Published in final edited form as:

J Alzheimers Dis. 2008 November ; 15(3): 461–464.

Administration and Scoring Variance on the ADAS-Cog

Donald J. Connor* and Marwan N. Sabbagh

Cleo Roberts Center for Clinical Research, Sun Health Research Institute, Sun City, AZ, USA

Abstract

The Alzheimer's Disease Assessment Scale – Cognitive (ADAS-Cog) is the most commonly used primary outcome instrument in clinical trials for treatments of dementia. Variations in forms, administration procedures and scoring rules, along with rater turnover and intra-rater drift may decrease the reliability of the instrument. A survey of possible variations in the ADAS-Cog was administered to 26 volunteer raters at a clinical trials meeting. Results indicate notable protocol variations in the forms used, administration procedures, and scoring rules. Since change over time is used to determine treatment effect in clinical trials, standardizing the instrument's ambiguities and addressing common problems will greatly increase the instrument's reliability and thereby enhance its sensitivity to treatment effects.

Keywords

ADAS-Cog; Administration; Alzheimer's disease; clinical trials; dementia

INTRODUCTION

In clinical trials on dementia, the Alzheimer's Disease Assessment Scale-Cognitive section (ADAS-Cog) [5] is the most commonly used objective measure of cognitive change. It was designed to measure cognitive areas commonly seen to decline in Alzheimer's disease (AD), specifically learning (word list), naming (objects), following commands (1 to 5 elements), constructional Praxis (copy 4 figures), ideational Praxis (mail a letter), orientation (person, time and place), recognition memory (from a second word list), and remembering test instructions (from the recognition subtest). The test usually includes three additional subjective scales containing impression of spoken language ability, word finding difficulty, and comprehension. A final subjective scale measuring concentration and distractibility is sometimes added. The ADAS-Cog has been shown to be sensitive to change across various levels of dementia, with minimal floor effects, but some question of ceiling effects [4]. The issue of sensitivity to very mild impairments and MCI has been addressed in some versions of the instrument by adding a delayed free recall of the first word list (short-term memory).

Despite being the gold standard in clinical dementia trials since the early tacrine studies [2], the administration procedures, work sheets, and scoring procedures of the ADAS-Cog were not precisely defined in the original article [5]. Further, the responses to situations that are often encountered when working with a demented population were not well operationalized and seldom addressed in the training sessions at clinical trial meetings. This may lead to some raters utilizing their individual judgments, thus reducing the inter-rater reliability. In 1998, the

Alzheimer's Disease Cooperative Study group (ADCS), under the guidance of Kimberly Schafer, developed an ADAS-Cog kit that included high-quality items for the naming task, card sets for the word lists, and a revision of the original manual [1,3] that attempted to operationalize much of the ADAS-Cog administration and scoring procedures. However, even this standardized version has shown modification over time (personal communication). This study attempts to determine if there remains variance in the ADAS-Cog forms and procedures between different clinical trials.

METHODS

A 22 item survey was distributed to ADAS-Cog raters who were attending an Elan investigator training meeting in 2007 for a potential treatment of AD. Participation was voluntary. The survey covered topics ranging from the test materials to administration procedures and scoring rules (see Table 1). Participants were asked to, "Please answer the following questions based on the way(s) you have been instructed to give the ADAS-Cog or seen it given in the various clinical trials protocols you have worked with. Please circle *all* answers that you think are correct or *all* the ways you have seen it done." For the purposes of summarizing the amount of variability, if a question asked about specific procedures and the participant circled two or more answers, then the response was counted as their being trained in more than one way to administer or score the item. This survey sought to discover the percent of individual raters who were being trained to administer the instrument differently between clinical trials (intra-rater difference), and was not specifically designed to assess if different raters were administering the test differently (inter-rater difference).

RESULTS

Twenty-six raters volunteered to fill out the survey. Rater experience varied from 1–12 years of administering the ADAS-Cog (5.94 ± 3.64 yrs; $n = 26$), and number of administrations varied from 40–1000 (296 ± 288 ; $n = 21$). The number of different protocols in which the ADAS-Cog was used ranged from 3–20 (10.6 ± 5 ; $n = 21$) with most raters having a mix of pharmaceutical industry sponsored and government sponsored (ADCS) protocols (68%; $n = 19$). Education background also varied widely from 14–20 years of education (associate's degree to Ph.D.; $n = 24$).

