Cognitive screening instruments for dementia: comparing metrics of test limitation

Andrew J. Larner

ABSTRACT. Cognitive screening instruments (CSIs) for dementia and mild cognitive impairment are usually characterized in terms of measures of discrimination such as sensitivity, specificity, and likelihood ratios, but these CSIs also have limitations. **Objective:** The aim of this study was to calculate various measures of test limitation for commonly used CSIs, namely, misclassification rate (MR), net harm/net benefit ratio (H/B), and the likelihood to be diagnosed or misdiagnosed (LDM). **Methods:** Data from several previously reported pragmatic test accuracy studies of CSIs (Mini-Mental State Examination, the Montreal Cognitive Assessment, Mini-Addenbrooke’s Cognitive Examination, Six-item Cognitive Impairment Test, informant Ascertain Dementia 8, Test Your Memory test, and Free-Cog) undertaken in a single clinic were reanalyzed to calculate and compare MR, H/B, and the LDM for each test. **Results:** Some CSIs with very high sensitivity but low specificity for dementia fared poorly on measures of limitation, with high MRs, low H/B, and low LDM; some had likelihoods favoring misdiagnosis over diagnosis. Tests with a better balance of sensitivity and specificity fared better on measures of limitation. **Conclusions:** When deciding which CSI to administer, measures of test limitation as well as measures of test discrimination should be considered. Identification of CSIs with high MR, low H/B, and low LDM, 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 characterized 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 Table 1 for definitions of metrics discussed in this study, their formulae, and score ranges).

This study was conducted at the Cognitive Function Clinic, Walton Centre for Neurology and Neurosurgery, Liverpool, United Kingdom

Disclosure: The authors report no conflicts of interest.

Funding: none.

Received on April 18, 2021. Accepted in final on June 14, 2021.
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. Sens and Spec are suggested key words for reports of diagnostic test accuracy studies in dementia (STARDdem) 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. Systematic reviews and meta-analyses of CSIs, for example, those produced by the Cochrane Dementia and Cognitive Improvement Group, 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−Sens; FPR=1−Spec). Other metrics of test limitation include inaccuracy (Inacc; also sometimes known as fraction incorrect or error rate) and error odds ratio, 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 2×2 contingency table of diagnostic test accuracy studies, form the subject of the current study. These are the misclassification rate (MR), the net harm/net benefit ratio (H/B), and the likelihood to be diagnosed or misdiagnosed (LDM).

The sum of FNR and FPR is used here to define the MR, following the usage of Perkins and Schisterman. (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.

### Table 1. Metrics, formulae, and ranges for measures of test discrimination and limitation.

| Metric | Formula | Range |
|--------|---------|-------|
| **Measures of test discrimination** | | |
| Sens, TPR | TP/(TP+FN) | 0–1, higher better |
| Spec, TNR | TN/(TN+FP) | 0–1, higher better |
| Y | Sens+Spec−1 | 0–1, higher better* |
| LR+ | Sens/(1−Spec)=TPR/FPR | 1–∞, higher better |
| LR− | (1−Sens)/Spec=FNR/TNR | 0–1, lower better |
| NND | 1/Y | 1–∞, lower better |
| NNM | 1/Inacc | 1–∞, higher better |
| **Measures of test limitation** | | |
| FNR | FN/(TP+FN)=1−Sens | 0–1, lower better |
| FPR | FP/(TN+FP)=1−Spec | 0–1, lower better |
| Inacc | (FP+FN)/(TP+FP+FN+TN) | 0–1, lower better |
| EOR | TP×FP/FN×TN | 0–∞, lower better |
| MR | FNR+FPR=2−(Sens+Spec) | 0–2, lower better |
| net H/B ratio | pre-test odds×LR+ | Higher better if FN deemed worse than FP |
| LDM | NNM/NND=Y/Inacc | -1–∞, <1 favors misdiagnosis, >1 favors diagnosis |

Sens, TPR: Sensitivity; Spec, TNR: Specificity; Y: Youden index; LR+: positive likelihood ratio; LR−: negative likelihood ratio; NND: number needed to diagnose; NNM: number needed to misdiagnose; FNR: false negative rate; FPR: false positive rate; Inacc: Inaccuracy; EOR: error odds ratio; MR: misclassification rate; net H/B ratio: net harm/net benefit ratio; LDM: likelihood to be diagnosed or misdiagnosed. *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, that is, the test result was negatively associated with the true diagnosis.
The H/B may be defined as the net 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 net harm of a false negative result. The H/B ratio may be calculated from the Bayes’ equation as the product of the pretest odds of disease and the positive likelihood ratio at the specified test cutoff (which is equivalent to the slope of the ROC curve, TPR/FPR, at that point) and hence is equivalent to the post-test odds. 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”; to emphasize this point, henceforward it will be referred to as “net H/B ratio.”

