EXTRA Dementia and Geriatric Cognitive Disorders

Dement Geriatr Cogn Disord Extra 2014;4:236–241
DOI: 10.1159/000363735
Accepted after revision: May 17, 2014
Published online: July 3, 2014

1664–5464/14/0042–0236$39.50/0
www.karger.com/dee

Karger Open access

Original Research Article

Effect Size (Cohen's d) of Cognitive Screening Instruments Examined in Pragmatic Diagnostic Accuracy Studies

Andrew J. Larner

Cognitive Function Clinic, Walton Centre for Neurology and Neurosurgery, Liverpool, UK

Key Words

Diagnostic accuracy · Effect size · Cohen's d · Mini-Mental State Examination · Mini-Mental Parkinson · Six-Item Cognitive Impairment Test · Montreal Cognitive Assessment · Test Your Memory test · Addenbrooke's Cognitive Examination-Revised

Abstract

Background/Aims: Many cognitive screening instruments (CSI) are available to clinicians to assess cognitive function. The optimal method comparing the diagnostic utility of such tests is uncertain. The effect size (Cohen's d), calculated as the difference of the means of two groups divided by the weighted pooled standard deviations of these groups, may permit such comparisons. ***Methods:*** Datasets from five pragmatic diagnostic accuracy studies, which examined the Mini-Mental State Examination (MMSE), the Mini-Mental Parkinson (MMP), the Six-Item Cognitive Impairment Test (6CIT), the Montreal Cognitive Assessment (MoCA), the Test Your Memory test (TYM), and the Addenbrooke's Cognitive Examination-Revised (ACE-R), were analysed to calculate the effect size (Cohen's d) for the diagnosis of dementia versus no dementia and for the diagnosis of mild cognitive impairment versus no dementia (subjective memory impairment). ***Results:*** The effect sizes for dementia versus no dementia diagnosis were large for all six CSI examined (range 1.59–1.87). For the diagnosis of mild cognitive impairment versus no dementia, the effect sizes ranged from medium to large (range 0.48–1.45), with MoCA having the largest effect size. ***Conclusion:*** The calculation of the effect size (Cohen's d) in diagnostic accuracy studies is straightforward. The routine incorporation of effect size calculations into diagnostic accuracy studies merits consideration in order to facilitate the comparison of the relative value of CSI.

Andrew J. Larner
Cognitive Function Clinic
Walton Centre for Neurology and Neurosurgery
Lower Lane, Fazakerley, Liverpool L9 7LJ (UK)
E-Mail a.larner@thewaltoncentre.nhs.uk

Dement Geriatr Cogn Disord Extra 2014;4:236–241
DOI: 10.1159/000363735

Introduction

The utility of cognitive screening instruments (CSI) for the diagnosis of dementia and lesser degrees of cognitive impairment may be indicated by a number of summary parameters, of which the most familiar are probably sensitivity and specificity. Predictive values, likelihood ratios, clinical utility indexes, agreement between tests (kappa statistic), and the area under the receiver operating characteristic curve (AUC ROC) may also be used as parameters of diagnostic utility [1], but all these measures have potential shortcomings. For example, sensitivity and specificity may be difficult to apply to individual patients, predictive values are influenced by the prevalence of the disease in the population being tested, and AUC ROC combines the test accuracy over a range of thresholds, which may be both clinically relevant and clinically nonsensical.

Another metric that may be used to demonstrate utility is the effect size. The effect size may be denoted by a variety of summary indices, of which Cohen's d is probably the most commonly used in the medical literature [2]. This parameter is calculated as the difference of the means of two groups divided by the weighted pooled standard deviations of these groups (fig. 1). Cohen [3] suggested that effect sizes of 0.2–0.3 were small, 0.5 medium, and ≥0.8 large.

One example of the potential utility of this approach in clinical studies of cognitive impairment was demonstrated by Brønnick [4] who compared standardized effect sizes of cognitive functions between groups of patients diagnosed with Parkinson's disease, Parkinson's disease dementia, Alzheimer's disease, and normal controls to identify differences in group mean values (Cohen's d). This study showed larger effect sizes for tests of memory in Alzheimer's disease and of executive and visuospatial function in Parkinson's disease dementia, indicating greater impairments in these domains. Hence, testing of these selected cognitive functions may be of particular utility for the differential diagnosis of these conditions.

