

Scale Rating Variability in Alzheimer's Disease Clinical Trials: A Literature Review

Weber, CJ¹, De Santi, S^{2,5}, Kirby, LC⁴, Cisneros, W¹, Patrick, K¹, Wolanski, K¹, Williams, JBW^{1,3}
MedAvante, Inc., Hamilton, NJ¹, NYU Langone Medical Center², College of Physicians and Surgeons, Columbia University³, Critical Path Institute, Phoenix, AZ⁴, Bayer Healthcare Pharmaceuticals⁵

ABSTRACT

Background: Clinician-administered cognitive assessments in clinical trials of Alzheimer's disease (AD) have multiple sources of variability that can collectively diminish or obscure the treatment effect. Both structured (e.g., Alzheimer's Disease Assessment Scale-Cognitive subscale or ADAS-cog) and semi-structured assessments (e.g., Clinical Dementia Rating Scale or CDR) have significant intra- and inter-site variability in test administration and interpretation of subject responses and scoring. Too much variability in key selection and efficacy/outcome measures can obscure a true treatment effect.

Methods: We reviewed the literature from PsychMed, PsychInfo, and Medline regarding assessment variability with the ADAS-cog, the MMSE and the CDR (i.e., scale administration and scoring deviations). The review focused on how this variability can potentially impact signal detection, including effect size, subject selection, power calculations, costs, and the success or failure of a study.

Results: The ADAS-cog has significant sources of administration and scoring deviations. For example, site raters working on multiple dementia protocols have a substantial likelihood of failing to administer and score according to each protocol's specified guidelines.¹⁴ Tractenberg, et al.²¹ found CDR scoring inconsistencies between experienced and inexperienced raters when rating cognitively normal subjects and those with questionable dementia. The CDR domains most susceptible to increased rater variability included memory, judgment, and problem solving. Furthermore, long AD trials incur increased rater turnover and rater drift, which represent additional sources of scoring variability in efficacy ratings.

Conclusions: Low reliability of outcome assessments may have contributed to recent failed AD drug trials. Ratings variability, which reduces treatment effect size, requires increased subject numbers and incurs increased costs. Identified causes of rater variability have implications for rater training and novel ongoing rater monitoring.

Disclosures: Weber C, Williams JBW, Cisneros W, Patrick K, Wolanski K are employed by MedAvante. Kirby L is a consultant for MedAvante. De Santi S is a senior scientist at Bayer and former employee of MedAvante.

INTRODUCTION

Clinician-administered cognitive assessments in clinical trials of Alzheimer's disease (AD) have multiple sources of variability that can collectively diminish or obscure a treatment effect. Both structured (e.g., Alzheimer's Disease Assessment Scale-Cognitive subscale; ADAS-cog) and semi-structured assessments (e.g., Clinical Dementia Rating Scale; CDR) have significant intra- and inter-site variability in test administration, interpretation of subject responses and scoring.

Too much variability in key selection and efficacy/outcome measures can obscure detection of a true treatment effect. The frequent use in AD trials of these instruments underscores the importance of understanding the sources and magnitude of variability.

AD clinical trials have failed for a variety of cited reasons and among them are ratings variability that may obscure a real treatment effect. Efforts to compensate for this variability can include increased sample size; however, Gold concluded that "by increasing the number of sites, the heterogeneity of patients, investigators, and raters is also increased and leads to a loss of accuracy."¹

By comparison to other reasons for trial failure, relatively little attention has been given to how trial inclusion and subject efficacy ratings performance impacts the accuracy of the data, and ultimately the success of a clinical trial. This review focuses on how scale administration variability can impact subject selection, signal detection, and consequently the success or failure of a study.

METHODS

We queried PsychMed, PsychInfo, and Medline regarding AD assessment variability (i.e., scale administration and scoring problems), reviewed the relevant articles, and collated the responses for presentation here.