More recently, another metric attempting to denote test limitation has been introduced: the LDM. 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, which is the inverse of Inacc, to the number needed to diagnose, which is the inverse of Youden index. Hence, LDM may also be conceptualized as a ratio of harms (misdiagnosis) and benefits (diagnosis) and hence of the “fragility” of screening and diagnostic tests. LDM ranges from -1 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>1 is desirable and LDM=∞ 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, net H/B ratio, and LDM) for several brief CSIs in common clinical usage for dementia diagnosis, namely the Mini-Mental State Examination (MMSE), the Montreal Cognitive Assessment (MoCA), the Mini-Addenbrooke’s Cognitive Examination (MACE), the Six-item Cognitive Impairment Test (6CIT), Free-Cog.

**METHODS**

**Participants**

Data from previously undertaken and reported pragmatic prospective test accuracy studies in consecutive patient cohorts from a single clinic were reanalyzed (Table 2). In all studies, subjects had given informed consent and study protocol was approved by the institute’s committee on human research.

**Procedures**

The studies examined seven CSIs which were in routine use in a dedicated cognitive disorders clinic at different times: MMSE, MoCA, MACE, 6CIT, iAD8, TYM, and Free-Cog. Each of these base studies was undertaken using a standardized methodology in the cognitive disorders clinic which was located in a regional neuroscience center. Criterion diagnosis of dementia followed standard diagnostic criteria (DSM-IV) and was made independent of scores on CSIs to avoid review bias. Cross classification of criterion diagnosis with CSI test result, dichotomized by test cutoff, in a standard 2×2 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 cutoffs documented in the respective index studies for each instrument were used to avoid bias.

| Table 2. Study demographics. |
|-----------------------------|
| CSI | n | Age, median (years) | Gender (F:M; %F) | Pre-test odds of dementia |
|-----|---|-------------------|-----------------|---------------------------|
| MMSE | 244 | 60 | 117:127; 48 | 0.22 |
| MoCA | 260 | 59 | 118:142; 45 | 0.21 |
| MACE | 755 | 60 | 352:403; 47 | 0.18 |
| 6CIT | 245 | 59 | 121:124; 49 | 0.25 |
| iAD8 | 212 | 64.5 | 106:106; 50 | 0.43 |
| TYM | 224 | 62 | 94:130; 42 | 0.54 |
| Free-Cog | 141 | 62 | 61:80; 43 | 0.12 |

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; iAD8: Ascertain Dementia 8; TYM: Test Your Memory test.
Statistical analysis
All studies followed either the STAndards for the Reporting of Diagnostic accuracy studies (STARD)26 or the derived guidelines specific for dementia studies, STARDdem,7 dependent on the exact date at which each test accuracy study was undertaken. Standard summary measures of test discrimination were calculated, namely, sensitivity and specificity, and positive and negative likelihood ratios (LR+, LR−; classified according to Jaeschke et al.).1 In addition, summary measures of test limitation were calculated, namely, MR,5 net H/B ratio,7 and LDM.8,9

RESULTS
Examining measures of test discrimination (Table 3), many were highly sensitive (MoCA, Free-Cog, MACE, and AD8) but had low specificity (MoCA, MACE, and AD8). Positive likelihood ratios were qualitatively either slight (MoCA, MACE, AD8, and TYM) or moderate (6CIT, Free-Cog, and MMSE), none achieving the large or very large classification.

Examining measures of test limitation (Table 3), few achieved a MR of ≤0.5 (Free-Cog, 6CIT, and MMSE). Only one test (6CIT) achieved net H/B ratio of 1. LDM values of <1 (likelihood of misdiagnosis greater than correct diagnosis) were recorded for some tests (MoCA, MACE, and AD8). Of note, the tests with high sensitivity but low specificity generally fared worse on these metrics examining test limitation, while 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).