The survey results are summarized in Table 1. The percent of raters that reported being trained differently on the scale or an item not being addressed in training varied from 19.2% who reported being told to use different topics for the initial interview, to 73% who reported variance in the maximum amount of time allowed for stimulus exposure in the word list task. Because of missing demographic data, a direct comparison of rater experience to the number of variations acknowledged was not done. However, approximately half of the respondents had 5 or more years experience with the ADAS-Cog and 65% indicated that they had given it for 7 or more different clinical trial protocols. Fifteen percent indicated using the scale in 6 or fewer trials (20% did not respond).

DISCUSSION

This brief survey indicates significant variance in intra-rater experience with the ADAS-Cog administration procedures, scoring rules, and materials used in clinical trials. While the impact of this variability on clinical trial outcome results is unknown, such variance is a threat to test-retest reliability. Any decrease in reliability can lead to an increase in nonspecific variability of the data (background noise) and therefore a decrease in ability to detect an effect of the compound under study. Utilization of different methodologies between clinical trials may also

affect the validity of comparing outcomes between different protocols, or of pooling the results in meta-analytic studies.

Since this survey was meant to assess intra-rater variance, it may underestimate the true variance in training by not assessing inter-rater training variance. For example, the item that assessed what raters had been taught to do if the subject makes an error in the Command section of the ADAS-Cog indicated that 38% of the raters were taught different ways to respond on different protocols (intra-rater variance). However, of the 62% that indicated they had only been taught one way to respond, half (31% of the total sample) indicated they were instructed to give a second chance and the other half indicated they were instructed to count the response as an error and move to the next item (inter-rater variance). Similarly, when asked whether a circumlocution would be counted as an error under the “Spoken Word” section, 52% of the raters indicated they had been instructed differently in different protocols; but of the 48% that indicated they had only been told one way to score it, one-third of them would count it as an error and two-thirds would not. A similar pattern was found for items assessing whether a second chance should be given after an error in the naming section; scoring criteria for the rhombus; and acceptable responses to the “place” question in the orientation section. Again, while evaluation of inter-rater differences is limited by the study design, this pattern is consistent with variations in clinical training.

The length of the clinical trials – which have expanded noticeably since the original cholinesterase inhibitor trials – also increases the chance that the rater who started a trial may not be the same one who is rating the patient at the end of the trial. While some protocols try to maintain the same ADAS-Cog rater-patient pairs for the duration of a trial, many consider the ADAS raters interchangeable. Consequently, good inter-rater reliability is critical to the trial outcome. The increase in trial duration may also lessen *intra*-rater reliability. There may be a tendency for administration and scoring standardization to “drift” over time as a rater becomes more experienced, again emphasizing the necessity for well operationalized instruments. This threat to intra-rater reliability is further increased when a particular rater may be working on several clinical trials, each using the ADAS-Cog but with various modifications in administration, scoring, source documents and case report forms. Trying to track the variations on the instrument in each protocol (especially if several different clinical trial evaluations are administered in the same day) increases the likelihood of rule substitution mistakes or merging of the different protocol rules.

Test materials have also varied over time, including a wide range in the quality of the naming materials, word card decks, instruction manuals, and worksheets. For example, the administration forms have ranged from 14 pages in the original format (including administration and scoring instructions), to two pages just containing item scores plus 4 recording pages for drawings. While the latter is space efficient and perhaps increases ease of data entry, this format may make the test both more difficult to reliably administer (no administration instructions on page) and to properly monitor (actual responses not recorded) in clinical trials. This may exacerbate the problem of adhering to differences in administration rules between protocols. Despite the revised manual’s attempt to address many common administration problems, there remain many issues that were not fully explored [1]. This has been exacerbated by variations of the standardized version based on sponsor preferences and other factors.

Since there was a large range of rater experience with the instrument and the number of protocols given, this study may underestimate or overestimate the true variability in the way the ADAS-Cog is currently being taught. While those raters with limited experience may not be exposed to enough trials to reflect the diversity in administration and scoring instruction, those raters with many years of experience may overestimate the variability if the

standardization of the ADAS-Cog has increased in recent years. The latter seems somewhat unlikely since the majority of raters started testing after the standardized kits from the ADCS became available. There is some indication that the ADAS-Cog shows significant variation between European countries [6] so a multi-national sampling may show even greater levels of variation.

While this study is limited by small numbers, a select group of raters (sites registered at an Elan investigators training meeting in North America), and there may be an issue of sampling bias (e.g., those who experienced the most variance in training may have been more likely to volunteer), the findings appear intriguing. In keeping with the current trend in dementia clinical trials, a larger multinational survey of these issues in experienced raters should be executed to confirm the results. If the results of this survey are replicated in larger studies, then an argument for a single, standardized version of the ADAS-Cog with expanded administration and scoring rules could be made. While clinical trial sponsors, trainers, and others may believe certain modifications to the ADAS-Cog may improve its sensitivity and specificity – and indeed they may – the impact of these changes on rater reliability and the ability to compare outcomes between studies should be carefully considered.