RESULTS

The goal in AD trials is to detect whether there is a divergence in test scores between the treatment group and the placebo group. The past few years have seen published examples of how major phase III trials have failed due to lack of separation between the treatment and placebo groups on measured endpoints. Ratings variability in the assessments themselves is one potential explanation for these results. Examples include:

Recent Failed Phase III Clinical Trials	
Phenserine	No significant differences were found between treatment and placebo groups on the ADAS-cog or CIBIC. ²
Tarenflurbil	No significant difference between the Tarenflurbil and placebo groups on the ADAS-cog or ADL. ³
Tramiprosate	No significant difference between Tramiprosate and placebo on the ADAS-cog or CDR. ⁴
Xaliprioden	No significant difference between Xaliprioden and placebo on the ADAS-cog or CDR. ⁵
Dimebon	No significant difference between Dimebon and placebo on the ADAS-cog, CIBIC-plus, or ADL. ⁶

The ability of an assessment instrument to show a treatment effect by demonstrating a separation of the placebo group from the treatment group can be challenging when evaluating relatively small amounts of decline over timespans extending over many months. Accurate measurement of cognitive status throughout a clinical trial requires sensitive assessment tools correctly administered and scored by well-trained clinicians. For example, Schneider and Sano⁷ examined ADAS-cog and CDR changes in the placebo groups of seven completed phase III, 18-month clinical trials, showing that the rate of decline in placebo groups varies widely from study to study. The mean ADAS-cog worsening was from 4.34 to 8.14, and the mean CDR sum of box change was from 2.05 to 2.74. In all instances, standard deviations exceeded their accompanying means. In a 70-point ADAS-cog or 18-point CDR Sum of Boxes, these are relatively subtle change scores, and demonstrating a significant separation from the treatment group leaves little room for noise due to scale administration variance.

Several articles highlight an array of design issues in AD clinical trials^{1,7,8,10,11}. They discuss potential problem areas including disease state severity, trial length, and choice of dosage, sample size, and selection of primary and secondary endpoints. All these play a critical role in designing a trial, with any single one threatening the success of a trial if not well chosen and managed. In addition, an inadequately administered primary endpoint rating scale will undermine even a well-designed trial. Interview quality and ratings performance is seldom brought up in the context of clinical trial methodology and factors that affect success or failure^{17,8,9,11}, yet a lack of quality assurance measures suggests that the administration and scoring quality of these assessments can not be verified: "Test bias is often not adequately addressed, and can be problematic..."¹⁰ Black et al. noted that "test bias is not adequately addressed..." (p.325) in clinical trials. They specifically recognized the potential test bias resulting from changing psychometric raters during trials, a factor that is difficult to manage as trials become increasingly longer¹⁰.

Selection of the proper study subject is also critical to effect size measurement. Clinical trials introduce selection bias by allowing the site rater to know the thresholds needed for inclusion into the study¹². This knowledge can allow unqualified subjects entry into a trial, an issue widely documented in mood trials¹³, and which parallels the issues encountered in AD.

Endpoints for AD trials must cover a broad spectrum of domains including cognitive functioning, functional status, and neuropsychiatric symptoms. We focus on 3 scales commonly used in screening, and tracking dementia symptoms in AD clinical trials, and discuss how variability in their use (administration and/or scoring) can potentially impact research outcomes.

The **ADAS-cog** is widely used as the primary or co-primary outcome measure of cognitive functioning in AD clinical trials. Its total score is tallied by the number of errors, where a higher score indicates a worse performance.

Despite its widespread use, however, the ADAS-cog's administration, scoring, and test materials have become increasingly diverse, as clinical trial sponsors have adapted the scale to suit individual trials. The result of each change to this original scale, however small, requires additional rater training. This reduces inter-rater reliability, as raters working on multiple protocols may have difficulty tracking the differences from one protocol to the next.

