

The MARQ-AD: A New Measure of Interview Quality in Alzheimer's Disease Trials

Weber, C¹ Williams, JBW^{1,2} Kirby, L^{1,3}
 MedAvante, Inc.¹, College of Physicians and Surgeons, Columbia University², Critical Path Institute³

ABSTRACT

Background: Clinician-administered neurocognitive and interview-based assessments in clinical trials of Alzheimer's disease (AD) have multiple sources of variability that can collectively diminish or obscure a treatment effect. It has been demonstrated that interview quality is directly related to better detection of a treatment effect. A quality measurement tool to assess and improve test instrument administration would be a desirable goal given the high stakes in Alzheimer's disease (AD) clinical trials. The MedAvante Analysis of Rating Quality - Alzheimer's Disease (MARQ-AD) was developed specifically for Alzheimer clinical trial endpoints to quantify critical domains of rating/interview quality and identify administration and scoring deviations encountered in AD assessments in randomized clinical trials. This poster presents the development of the MARQ-AD scale and a review of clinician training on the scale. The MARQ-AD is an integral component of the MedAvante quality control process, which mirrors the successful use of a similar process applied to psychiatric trials.

Methods: The MARQ-AD was developed for Alzheimer's rating scales from the Rater Applied Performance Scale (RAPS), a scale developed for psychiatric ratings³. Seven doctoral level psychometricians with at least 7 years of AD assessment experience ("Trainers") were selected to perform independent rater evaluations using the MARQ-AD. Each Trainer learned the administration and scoring parameters of the MARQ-AD as applied to the MMSE, ADAS-cog, CDR, ADCS-ADL, and ADCS-CGIC. Trainers then reviewed and rated a series of videos of outside clinicians performing actual patient and/or caregiver interviews. The Trainers independently scored the interviews along the specified MARQ domains, and the scores were collated. After each set of videos, the group of trainers compared scores and discussed their ratings. Variability of ratings prompted revisions to the MARQ-AD that were then tested in the next set of interviews. We present data from three calibration exercises, where Trainers viewed and scored the MMSE, ADAS-cog, CDR, ADCS-ADL, and ADCS-CGIC with accompanying MARQ-AD scores for an AD patient presenting with mild dementia symptomatology. An average domain score for each MARQ-AD was calculated, and the assessment was assigned a "Meets Criteria" or "Does Not Meet Criteria" rating, the latter of which may indicate the need for rater remediation.

Results: The independently-derived MARQ-AD scores (n=105) showed good agreement across the 7 Trainers. By group consensus, 14 of the 15 assessments were rated as "Meets Criteria" and considered to be of good interview quality (the MMSE administration from calibration exercise #1 was rated "Does Not Meet Criteria"). There was 100% agreement among Trainers in all but four assessments in this calibration exercise.

Conclusions: The MARQ-AD is derived from the RAPS, a validated instrument for assessment of the quality of administration of psychiatric rating scales. After group calibration, qualified Trainers who participated in MARQ-AD training achieved nearly good agreement when reviewing a video recording of common AD scales. The MARQ-AD could become integral to an ongoing quality control process incorporated into an AD clinical trial with the goal of reducing assessment scale variability, making detection of a true treatment effect more possible.

Disclosures: Weber C and Williams JBW are employed by MedAvante. Kirby L is a consultant for MedAvante and Director of Critical Path Institute.

INTRODUCTION

Clinician-administered neurocognitive and interview-based assessments in clinical trials of Alzheimer's disease (AD) have multiple sources of variability that can collectively diminish or obscure a treatment effect.

Possible sources of variability that can compromise standardization include:

- Inconsistent application of administration and scoring rules¹
- Inadequate experience on scales or with AD patients
- Poor consistency of scale administration, as measured by inter-rater reliability (intra-class correlation; ICC)
- Intra-rater reliability issues (e.g., rater drift) particularly in longer trials¹
- Confusion or apathy about attending to scale-specific instructions across several clinical trials (e.g., WORLD backwards vs. serial 7s)
- Increased rater turnover in long AD trials (≥18 months)¹

Interview quality determines the reliability of the rating. Ongoing ratings, monitoring, and feedback have been shown to improve rating quality in psychiatric clinical trials² and to improve signal detection³.