ACKNOWLEDGMENTS

This survey was only possible through the permission of Elan Pharmaceuticals to disseminate it at an investigators meeting. We would especially like to thank Dr. Enchi Liu for her support.

This work is supported by Arizona State Grants AGR2007-07 and ABRC-0011; NIA GrantAG019610; Alzheimer's Association NIRG-04-1159; and the Michael J. Fox Foundation.

Dr. Sabbagh receives consulting fees, lecture fees and/or grant support from Eisai, Elan, Forest, GSK, Lilly, Novartis, Pfizer, and Wyeth.

References

1. Connor, DJ.; Schafer, K. Administration Manual for the Alzheimer's Disease Assessment Scale. San Diego: Alzheimer's Disease Cooperative Study; 1998.
2. Davis KL, Thal LJ LJ, Gamzu ER, Davis CS, Woolson RF, Gracon SI, Drachman DA, Schneider LS, Whitehouse PJ, Hoover TM, Morris JC, Kawas CH, Knopman DS, Earl NL, Kumar V, Doody RS. and the Tacrine Collaborative Study Group. A double-blind, placebo-controlled multi-center study of tacrine for Alzheimer's disease. *N Engl J Med* 1992;327:1253–1259. [PubMed: 1406817]
3. Mohs, RC. Administration and Scoring Manual for the Alzheimer's Disease Assessment Scale. New York: Mount Sinai School of Medicine; 1994.
4. Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C, Grundman M, Sano M, Bieliauskas L, Geldmacher D, Clark C, Thal LJ. Development of cognitive instruments for use in clinical trials of antidementia drugs: Additions to the Alzheimer's Disease Assessment Scale that broaden its scope. *Alzheimer Dis Assoc Disord* 1997;11:13–21. [PubMed: 9194962]
5. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. *Am J Psychiatry* 1984;141:1356–1364. [PubMed: 6496779]
6. Verhey FR, Houx P, Van Lang N, Huppert F, Stoppe G, Saerens J, Böhm P, De Vreese L, Nordlund A, DeDeyn PP, Neri M, Peña-Casampva J, Wallin A, Bollen E, Middelkoop H, Nargeot MC, Puel M, Fleischmann UM, Jolles J. Cross-national comparison and validation of the Alzheimer's Disease Assessment Scale: Results from the European Harmonization Project for Instruments in Dementia (EURO-HARPID). *Int J Geriatr Psychiatry* 2004;19:41–50. [PubMed: 14716698]

Table 1

Summary of rater endorsement of inconsistencies in training of the ADAS-Cog between different clinical trial protocols (intra-rater, inter-trial)

Item	Variance
Variations in the length of the work sheets	65.4%
Variation in initial interview form	19.2%
Used an unstructured initial interview	73%
Never received topics for the initial interview	52%
Some protocols provide topics for interview and others do not	44%
ADAS scored either for errors or correct responses depending on the protocol	61.5%
Differences in the ADAS word lists between protocols	69.2%
Different minimum exposure times for the words on the word lists	42.3%
Different maximum exposure times for the words on the word lists, or not addressed in training	73%
Word Recall subtest scored for errors or correct responses in different protocols	56%
Total Word Recall score calculated in various ways in different protocols	53.8%
Variation between protocols on whether to give the semantic cue after a naming error on the naming task	30.8%
Variation between protocols on whether to give the semantic cue if the participant only names the object's function on the naming task	50%
Report having been given different instructions if an error was made in the Command section	38%
Variation between protocols on what to do if the participant makes an error on the drawing task	46%
Report being taught different criteria for scoring the Rhombus	46%
Maximum number of folds acceptable (Ideational Praxis) was not addressed in training	34.6%
Variation between protocols on what to do if the participant writes the address on the back of the envelope (or not addressed in training)	30.7%
Variation between protocols on criteria for "place" in different studies	38%
Report a different number of trials for word recognition between protocols	64%
Variation between protocols for when to give the reminder cue in word recognition, or it was not defined in training	62%
Variation between protocols (or not addressed in training) as to whether a circumlocution is an error in the Spoken Word section	52%
Variation between protocols (or not addressed) as to whether a circumlocution is an error in the "Word Finding" section	41%
Variation between protocols as to whether the Concentration/distractibility section was included in the ADAS	38%