ELSEVIER

Alzheimer's & Dementia 9 (2013) S4–S9

Alzheimer's & Dementia

Featured Articles

Putting the Alzheimer's cognitive test to the test I: Traditional psychometric methods

Jeremy Hobart[a,*,†], Stefan Cano[a,†], Holly Posner[b], Ola Selnes[c], Yaakov Stern[d], Ronald Thomas[e], John Zajicek[a]; for the Alzheimer's Disease Neuroimaging Initiative[‡]

[a]*Clinical Neurology Research Group, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK*
[b]*Clinical R&D, Pfizer, Inc., New York, NY, USA*
[c]*Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[d]*Cognitive Neuroscience Division, Columbia University, New York, NY, USA*
[e]*Department of Family and Preventative Medicine, University of California, San Diego, CA, USA*

Abstract

Background: The Alzheimer's Disease Assessment Scale—Cognitive Behavior section (ADAS-Cog) is the most commonly used cognitive test in AD clinical trials. However, there are concerns about its use in early-stage disease. Herein we examine those concerns using traditional psychometric methods.
Methods: We analyzed ADAS-Cog data (n = 675) based on six psychometric properties: data completeness; scaling assumptions; targeting; reliability; validity; and responsiveness.
Results: At the *scale-level*, criteria tested for data completeness, scaling assumptions (item total correlations 0.33–0.59), targeting (no floor/ceiling effects), reliability (Cronbach's $\alpha = 0.74$), and validity (correlation with MMSE = −0.70) were satisfied. Responsiveness (baseline to 12 months; n = 145) was moderate to high (effect size = −0.73). However, 8 of 11 ADAS-Cog *components* had substantial ceiling effects (range 32%–83%), and decreased responsiveness associated with low to moderate effect sizes (0.14–0.65).
Conclusion: In our study, many patients with AD found large portions of the ADAS-Cog too easy. Future research should consider modifying the ADAS-Cog or developing a new test.

Keywords: Alzheimer's disease; Clinical trials; Psychometrics; Reliability; Validity

1. Introduction

Alzheimer's disease (AD) is an incurable progressive neurodegenerative disease that impacts primarily cognition [1]. It is the most common dementia, affecting approximately 27 million people worldwide [2,3]. Incidence rates are expected to quadruple by 2050 [2]. Considerable resources have been targeted at slowing AD progression, and the number of clinical trials is increasing [1,4].

The most widely used primary outcome test has been the AD Assessment Scale—Cognitive Behavior Section (ADAS-Cog) [5]. It was published in 1984 specifically for clinical trials of people with dementia of the AD type and has been used in approximately 170 trials. However, although the ADAS-Cog has had a critical role as a primary outcome measure in numerous clinical trials, its suitability to the changing face of AD studies needs review.

Concerns have been raised regarding the ability of the ADAS-Cog to detect change in the mildest stages of AD [6]. In particular, when examined closely, the 11 components that make up the total score of the ADAS-Cog have been found to have significant ceiling effects, which reduces their ability to measure changes and differences in higher functioning patients [7]. However, most traditional psychometric evaluations of the ADAS-Cog have been incomplete [8,9], and the most

The authors have no conflicts of interest to report.

[†]J.H. and S.C. contributed equally to this article and share first author status.[‡]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/-ADNI_Acknowledgement_List.pdf.

*Corresponding author. Tel.: +44(0)1752-315272; Fax: +44(0)1752-315254.

E-mail address: Jeremy.Hobart@pms.ac.uk

1552-5260/$ - see front matter © 2013 The Alzheimer's Association. All rights reserved.
http://dx.doi.org/10.1016/j.jalz.2012.08.005

comprehensive did not examine responsiveness. In our previous psychometric evaluation of the ADAS-Cog we focused on a sample of patients with AD recruited for an industry-funded, randomized, controlled clinical trial [7]. However, the extent to which samples of participants recruited for such trials differ from those in observational studies (e.g., the ADNI, which may be considered more "real world") is an empirical question. As such, we considered it important to examine the reproducibility of our previous findings, especially given the widespread use of the ADAS-Cog. Thus, the aim of this study was to provide clinicians and researchers with a thorough traditional psychometric evaluation, including responsiveness, in general-population, non-industry ADAS-Cog data.