Connor and Sabbagh¹⁴ reported numerous areas of variance in inter- and intra-rater experience with the ADAS-cog. Many of the items reported have obvious and significant scoring implications. For example, 52% of raters surveyed noted that they were instructed to score circumlocution in the "Spoken Word" section differently across various protocols; however, of the 48% that reported having been instructed to score it one way, one-third of the raters would count this as an error and the others would not. These scoring and administration variations are a major concern if the ADAS-cog is a primary endpoint.

In addition, the ADAS-cog's sensitivity to change varies depending on the degree of AD severity¹⁵. Measured deterioration is slower for mildly and severely demented patients (less than 4 points per year) than for patients with moderate dementia (7-11 points per year)^{15,16}. In the increasingly common trials in milder dementia, reducing ratings variability becomes critical to treatment effect detection.

The **MMSE** was developed to help standardize the bedside examination of the cognitive state and is used as a screening qualification measure for most dementia trials. The MMSE clinical guide¹⁷ states that one training session with an experienced administrator is usually sufficient to achieve inter-rater reliability.

The MMSE manual published by PAR, Inc.¹⁷ reports inter-rater reliability coefficients (Pearson product moment correlations) from .83 to .95¹⁸. Other publications, however, have shown the inter-rater reliability of the MMSE to be more variable.

In a study by Bowie et al.,¹⁹ 40 clinicians were asked to administer the MMSE to a standardized patient who gave the same "scripted" responses to each participant yielding an expected MMSE total of 20 (out of 30). MMSE scores for all 40 participants ranged from 14-27, illustrating the magnitude of the variability.

The **CDR** is a widely used, semi-structured global assessment for dementia²⁰. It is regularly used as a primary or secondary outcome measure in AD clinical trials. Information obtained from the subject and caregiver (informant) interviews is used to score functioning in six domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Training clinicians on the CDR involves instruction on rating each of the domains (box score) and how to use the semi-structured format to obtain information in interviews with the informant and patient. The CDR sum of boxes can vary from 0-18.

The CDR has been shown to be quite sensitive to even small degrees of clinical change, but requires clinician judgment to accurately rate the various domains. Furthermore, the semi-structured nature of this interview requires familiarity and a comfort level in presenting the items in a conversational, interview style. Therefore, interview skills as well as clinical judgment are critical to the validity of this assessment¹⁰.

Discrepancies in CDR ratings can have a major impact on the outcome of a clinical trial where the CDR is used as a screening criterion or primary measure. Tractenberg et al.²¹ identified CDR Box Score disagreements among raters across various levels of CDR administration experience. For example, raters-in-training experienced the most difficulty with rating normal and questionable dementia patients, and also had the most difficulty scoring the Memory domain against the gold standard. The authors note the importance of customized training programs to increase agreement among all experience levels²¹, but do not address how to sustain this agreement over a 1-2 year (or longer) clinical trial.

SUMMARY & CONCLUSIONS

Commonly used rating scales in AD trials have variability as documented by a number of authors. The variability in the cited scales may have contributed to negative AD efficacy trials and therefore has implications for planning future trials. Additional research is warranted. The findings have implications for training and monitoring raters before and during the course of an AD trial.