Need for measurement of AD scale administration quality

A measurement tool is needed to objectively quantify AD scale administration quality. The RAPS³ is a validated instrument to assess interview quality in psychiatric scales. Incorporated into an overall rater quality process, it has been used as a tool to improve assessment quality, demonstrating improvement in signal detection³. AD scales have distinct differences from psychiatric scales and the MARQ-AD was derived from the RAPS to reflect those distinctions in the AD scale setting.

- The MARQ-AD utilizes specific rating domains important to the correct administration of the scale.
- Because of the unique character of AD scales, different versions of the MARQ-AD have been developed.
- Different versions have overlapping but distinct domains to reflect the differences in structured scales (e.g., ADAS-cog) and semi-structured scales (e.g., CDR).

METHODS

Development of the MARQ-AD

- MARQ-AD consists of 3 versions, each containing scoring domains designed to capture the key elements in:
 - Structured AD scales (e.g., MMSE, ADAS-cog)
 - Semi-structured AD scales (e.g., CDR, ADCS-ADL)
 - Unstructured AD scales (e.g., ADCS-CGIC)
- Each version is independent, with individual detailed scoring guidelines, including descriptions of each domain and anchor points.
- The MARQ-AD is used for the administration of a single scale. For example, in a typical clinical trial setting, an ADAS-cog, an ADCS-ADL and a CDR would be administered to a patient in a single visit. This would generate three MARQ-AD scores and each MARQ-AD instrument used would be the one developed for that type of AD scale.

Domains of the MARQ-AD

- The MARQ-AD includes these domains that are critical components of assessment quality:
 - Adherence
 - Support and engagement
 - Pacing
 - Scoring accuracy
 - Interview comprehensiveness
 - Follow-up and clarification
 - Neutrality
 - Assessment environment*
- Not all domains are included in each version of the MARQ-AD, as not all domains are relevant to every scale. For example, in a structured assessment such as the MMSE, there is no need to evaluate Neutrality.
- Each domain is scored independently on a 4-point Likert scale: Excellent, Good, Marginal, Unacceptable.

MARQ-AD Clinician Training

- The goal of the MARQ-AD training was to calibrate a group of MARQ-AD "Trainers" so the ratings and feedback to raters are closely correlated. We undertook the task of identifying and subsequently training the clinicians to achieve the required calibration.
- Doctoral level psychometricians (n=7) with at least 7 years of psychometric experience in AD trials participated in training and calibration on the MARQ-AD.
- Trainers participated in an initial two-day training workshop to review the administration and scoring conventions for the MMSE, ADAS-cog, CDR, ADCS-ADL, and ADCS-CGIC, and to introduce the MARQ-AD domains and scoring guidelines for each version.

Training Sessions

- The development of the MARQ-AD and accompanying scoring conventions was an iterative process, and the data presented represent the culmination of monthly calibration exercises.
- Each Trainer independently reviewed the 5 videotaped assessments and completed the MARQ-AD for each assessment. MARQ-AD scoresheets were collected and collated prior to the group discussion.
- Training sessions were conducted to improve the scoring guidelines of the MARQ-AD, and therefore improve calibration of Trainer agreement and consistency among their MARQ-AD ratings. For example, following one calibration exercise, the scoring guidelines between MARQ-AD ratings (i.e., Excellent vs. Good) were revised on the MMSE. It was decided that one or two administration or scoring errors could significantly impact the total score, whereas similar errors committed on the ADAS-cog would have less of an impact on the total score.
- In the three most recent training sessions (data presented here), each Trainer independently reviewed and scored 3 "sets" of video taped assessments and completed the MARQ-AD for each scale (in training sessions #2 and #3, half of Trainer group performed audio-only review of all assessments). MARQ-AD scoresheets were collected and collated prior to the group discussion. A moderated group training telephone conference was conducted to discuss the ratings and the MARQ scores. Deviations from the group average were collectively discussed. MARQ-AD feedback was not provided to the site raters.

METHODS *continued*

Training Materials

- Video-recorded interviews were produced by an Alzheimer's disease research center (Banner Alzheimer's Institute, Phoenix, AZ) using their own selected raters, patients and caregivers.
- All patients/caregivers were each interviewed on five AD scales (MMSE, ADAS-cog, CDR, ADCS-ADL, and ADCS-CGIC) while a video-recording was obtained. After the interview, the site faxed all rater scoring sheets (e.g., drawings and source documentation) to the authors, which were then distributed as pdf files to the Trainers. The video recordings were downloaded from MedAvante's central server for review by the Trainers.