2. Methods

2.1. Setting and participants

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of the ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biologic markers, and clinical and neuropsychologic assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to help researchers and clinicians develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The principal investigator of this initiative is Michael W. Weiner, MD (VA Medical Center and University of California, San Francisco). The ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the USA and Canada. The initial goal of the ADNI was to recruit 800 adults, 55–90 years of age, to participate in the research, including approximately 200 cognitively normal older individuals to be followed for 3 years, 400 with MCI to be followed for 3 years, and 200 with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org. For the present study, anonymized, longitudinal ADAS-Cog data on AD subjects from the ADNI central database were made available for data analysis.

In this study we included only the subsample of ADNI subjects identified as having mild AD (n = 193 at baseline), diagnosed in those having MMSE scores of between 20 and 26 (inclusive), a Clinical Dementia Rating (CDR) score of 0.5 or 1.0, and meeting NINCDS/ADRDA criteria for probable AD. The ADNI data set was downloaded on April 9, 2008. All analyses were conducted on the entire data set, including all data points from the mild AD subsample (i.e., 675 measurements across all time-points), except the responsiveness analysis that included n = 191 people (baseline compared with 12 months).

2.2. Procedures: ADAS-Cog content, scoring, and analysis

The ADAS-Cog includes 11 cognitive components[1], which are summed into a total score. Low scores indicate better cognitive performance. We used traditional psychometric methods to test the ADAS-Cog, as these are the most widely used [10]. We examined six properties (Table 1, more fully explained elsewhere [7]): data availability; scaling assumptions; targeting; reliability; validity; and responsiveness. Analyses were conducted separately for the total scale score and each of the 11 components. We used SPSS (version 19.0) software to conduct the analyses.

3. Results

3.1. Sample

At the time we accessed the ADNI data set there were a total of 675 measurements from 193 patients with AD at four time-points: 0, 6, 12, and 24 months (mean age 74 [SD 8, range 53–80] years, 47% female, 94% white). The mean Mini-Mental State Examination (MMSE) score for the group was 23 (SD 8) across all time-points.

3.2. Psychometric properties

3.2.1. Data completeness

Data completeness was high. The proportion of component-level missing data was low (≤4%). ADAS-Cog total scores could be computed for 96% of the sample (Table 2).

3.2.2. Scaling assumptions

The ADAS-Cog satisfied most criteria for scaling assumptions. Component (item)-total correlations for the 11 ADAS-Cog components ranged from 0.33 to 0.59, satisfying the recommended criteria (Table 2).

3.2.3. Targeting

ADAS-Cog total scores spanned approximately 87% of the entire scale range (6–57), with no floor (score = 70) or ceiling (score = 0) effects. Their distribution was only slightly skewed (skewness = 1.2) (Table 2).

Eight of 11 components (all except word recall, word recognition, and orientation) had significant floor/ceiling

[1] In fact, each component is a cognitive test in its own right. We use the term "component" or "item" here to distinguish these tests from the total ADAS-Cog "test" score.

Table 1
Brief definitions of psychometric properties*

Psychometric property	Definition/criteria for acceptability
Data completeness	The extent to which ADAS-Cog components and total scores can be computed. This is assessed by percent of missing data for each component, and the percent of subjects for whom a scale score can be computed. The criterion for item-level missing data was <10%, with computable scale scores >50% completed components.
Scaling assumptions	The extent to which it is legitimate to sum a set of component scores, without weighting or standardization, to produce a single total score. Summing ADAS-Cog component scores is considered legitimate when the components: (1) are measured at the same point on the scale (criterion—components have similar mean scores); (2) contribute similarly to the variation of the total score (criterion—components have similar standard deviations); (3) measure a common underlying construct, here cognitive performance (criterion—components have adequate corrected item–total correlation [ITC] ≥0.30); and (4) contain a similar proportion of information with regard to the construct being measured (criterion—components have similar ITCs).
Targeting	The extent to which the range of cognitive performance measured by the scale matches the range of that cognitive ability in the study sample. Scale scores should span the entire range; floor (proportion of the sample at the maximum scale score for the ADAS-Cog) and ceiling (proportion of the sample at the minimum scale score) effects should be low (<15%); and skewness statistics should range from −1 to +1.
Reliability	The extent to which scale scores are not associated with random error. The precision of the scale is based on the homogeneity (intercorrelations) of items at a single point in time. Assessed using criterion - Cronbach's α >0.70 (but minimum desired >0.80), mean item–item correlations (known as the homogeneity coefficient) ≥0.30, and item–total correlations ≥0.30.
Validity	The extent to which a scale measures what it intends to measure. This is essential for the accurate and meaningful interpretation of scores. Two aspects of construct validity were tested: (1) convergent construct validity was examined by computing correlations (Pearson's *r*) between ADAS-Cog and the Mini-Mental State Examination (MMSE†) (criterion—correlation >0.70 due to similarity in constructs between ADAS-Cog and MMSE); and (2) discriminant construct validity was examined by computing correlations between the ADAS-Cog and sociodemographic variables (age and gender; these were selected as we would not expect these variable to significantly influence measurement performance; criterion—correlations <0.30).
Responsiveness	Examined at the group level by comparing baseline and 12-month scores using two effect-size calculations (Kazis effect size [ES] and standardized response mean [SRM]). ES/SRMs are interpreted as follows: 0.20 (small change); 0.50 (moderate change); or ≥0.80 (large change).