References

- Gold M. Study design factors and patient demographics and their effect on the decline of placebo-treated subjects in randomized clinical trials in Alzheimer's disease. *J Clin Psychiatry*. 2007 Mar;68(3):430-8.
- Axonys Announces That Phenserine Did Not Achieve Significant Efficacy in Phase III Alzheimer's Disease Trial. *Business Wire*. FindArticles.com. 10 Jun, 2010. http://findarticles.com/p/articles/mi_m0ERN/vs_2005_Feb_7/ai_n9494083/.
- Gordon W, Sandra B, Alfred HB, David A, Andrew B, Lon S, Robert G, Edward S, Kenton Z. Safety and efficacy of tarenflurbil in subjects with mild Alzheimer's disease: Results from an 18-month international multi-center phase 3 trial. *Alzheimer Dement*. 2009 Jul; 5(4):86.
- Sullivan MG. Tramiprosate fails short in phase III Alzheimer's trial: unusually large placebo effect could be a recurring problem in studies that allow concomitant medications. *Clinical Psychiatry News*. FindArticles.com. 10 Jun, 2010 http://findarticles.com/p/articles/mi_n04345/v1_11_35/ai_n29399056/.
- Douville P, Gogozou JM. What we have learned from the Xaloprioden Sano1-Aventis trials. *J of Nutrition, Health, & Aging*. 2009 13(4): 365.
- Jeffery S. Dimebon Disappoints: Is There Hope for Novel Alzheimer's Agent? *Medscape Medical News*. 10 Jun, 2010. <http://www.medscape.com/viewarticle/718401>.
- Schneider LS, Sano M. Current Alzheimer's disease clinical trials: methods and placebo outcomes. *Alzheimer Dement*. 2009 Sep;5(5):388-97.
- Knopman DS. Clinical trial design issues in mild to moderate Alzheimer disease. *Cogn Behav Neurol*. 2008 Dec;21(4):197-201.
- Prins ND, et al: Can novel therapeutics halt the amyloid cascade? *Alzheimer's Research and Therapy* 2010, 2:5.
- Black R, Greenberg B, Ryan JM, Posner H, Seeburger J, Amatniek J, Resnick M, Mohs R, Miller DS, Saumier D, Carrillo MC, Stern Y. Scales as outcome measures for Alzheimer's disease. *Alzheimer's Dement*. 2009 Jul;5(4):324-39.
- Vellas B, Andrieu S, Sampaio C, Wilcock G. European Task Force group. Disease-modifying trials in Alzheimer's disease: a European task force consensus. *Lancet Neurol*. 2007 Jan;6(1):56-62.
- Kirby L, Borwege S, Christensen J, Weber C, McCarthy C. Reducing Placebo Response: Triple Blinding & Setting Expectations; Strategies for eliminating factors that influence the placebo response during the clinical trial process. *Applied Clinical Trials*. 2005 Nov; 14(11): 48-52.
- Detke M, Williams J, Koback K, Ellis A, Giller E, Leon A, Reines S, Kane J. The Challenge of Patient Ascertainment in Clinical Trials -- New Data. Poster presented at the International Society for CNS Clinical Trials and Methodology, Autumn Conference, San Diego, CA (October 2009).
- Connor DJ, Sabbagh MN. Administration and scoring variance on the ADAS-Cog. *J Alzheimers Dis*. 2008 Nov;15(3):461-4.
- Rogers SL, Farlow MR, Doody RS, Mohs R, Friedhoff LT. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. *Donepezil Study Group*. *Neurology*. 1998 Jan;50(1):136-45.
- Stern RG, Mohs RC, Davidson M, Scheidegger J, Silverman J, Kramer-Ginsberg E, Searcy T, Bierer L, Davis RL. A longitudinal study of Alzheimer's disease: measurement, rate, and predictors of cognitive deterioration. *Am J Psychiatry*. 1994 Mar;151(3):390-6.
- Folstein MF, Folstein SE, Folstein G. The Mini-Mental State Examination Clinical Guide. Psychological Assessment Resources, Inc. 2001. www.parin.com
- Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": A practical method for grading the clinician. *J Psychiatry Res* 1975; 12: 129-138.
- Bowie P, Branton I, Holmes J. Should the Mini Mental State Examination be used to monitor dementia treatments? *The Lancet*. 1999. 354, 1527-1528.
- Hughes CP, Berg L, Danziger WL, Cohen LA, Martin RL. A new clinical scale for the staging of dementia. *Br J Psychiatry*. 1982 Jun;140:566-72.
- Tractenberg RE, Schafer K, Morris JC. Interobserver disagreements on clinical dementia rating assessment: interpretation and implications for training. *Alzheimer Dis Assoc Disord*. 2001 Jul-Sep;15(3):155-61.