Table 3. Comparing metrics of test discrimination and test limitation for CSIs for diagnosis of dementia versus no dementia.

| CSI (Cutoff) | Measures of test discrimination | Measures of test limitation |
|-------------|---------------------------------|-----------------------------|
|             | Sens   | Spec   | LR+           | MR   | Net H/B ratio | LDM |
| MMSE (<26/30) | 0.86   | 0.64   | 2.37=moderate | 0.50 | 0.52          | 1.56 |
| MoCA (<26/30) | 1.00   | 0.31   | 1.46=slight   | 0.69 | 0.30          | 0.54 |
| MACE (<25/30) | 0.99   | 0.32   | 1.45=slight   | 0.69 | 0.26          | 0.53 |
| 6CIT (>9/28)  | 0.88   | 0.78   | 4.00=moderate | 0.34 | 1.00          | 3.30 |
| AD8 (>2/8)    | 0.98   | 0.10   | 1.09=slight   | 0.92 | 0.47          | 0.12 |
| TYM (<42/50)  | 0.95   | 0.45   | 1.73=slight   | 0.60 | 0.93          | 1.08 |
| Free-Cog (<22/30) | 1.00 | 0.67 | 3.07=moderate | 0.33 | 0.38 | 2.31 |

CSIs: 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 test; 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.

Table 4. Ranking of cognitive screening instruments by outcome measures of test discrimination and test limitation (1=best, 7=worst).

| Rank | Measures of test discrimination | Measures of test limitation |
|------|---------------------------------|-----------------------------|
|      | Sens   | Spec   | LR+ | MR | Net H/B ratio | 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 | AD8 | TYM |
| 5    | TYM    | MACE   | MoCA | MACE, MoCA | Free-Cog | MoCA |
| 6    | 6CIT   | MoCA   | MACE | – | MoCA | MACE |
| 7    | MMSE   | AD8    | AD8 | AD8 | MACE | AD8 |

CSIs: 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 test; 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.
DISCUSSION

The metrics examined here explicitly acknowledge test shortcomings, hence their designation as measures of test limitation in distinction from measures of test discrimination. Although limitation may be implicit in the latter (e.g., FNR in Sens, FPR in Spec), this inherent quality may not be apparent on a cursory examination. Moreover, some test metrics choose the best quality of a test and largely ignore its weaknesses (e.g., diagnostic odds ratio, area under the ROC curve) giving the most optimistic results. 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 utilities’ or cost ratios.27

This study has various shortcomings. The findings are of course dependent upon the diagnostic test accuracy studies upon which they are based.18-25 These base studies are of course dependent upon the diagnostic test accuracy but their performance.28 Nevertheless the findings suggest 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 utilities’ or cost ratios.27

In summary, CSIs have shortcomings which may be corroborated by undertaking similar analyses with data reported in systematic reviews of these CSIs where available.

For MR and the net H/B ratio, lower or higher values, respectively, may be better, but precisely how high or 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, but as it is based 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. Lower 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. 1994;271(9):703-7. https://doi.org/10.1001/jama.1994.03500930189
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 STARIDdem Initiative. Neurology. 2014;83(4):364-73. https://doi.org/10.1212/WNL.0000000000000621
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; 2018.
4. Davis DH, Creavin ST, Noel-Storr A, Quinn TJ, Smalagic N, Hyde CH, et al. Neuropsychological tests for the diagnosis of Alzheimer’s dementia and other dementias: a generic protocol for cross-sectional and delayed-verification studies. Cochrane Database Syst Rev. 2013;3:CD010460. https://doi.org/10.1002/14651858.CD010460
5. Perkins NJ, Schisterman EF. The inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006;163(7):670-5. https://doi.org/10.1093/aje/kwj063
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). 2016;26(3):297-307. https://doi.org/10.11613/BM.2016.024
7. Hunink MGM, Weinstein MC, Wittenberg E, Drummond MF, Pilsin J, Wong JB, et al. Decision making in health and medicine. Integrating evidence and values. 2nd ed. Cambridge: Cambridge University Press; 2014. https://doi.org/10.1017/CBO978113906779.004
8. Lamer AJ. Number needed to diagnose, predict, or misdiagnose: useful metrics for non-canonical signs of cognitive status? Dement Geriatr Cogn Dis Extra. 2018;8(3):321-7. https://doi.org/10.1159/000492783
9. Lamer AJ. Evaluating cognitive screening instruments with the “likelihood to be diagnosed or misdiagnosed” measure. Int J Clin Pract. 2019;73(2):e13265. https://doi.org/10.1111/ijcp.13265
10. Habibzadeh F, Yadollahie M. Number needed to misdiagnose: a measure of diagnostic test effectiveness. Epidemiology. 2013;24(1):170. https://doi.org/10.1097/EDE.0b013e3182182792
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. 1975;12(3):189-98. https://doi.org/10.1016/0022-3956(75)90026-6
12. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Colín I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005;53(4):686-9. https://doi.org/10.1111/j.1532-5415.2005.53221.x