The aim of the study presented here was to calculate the Cohen's d metric from the datasets of a number of pragmatic prospective diagnostic accuracy studies undertaken in dedicated secondary care memory clinics to calculate effect sizes for several CSI, specifically the Mini-Mental State Examination (MMSE) [5], the Mini-Mental Parkinson (MMP) [6], the Six-Item Cognitive Impairment Test (6CIT) [7], the Montreal Cognitive Assessment (MoCA) [8], the Test Your Memory (TYM) test [9], and the Addenbrooke's Cognitive Examination-Revised (ACE-R) [10]. The calculation of the effect size was undertaken for both the diagnosis of dementia versus no dementia and of mild cognitive impairment versus no dementia.

Materials and Methods

Data from five previous pragmatic diagnostic accuracy studies [11–15], which examined six different CSI [5–10], namely the MMSE [11], MMP [11], 6CIT [12], MoCA [13], TYM [14], and ACE-R [15], were reanalysed. Study details (setting, sample size, dementia prevalence, sex ratio, and age range) are shown in table 1.

In each of these studies, the criterion diagnosis was established by the judgment of an experienced clinician based on widely accepted clinical diagnostic criteria. The mean test scores for demented and non-demented groups as well as for mild cognitive impairment and non-demented groups, along with their standard deviations, were applied to the Cohen's d formula (fig. 1) to calculate effect sizes. Because these were clinic-based pragmatic studies, there was no normal control group, the non-demented cases consisting of patients with at least subjective memory impairment as well as patients with mild cognitive impairment insufficient to mandate a dementia diagnosis.

Cohen's d formula:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^3 + s_2^3}{2}}}$$

Fig. 1. d = Cohen's d effect size; X_1 and X_2 = means of the two groups; s_1 and s_2 = standard deviations of the two groups.

Table 1. Study demographics

CSI	Setting	n	Dementia prevalence, %	M:F (% male)	Age range, years	Ref.
MMSE, MMP	Cognitive function clinic	225	21	130:95 (58)	20–86 (median 62)	11
6CIT	Cognitive function clinic	186	19	96:90 (52)	16–94 (median 59)	12
MoCA	Cognitive function clinic	150	24	93:57 (62)	20–87 (median 61)	13
TYM	Cognitive function clinic and old age psychiatry memory clinic	224	35	130:94 (58)	20–90 (mean 63)	14
ACE-R	Old age psychiatry memory clinic	183	40	105:78 (57)	37–89 (median 67)	15

Table 2. Cohen's d effect size for dementia versus no dementia and for mild cognitive impairment versus no dementia

CSI	Cohen's d: dementia vs. no dementia	Cohen's d: mild cognitive impairment vs. no dementia
MMSE	1.59 (large)	0.69 (medium)
MMP	1.78 (large)	0.81 (large)
6CIT	1.84 (large)	0.65 (medium)
MoCA	1.80 (large)	1.45 (large)
TYM	1.62 (large)	0.48 (medium)
ACE-R	1.87 (large)	0.73 (medium)

Results

The calculation of Cohen's d comparing patients with and without dementia suggested large but similar effect sizes for all CSI examined (table 2, left-hand column), using the classification suggested by Cohen [2, 3]. These values suggested a consistent difference in test scores between demented and non-demented individuals.

The calculation of Cohen's d comparing patients with mild cognitive impairment and no dementia (subjective memory impairment) suggested smaller effect sizes for all CSI examined than in the dementia versus no dementia distinction (table 2, right-hand column), using the classification suggested by Cohen [2, 3]. However, one effect size, namely that for the MoCA, was clearly larger than all the others. These values suggested a consistent difference in test scores between MCI and non-demented individuals, but with the MoCA performing best.

For comparative purposes, other parameters of diagnostic utility derived from the sampled diagnostic accuracy studies [11–15] are shown, namely sensitivity, specificity, predictive values, likelihood ratios, and AUC ROC (table 3).

