 10.1111/j.1532-5415.2005.53221.x (2005).
13. Hsieh S, McGrory S, Leslie F, Dawson K, Ahmed S, Butler CR, *et al.* The Mini-Addenbrooke's Cognitive Examination: a new assessment tool for dementia. *Dement Geriatr Cogn Disord* **39**: 1-11, doi: 10.1159/000366040. (2015).

14. Brooke P, Bullock R. Validation of a 6 item cognitive impairment test with a view to primary care usage. *Int J Geriatr Psychiatry* **14**: 936-940, doi: 10.1002/(SICI)1099-1166(199911)14:11<936::AID-GPS39>3.0.CO;2-1 (1999).
15. Galvin JE, Roe CM, Powlishta KK, Coats MA, Muich SJ, Grant E, *et al.* The AD8. A brief informant interview to detect dementia. *Neurology* **65**: 559-564, doi: 10.1212/01.wnl.0000172958.95282.2a (2005).
16. Brown J, Pengas G, Dawson K, Brown LA, Clatworthy P. Self administered cognitive screening test (TYM) for detection of Alzheimer's disease: cross sectional study. *BMJ* **338**: b2030, doi: 10.1136/bmj.b2030 (2009).
17. Lerner AJ. *Manual of screeners for dementia. Pragmatic test accuracy studies*. London: Springer. doi: 10.1007/978-3-030-41636-2 (2020).
18. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis GA, Glasziou PP, Irwig LM, *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* **49**: 7-18, doi: 10.1373/49.1.7 (2003).
19. Kraemer HC. *Evaluating medical tests. Objective and quantitative guidelines*. Newbery Park, California: Sage (1992).
20. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol* **62**: 5-12, doi: 10.1016/j.jclinepi.2008.04.007 (2009).

Box 1: Metrics, formulae, and ranges for measures of test effectiveness and limitation

Metric	Formula	Range
Measures of test discrimination:		
Sensitivity (Sens, TPR)	$TP/(TP + FN)$	0-1, higher better
Specificity (Spec, TNR)	$TN/(TN + FP)$	0-1, higher better
Youden index (Y)	$Sens + Spec - 1$	0-1, higher better*
Positive likelihood ratio (LR+)	$Sens/(1 - Spec) = TPR/FPR$	1-∞, higher better
Negative likelihood ratio (LR-)	$(1 - Sens)/Spec = FNR/TNR$	0-1, lower better
Number needed to diagnose (NND)	$1/Y$	1-∞, lower better
Number needed to misdiagnose (NNM)	$1/Inacc$	1-∞, higher better
Measures of test limitation:		
False negative rate (FNR)	$FN/(TP + FN) = 1 - Sens$	0-1, lower better
False positive rate (FPR)	$FP/(TN + FP) = 1 - Spec$	0-1, lower better
Inaccuracy (Inacc)	$(FP + FN)/(TP + FP + FN + TN)$	0-1, lower better
Error odds ratio (EOR)	$TP \times FP / FN \times TN$	0-∞, lower better
Misclassification rate (MR)	$FNR + FPR = 2 - (Sens + Spec)$	0-2, lower better
Net harm/net benefit ratio (H/B)	pre-test odds x LR+	higher better if FN deemed worse than FP
Likelihood to be diagnosed or misdiagnosed (LDM)	$NNM/NND = Y/Inacc$	0-∞, <1 favours misdiagnosis, >1 favours diagnosis*

*Values of Y may in fact range from -1 to +1, as is evident from the formula. However, most values of Y would be expected to fall within the range 0 (worthless test, no diagnostic value) to 1 (perfect diagnostic value, no false positives or false negatives), with negative values occurring only if the test were misleading, i.e. the test result was negatively associated with the true diagnosis. LDM could also theoretically have negative values

Table 1: Study demographics

CSI	N	Age, median (years)	Gender (F:M; %F)	Test threshold for dementia	Pre-test odds
MMSE	244	60	117:127; 48	<26/30	0.22
MoCA	260	59	118:142; 45	<26/30	0.21
MACE	755	60	352:403; 47	≤25/30	0.18
6CIT	245	59	121:124; 49	>9/28	0.25
AD8	212	64.5	106:106; 50	≥2/8	0.43
TYM	224	62	94:130; 42	≤42/50	0.54
Free-Cog	141	62	61:80; 43	≤22/30	0.12