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. 2015;39(1-2):1-11. https://doi.org/10.1159/000386604

14. Brooke P, Bullock R. Validation of a 6 item cognitive impairment test with a view to primary care usage. Int J Geriatr Psychiatry. 1999;14(1):928-40. https://doi.org/10.1002/(sici)1099-1166(199901)14:1<928::aid-gps39b>3.0.co;2-1

15. Galvin JE, Roe CM, Pwllhwaith KC, Coats MA, Much SJ, Grant E, et al. The AD8. A brief informant interview to detect dementia. Neurology. 2005;65(4):559-64. https://doi.org/10.1212/01.wnl.0000172958.95282.2a

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. 2009;338:b2030. https://doi.org/10.1136/bmj.b2030

17. Burns A, Harrison JR, Symonds C, Morris J. A novel hybrid scale for the assessment of cognitive and executive function: the Free-Cog. Int J Geriatr Psychiatry. 2021;36(4):566-72. https://doi.org/10.1002/gps.5454

18. Larner AJ. Mini-Addenbrooke’s Cognitive Examination: a pragmatic diagnostic accuracy study. Int J Geriatr Psychiatry. 2015;30(5):547-8. https://doi.org/10.1002/gps.4258

19. Larner AJ. Mini-Addenbrooke’s Cognitive Examination diagnostic accuracy for dementia: reproducibility study. Int J Geriatr Psychiatry. 2015;30(10):1103-4. https://doi.org/10.1002/gps.4234

20. Larner AJ. MACE versus MoCA: equivalence or superiority? Pragmatic diagnostic test accuracy study. Int Psychogeriatr. 2017;29(6):931-7. https://doi.org/10.1017/S10416120176002210

21. Larner AJ. MACE for diagnosis of dementia and MCI: examining cut-offs and predictive values. Diagnostics (Basel). 2019;9(2):51. https://doi.org/10.3390/diagnostics9020051

22. Abdel-Aziz K, Larner AJ. Six-item Cognitive Impairment Test (6CIT); pragmatic diagnostic accuracy study for dementia and MCI. Int Psychogeriatr. 2015;27(6):991-7. https://doi.org/10.1017/S1041610214002932

23. Larner AJ. AD8 informant questionnaire for cognitive impairment: pragmatic diagnostic test accuracy study. J Geriatr Psychiatry Neurol. 2015;28(3):198-202. https://doi.org/10.1177/0891988715573536

24. Hancock P, Larner AJ. Test Your Memory (TYM) test: diagnostic utility in a memory clinic population. Int J Geriatr Psychiatry. 2011;26(9):976-80. https://doi.org/10.1002/gps.2639

25. Larner AJ. Free-Cog: pragmatic test accuracy study and comparison with Mini-Addenbrooke’s Cognitive Examination, Dement Geriatr Cogn Disord. 2019;47(4-6):254-63. https://doi.org/10.1159/000500069

26. 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. 2003;49(1):7-18. https://doi.org/10.1373/49.1.7

27. Kraemer HC. Evaluating medical tests. Objective and quantitative guidelines. Newbury Park, CA: Sage; 1992.

28. Brucki SMD, Nitrini R, Caramelli P, Bertolucci PHF, Okamoto IH. Suggestions for utilization of the Mini-Mental State Examination in Brazil. Arq Neuro-Psiquiatr. 2003;61(3B):777-81. https://doi.org/10.1590/s0004-282x2003000500014

29. Brenner H, Geleller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997;16(9):981-91. https://doi.org/10.1002/(sici)1097-0258(19970515)16:9<981::aid-sim10>3.0.co;2-n