Table 3. Other diagnostic utility measures for dementia versus no dementia

CSI	Overall accuracy	Sens	Spec	PPV	NPV	LR+	LR–	AUC ROC	Ref.
MMSE	0.86	0.45	0.98	0.88	0.85	22.9	0.56	0.87	11
MMP	0.86	0.51	0.97	0.83	0.87	15.7	0.51	0.89	11
6CIT (n = 100)	0.79	0.80	0.79	0.48	0.94	3.8	0.25	0.88	12
MoCA	0.81	0.63	0.95	0.91	0.77	13.4	0.39	0.91	13
TYM	0.83	0.73	0.88	0.77	0.86	6.3	0.30	0.89	14
ACE-R	0.87	0.82	0.89	0.75	0.93	7.7	0.20	N/A	15

Sens = Sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; LR+ = positive likelihood ratio; LR– = negative likelihood ratio; N/A = not available.

Discussion

In this study, data from five pragmatic diagnostic accuracy studies of CSI [11–15] were re-analysed to calculate effect sizes (Cohen's d) for the diagnosis of dementia versus no dementia and for the diagnosis of mild cognitive impairment versus no dementia. These were observational studies, which examined unselected patient groups with cognitive complaints of unknown aetiology. Such pragmatic diagnostic accuracy studies [16] differ from experimental studies in which patient groups are selected by known diagnostic categories, often with a normal control group, and then have the test or intervention applied (essentially case-control studies). Since pragmatic diagnostic accuracy studies reflect the idioms of clinical practice in terms of the typical spectrum of patients seen more closely, it may be argued that their results are more broadly generalizable [1], although the groupings are more heterogeneous than in experimental studies.

Making comparisons between studies is, of course, problematic, notwithstanding the consistency of study protocols and authorship in the studies examined here. One potential shortcoming of the current analysis was the different settings and sample characteristics for each of the five studies (table 1). Inevitably, the case-mix in clinics run by neurologists and by old age psychiatrists will differ in terms of both patient age and dementia prevalence. One outcome of this heterogeneity was insufficient data to compare effect sizes by patient age.

Overall, the calculations suggested a large effect size for all CSI examined for the diagnosis of dementia versus no dementia (table 2, left-hand column). Since the purpose of these instruments is to screen for the diagnosis of dementia, this finding is perhaps not surprising.

Effect size calculations for the differential diagnosis of mild cognitive impairment versus no dementia (subjective memory impairment) gave lower values for all CSI (table 2, right-hand column), as might be anticipated for this more challenging clinical distinction. These calculations suggested that the MoCA performed best, as might be expected since this instrument was specifically designed to detect mild cognitive impairment [9]. This distinction between the performance of different CSI may be of clinical importance when disease-modifying drugs for cognitive impairment are developed.

How do effect sizes relate to other diagnostic utility measures (table 3)? As a global measure, encompassed in a single outcome number compared to criterion values [3], effect sizes are perhaps akin to AUC ROC, which likewise differed little between the various CSI examined in this study (table 3). The effect size thus gives an overall index of test diagnostic utility, but obviously this metric may conceal differences between tests in terms of, for example, their relative sensitivity (range 0.45–0.82) and specificity (range 0.79–0.98),

although these parameters are obviously dependent on the chosen test cutoff. Thus, effect sizes may need to be used in conjunction with other summary test metrics to decide which might be most appropriate for the given clinical situation, for example, whether the clinician seeks high sensitivity or specificity.

The calculated effect sizes for the CSI examined in this study permit some comparison between instruments, although other summary measures and analyses are also available for this purpose. Meta-analysis is perhaps the most favoured approach for its methodological rigor, but the outcomes are dependent on the stringency of study inclusion and exclusion criteria, which may sometimes give results that are not anticipated (e.g. comparability of the MMSE and ACE/ACE-R, due to a surprisingly high sensitivity of the MMSE in included studies [17]). The test of agreement between tests (kappa statistic) can be used to measure whether agreement between tests is perfect (kappa = 1) or due to chance alone (kappa = 0) [18]. The AUC ROC is a measure of diagnostic accuracy but has been criticised for combining test accuracy over a range of thresholds, which may be both clinically relevant and clinically nonsensical [19]. Weighted comparison may be used to indicate the net benefit of one test with respect to another and permits the calculation of the equivalent increase in the number of true positive patients identified per 1,000 patients tested [19]. One such weighted comparison study suggested the superiority of the ACE-R and MoCA over the MMSE and the inferiority of TYM to MMSE [20].