*Adapted from Cano et al [9]; see this article for more information and related psychometric criteria references.
†The MMSE is a 30-item rating scale used to assess aspects of cognitive performance (including arithmetic, memory, and orientation), and is commonly used in screening for dementia. It also classifies AD as mild (MMSE 21–26), moderate (MMSE 15–20), or severe (MMSE 10–14).

effects (32%–83%) and their distributions were notably skewed (+0.9 to +3.2). These findings indicate poor component-to-sample targeting. They imply that the range of cognitive performance measured by these eight components is considerably mismatched to the ranges of cognitive performance in this sample.

3.2.4. Reliability

Cronbach's α for the ADAS-Cog scale was acceptable (0.74; CI 0.72–0.76). This can be viewed as supporting the scale's reliability as it exceeded some recommended criteria of 0.70, but it did not exceed the desired minimum (0.80).

3.2.5. Validity

The correlation between the ADAS-Cog and MMSE was near our prediction (−0.70). Correlations between the ADAS-Cog at baseline and sociodemographic variables were very low (age = −0.04, gender = 0.03), implying that ADAS-Cog scores were not biased by these variables. Together, these findings provide evidence for the ADAS-Cog's convergent and discriminant construct validity.

3.2.6. Responsiveness

The mean change in ADAS-Cog total scores measured between baseline and 12 months was −4.3 points (SD 6.4; $P < .000$). Although this is only 6% of the available scale range it represents a moderate to large effect size (−0.73) and standardized response mean (−0.66).

The mean change in component scores measured between baseline and 12 months ranged from −0.1 to −1.1 (SD range 0.6–3.2; significance range $P < .000$ to $P < .2$). Component effect sizes ranged from −0.14 (constructional praxis) to −0.65 (orientation), with standardized response means of −0.11 and −0.63, respectively. These values reflected low to moderate effect size statistics.

4. Discussion

The major finding of this ADAS-Cog analysis in patients with AD-type dementia was that, despite adequately performing at the *scale-level*, at the *component-level*, three quarters of the ADAS-Cog's components had limited response distributions. This supports our previous evaluation in the pharmaceutical company clinical trial data [7], and means these components may underestimate cognitive performance differences in those with mild to moderate AD-type dementia. This may lead to problems in detecting clinical change.

The key issue is that caution is required for the apparent adequate responsiveness of the ADAS-Cog total score over 12 months. This is because traditional responsiveness

Table 2
ADAS-Cog scale and component-level analyses (n = 675)