Abbreviations: CSI = Cognitive screening instrument; MCI = mild cognitive impairment; TP = true positive; FN = false negative; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment; MACE = Mini-Addenbrooke's Cognitive Examination; 6CIT = Six-item Cognitive Impairment Test; AD8 = Ascertain Dementia 8; TYM = Test Your Memory

Table 2: Comparing metrics of test discrimination for CSIs for diagnosis of dementia vs no dementia (with ranking in brackets)

CSI	Sens	Spec	LR+ and qualitative classification
MMSE	0.86 (7)	0.64 (3)	2.37 = moderate (3)
MoCA	1.00 (1=)	0.31 (6)	1.46 = slight (5)
MACE	0.99 (3)	0.32 (5)	1.45 = slight (6)
6CIT	0.88 (6)	0.78 (1)	4.00 = moderate (1)
AD8	0.98 (4)	0.10 (7)	1.09 = slight (7)
TYM	0.95 (5)	0.45 (4)	1.73 = slight (4)
Free-Cog	1.00 (1=)	0.67 (2)	3.07 = moderate (2)

Abbreviations: CSI = Cognitive screening instrument; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment; MACE = Mini-Addenbrooke's Cognitive Examination; 6CIT = Six-item Cognitive Impairment Test; AD8 = Ascertain Dementia 8; TYM = Test Your Memory; Sens = Sensitivity; Spec = Specificity; LR+ = Positive likelihood ratio

Table 3: Comparing metrics of test limitation for CSIs for diagnosis of dementia vs no dementia (with ranking in brackets)

CSI	MR (lower better)	H/B (higher better)	LDM (higher better)
MMSE	0.50 (3)	0.52 (3)	1.56 (3)
MoCA	0.69 (5=)	0.30 (6)	0.54 (5)
MACE	0.69 (5=)	0.26 (7)	0.53 (6)
6CIT	0.34 (2)	1.00 (1)	3.30 (1)
AD8	0.92 (7)	0.47 (4)	0.12 (7)
TYM	0.60 (4)	0.93 (2)	1.08 (4)
Free-Cog	0.33 (1)	0.38 (5)	2.31 (2)

Abbreviations: CSI = Cognitive screening instrument; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment; MACE = Mini-Addenbrooke's Cognitive Examination; 6CIT = Six-item Cognitive Impairment Test; AD8 = Ascertain Dementia 8; TYM = Test Your Memory; MR = Misclassification rate; H/B = Net harm/net benefit ratio; LDM = Likelihood to be diagnosed or misdiagnosed

Table 4: Ranking of CSIs by outcome measures of test discrimination and limitation

Rank	Measures of test discrimination			Measures of test limitation		
	Sens	Spec	LR+	MR	H/B	LDM
1	MoCA, Free-Cog	6CIT	6CIT	Free-Cog	6CIT	6CIT
2	-	Free-Cog	Free-Cog	6CIT	TYM	Free-Cog
3	MACE	MMSE	MMSE	MMSE	MMSE	MMSE
4	AD8	TYM	TYM	TYM	AD8	TYM
5	TYM	MACE	MoCA	MACE, MoCA	Free-Cog	MoCA
6	6CIT	MoCA	MACE	-	MoCA	MACE
7	MMSE	AD8	AD8	AD8	MACE	AD8

Abbreviations: CSI = Cognitive screening instrument; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment; MACE = Mini-Addenbrooke's Cognitive Examination; 6CIT = Six-item Cognitive Impairment Test; AD8 = Ascertain Dementia 8; TYM = Test Your Memory; Sens = Sensitivity; Spec = Specificity; LR+ = Positive likelihood ratio; MR = Misclassification rate; H/B = Net harm/net benefit ratio; LDM = Likelihood to be diagnosed or misdiagnosed