As shown in this study, the calculation of effect size (Cohen's d) for CSI is straightforward. No previous analyses of the comparative diagnostic utility of CSI using this method have been identified. This study suggests that there is a case for the routine incorporation of effect sizes (it should be noted that there are effect size formulae other than Cohen's d [21]) as a measure of diagnostic test performance into diagnostic accuracy studies. Although this is not an explicit recommendation of the STARDdem guidelines (Standards for the Reporting of Diagnostic Accuracy studies specific to diagnostic accuracy studies in dementia; www.starddem.org), summary measures such as Cohen's d and weighted comparison might be considered in future iterations of these guidelines.

Disclosure Statement

The author has no conflicts of interest to declare.

References

1 Larner AJ: Dementia in Clinical Practice: A Neurological Perspective. Pragmatic Studies in the Cognitive Function Clinic, ed 2. London, Springer, 2014, pp 21–40.
2 Cohen J: Statistical Power Analysis for the Behavioral Sciences, ed 2. Hillsdale, Lawrence Erlbaum, 1988.
3 Cohen J: A power primer. Psychol Bull 1992;112:155–159.
4 Brønnick K: Cognitive profile in Parkinson's disease dementia; in Emre M (ed): Cognitive Impairment and Dementia in Parkinson's Disease. Oxford, Oxford University Press, 2010, pp 27–43.
5 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.
6 Mahieux F, Michelet D, Manifacier M-J, Boller F, Fermanian J, Guillard A: Mini-Mental Parkinson: first validation study of a new bedside test constructed for Parkinson's disease. Behav Neurol 1995;8:15–22.
7 Brooke P, Bullock R: Validation of a 6 item cognitive impairment test with a view to primary care usage. Int J Geriatr Psychiatry 1999;14:936–940.
8 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.
9 Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H: The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53:695–699.

10 Mioshi E, Dawson K, Mitchell J, Arnold R, Hodges JR: The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. Int J Geriatr Psychiatry 2006;21:1078–1085.
11 Larner AJ: Mini-mental Parkinson (MMP) as a dementia screening test: comparison with the Mini-Mental State Examination (MMSE). Curr Aging Sci 2012;5:136–139.
12 Abdel-Aziz K, Larner AJ: Six-Item Cognitive Impairment Test (6CIT) for detection of dementia and cognitive impairment. J Neurol Neurosurg Psychiatry, in press.
13 Larner AJ: Screening utility of the Montreal Cognitive Assessment (MoCA): in place of – or as well as – the MMSE? Int Psychogeriatr 2012;24:391–396.
14 Hancock P, Larner AJ: Test Your Memory test: diagnostic utility in a memory clinic population. Int J Geriatr Psychiatry 2011;26:976–980.
15 Larner AJ, Hancock P: ACE-R or MMSE? A weighted comparison. Int J Geriatr Psychiatry 2014;29:767–768.
16 Larner AJ: Pragmatic diagnostic accuracy studies. www.bmj.com/contents/345/bmj.e3999/rr/599970 (accessed March 23, 2014).
17 Larner AJ, Mitchell AJ: A meta-analysis of the accuracy of the Addenbrooke's Cognitive Examination (ACE) and the Addenbrooke's Cognitive Examination-Revised (ACE-R) in the detection of dementia. Int Psychogeriatr 2014;26:555–563.
18 Cohen J: A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
19 Mallett S, Halligan S, Thompson M, Collins GS, Altman DG: Interpreting diagnostic accuracy studies for patient care. BMJ 2012;344:e3999.
20 Larner AJ: Comparing diagnostic accuracy of cognitive screening instruments: a weighted comparison approach. Dement Geriatr Cogn Disord Extra 2013;3:60–65.
21 Ellis PD: The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge, Cambridge University Press, 2010.