Psychometric property	Word recall	Commands	Constructional praxis	Naming objects and fingers	Ideational praxis	Orientation	Word recognition	Remembering test instructions	Comprehension	Word finding	Spoken language	ADAS-Cog total
Data completeness												
MD (%)*	2	2	2	2	3	2	4	4	3	3	3	
Computable scale scores (%)	–	–	–	–	–	–	–	–	–	–	–	96
Scaling Assumptions												
Possible range	0–10	0–5	0–5	0–5	0–5	0–8	0–12	0–5	0–5	0–5	0–5	–
Range midpoint	5	2.5	2.5	2.5	2.5	4	6	2.5	2.5	2.5	2.5	–
Score range	2–10	0–4	0–4	0–5	0–5	0–8	0–12	0–5	0–5	0–5	0–4	–
Mean score	6.2	0.5	0.9	0.6	0.5	2.5	7.1	0.4	0.4	0.7	0.3	–
SD	1.5	0.8	0.8	0.8	0.9	1.9	3.0	1.0	0.8	1.0	0.7	–
Corrected ITC	0.59	0.41	0.33	0.39	0.49	0.50	0.39	0.52	0.55	0.38	0.41	0.33–0.59†
Targeting												
Possible range	0–10	0–5	0–5	0–5	0–5	0–8	0–12	0–5	0–5	0–5	0–5	0–70
Range midpoint	5	2.5	2.5	2.5	2.5	4	6	2.5	2.5	2.5	2.5	35
Score range	2–10	0–4	0–4	0–5	0–5	0–8	0–12	0–5	0–5	0–5	0–4	6–57
Mean score	6.2	0.5	0.9	0.6	0.5	2.5	7.1	0.4	0.4	0.7	0.3	20.0
SD	1.5	0.8	0.8	0.8	0.9	1.9	3.0	1.0	0.8	1.0	0.7	7.9
C/F effect (%)‡	0/2	66/0	32/0	58/0	66/1	16/0	0/9	83/0	73/0	57/0	79/0	0/0
Skewness	0.1	2.0	0.9	2.0	2.7	0.5	−0.1	3.1	2.2	1.4	2.4	1.2
Responsiveness												
Mean change score	−0.4	−0.2	−0.1	−0.2	−0.4	−1	−1.1	−0.3	−0.2	−0.3	−0.2	−4.3
SD	1.1	0.8	0.8	0.6	0.9	1.6	3.2	1.2	1	0.9	0.9	6.4
P-value	.000	.018	.192	.000	.000	.000	.000	.014	.022	.000	.008	.000
Effect size	−0.24	−0.27	−0.14	−0.29	−0.5	−0.65	−0.41	−0.33	−0.28	−0.32	−0.32	−0.73
Standardized response mean	−0.32	−0.2	−0.11	−0.34	−0.41	−0.63	−0.35	−0.21	−0.19	−0.3	−0.22	−0.66

Abbreviation: MD, missing data; ITC, item total correlation.
*<0.5% MD rounded to 0.
†Range of ITC.
‡C/F = ceiling/floor.

statistics are difficult to interpret and may even be misleading [11]. In addition, our scrutiny of component-level data, rarely undertaken, revealed a seemingly simple but significant weakness: Many subjects (often >75%) scored either 0 or 1 on the majority of ADAS-Cog components. This implies the detection of few or no cognitive problems. However, as there is almost certainly greater variance in patient ability, this finding points to a limitation in the ADAS-Cog score function—namely that the ADAS-Cog, in its current form, is not subtle enough to record and monitor variance in the mildest stages of AD-type dementia. This is important because, although component-level floor and ceiling effects will almost always exist to some extent, they should be minimized if the potential of the ADAS-Cog to detect change is to be maximized.

Our findings also have implications for mild cognitive impairment (MCI) populations, among whom it is likely that the ADAS-Cog's components would perform worse. In fact, a brief examination of those patients classified as having MCI in the ADNI study (n = 1150 person measurements, mean MMSE = 27) revealed larger component-level ceiling effects compared with the AD sample (52%–96% in 9 of 11 components [MCI] versus 32%–83% in 8 of 11 components [AD]), and poorer reliability (Cronbach's α [homogeneity coefficient range]: MCI, 0.50 [0.10–0.15]; AD, 0.74 [0.18–0.28]; data available from authors).

Our study has four key limitations. First, validity testing was limited. The aspects of convergent (i.e., MMSE) and discriminant (i.e., age, gender) validity examined in this study provide very broad assessments of validity at best and, at worst, can be considered weak. Further work would be valuable, such as assessment of construct validity including other cognitive performance measures and known groups validity testing against *a priori*, clinically driven hypotheses. Although the proviso is that these proposed analyses would not overcome the importance of the component-level ceiling effects. Second, our analyses were conducted on all of the available data as opposed to simply baseline data, which introduces the potential drawback of repeated measurements as within-person correlations can influence results. We selected this approach given the limited sample size of the baseline data. And, in fact, an analysis of the baseline data resulted in findings comparable to those presented herein (data available from the authors). However, it would be valuable to repeat these analyses on a larger cross-sectional data set.

The third limitation is that the interpretability of our responsiveness analysis is hampered by the lack of a "true change" criterion. Unfortunately, this is common to many psychometric studies, and the extent to which traditional responsiveness statistics are useful indicators of the ability of a rating scale to measure change is unclear. In this study our backdrop hypothesis was based on the clinical expectation that patients with AD will experience deteriorating cognitive performance over time. Despite this, our interpretation of the responsiveness statistics based on the ADAS-Cog data would be aided by supplemental analyses, such as the examination of relative responsiveness compared with concurrently collected rating scales, which purport to measure the same construct.

The final limitation is that the results of traditional psychometric analyses have many clinically relevant drawbacks, including sample and scale dependency and arbitrary criteria (detailed elsewhere [12]). Further examinations are required using new techniques, such as Rasch Measurement Theory [13], to *diagnose* the specific performance issues and potential areas for improvement in the ADAS-Cog.

Acknowledgments

This study was supported by grants from an anonymous foundation (to J.H.). Some of J.H., S.C., and J.Z.'s research time was funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0707-10124). The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The authors thank Teresa Driscoll for editorial assistance with this manuscript.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904). The ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from: Abbott Laboratories; the Alzheimer's Association; the Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences, Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec, Inc.; Bristol-Myers Squibb Co.; Eisai, Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Co.; F. Hoffmann-La Roche, Ltd., and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Servier; Synarc, Inc.; and Takeda Pharmaceutical Co. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of California, Los Angeles. This research was also supported by the National Institutes of Health (P30 AG010129 and K01 AG030514).

References

[1] Blennow K, de Leon M, Zetterberg H. Alzheimer's disease. Lancet 2006;368:387–403.

[2] Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement 2007;3:186–91.
[3] Alzheimer's Association. Alzheimer's disease facts and figures. Chicago: Alzheimer's Association; 2008.
[4] Aisen P, Schafer K, Grundman M, Pfeiffer E, Sano M, Davis K, et al. Effects of rofecoxib or naproxen vs placebo on Alzheimer disease progression: A randomized controlled trial. JAMA 2003; 289:2819–26.
[5] Mohs K, Rosen W, Davis K. The Alzheimer's Disease Assessment Scale: An instrument for assessing treatment efficacy. Psychopharmacol Bull 1983;19:448–50.
[6] Dichgans M, Markus H, Salloway S, Verkkoniemi A, Moline M, Wang Q, et al. Donepezil in patients with subcortical vascular cognitive impairment: A randomised double-blind trial in CADASIL. Lancet Neurol 2008;7:310–8.
[7] Cano S, Posner H, Moline M, Hurt S, Swartz J, Hsu T, et al. The ADAS-cog in Alzheimer's disease clinical trials: Psychometric evaluation of the sum and its parts. J Neurol Neurosurg Psychiatry 2010; 81:1363–8.
[8] Weyer G, Erzigkeit H, Kanowski S, Ihl R, Hadler D. Alzheimer's Disease Assessment Scale: Reliability and validity in a multicenter clinical trial. Int Psychogeriatr 1997;9:123–38.
[9] Doraiswamy P, Kaiser L, Bieber F, Garman R. The Alzheimer's Disease Assessment Scale: Evaluation of psychometric properties and patterns of cognitive decline in multicenter clinical trials of mild to moderate Alzheimer's disease. Alzheimer Dis Assoc Disord 2001; 15:174–83.
[10] U.S. Food and Drug Administration. Patient reported outcome measures: Use in medical product development to support labelling claims. Available at: www.fda.gov/cber/gdlns/prolbl.pdf. Accessed April 12, 2012.
[11] Hobart J, Cano S, Thompson A. Effect sizes can be misleading: Is it time to change the way we measure change? J Neurol Neurosurg Psychiatry 2010;81:1044–8.
[12] Hobart J, Cano S, Zajicek J, Thompson A. Rating scales as outcome measures for clinical trials in neurology: Problems, solutions, and recommendations. Lancet Neurol 2007;6:1094–105.
[13] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcome Res 2011;11:571–85.