RESULTS

- Average MARQ-AD scores (n=105) were examined for each completed assessment (21 MMSE, 21 ADAS-cog, 21 CDR, 21 ADCS-ADL, and 21 ADCS-CGIC), and appropriate changes were made to the MARQ-AD.
- An algorithm for determining whether or not a site rater needs remediation ("Meets Criteria" or "Does Not Meet Criteria") was developed. Each domain has an attached score. Excellent scored a 4, Good a 3, Marginal a 2 and Unacceptable a 1. We calculated the average of the MARQ domains, which yielded an overall MARQ-AD score. Subsequently, a score of
 - ≥ 2.5 = "Meets Criteria"
 - < 2.5 = "Does Not Meet Criteria"
- Also, if any MARQ-AD domain was rated a "1", a "Does Not Meet Criteria" rating was given, regardless of the average domain score.
- In a clinical trial, these criteria would determine if rater remediation is necessary.

- The table presents the Trainers' Meets Criteria/Does Not Meet Criteria Ratings for the calibration exercise using the final MARQ-AD.

MARQ-AD Meets Criteria/Does Not Meet Criteria Ratings

● = Meets Criteria
 ● = Does Not Meet Criteria

		Trainer						Agreement
		1	2	3	4	5	6	
Calibration exercise #1	MMSE	●	●	●	●	●	●	100%
	ADAS-cog	●	●	●	●	●	●	100%
	CDR	●	●	●	●	●	●	86%
	ADCS-ADL	●	●	●	●	●	●	100%
Calibration exercise #2	ADCS-CGIC	●	●	●	●	●	●	100%
	MMSE	●	●	●	●	●	●	100%
	ADAS-cog	●	●	●	●	●	●	100%
	CDR	●	●	●	●	●	●	100%
Calibration exercise #3	ADCS-ADL	●	●	●	●	●	●	86%
	ADCS-CGIC	●	●	●	●	●	●	100%
	MMSE	●	●	●	●	●	●	100%
	ADAS-cog	●	●	●	●	●	●	100%
	CDR	●	●	●	●	●	●	100%
	ADCS-ADL	●	●	●	●	●	●	72%
	ADCS-CGIC	●	●	●	●	●	●	86%

- Trainers indicated 100% agreement on the Meets Criteria/Does Not Meet Criteria for 11 of the 15 assessments. Of those assessments in which there was less than perfect agreement (n = 4), Trainers still agreed between 72 - 100% of the time.
- In cases where 1 or 2 Trainers rated the assessment as "Does Not Meet Criteria," the rater did not meet the established minimum interview quality criteria and would participate in remediation training.

DISCUSSION

- To date, there has not been an adequate scale to measure AD rating scale administration quality. The MARQ was developed specifically to assess AD scale interview quality. In practice, the MARQ domains proved useful in rating AD interview quality.
- In this set of ratings, there is a high percentage of overall agreement with the Trainer's MARQ-AD scores in identifying interviews of good or poor quality.
- Practical aspects of the process were easily achieved. The trainers uniformly felt the technical aspects of the site recordings (e.g., quality of the video recording and transfer of drawings and source documentation) were sufficient for their rating purposes.
- In support of the overall goal of the MARQ-AD, the trainers identified rater mistakes that would have clearly degraded the accuracy of the score. These included errors of scoring, administration errors (skipping items, changing item order), inadequate follow-up questioning, and so forth. In the course of a clinical trial, these errors or omissions would be addressed by specific feedback to the rater.
- As employed during the course of a trial, each Trainer would continue their calibration training through periodic supervisor reviews of a Trainer's MARQ rating and site feedback with remediation as needed.

SUMMARY & CONCLUSIONS

- Until now, there has not been an instrument to measure the adequacy of raters administering AD scales. The MARQ-AD is a new tool to assess rater administration and scoring quality of AD clinical scales. The purpose of the MARQ-AD is to provide a discussion framework and metrics for feedback to raters and the sponsor during the conduct of a clinical trial.
- The three versions of the MARQ-AD are designed to work with structured (e.g., MMSE and ADAS-cog), semi-structured (e.g., CDR and ADCS-ADL) and unstructured scales (e.g., ADCS-CGIC). In this poster we discuss the training on the MARQ-AD and describe achieving required consistency in the MARQ-AD ratings. Recording ratings for later review is feasible.
- The benefit of the MARQ-AD to clinical trials in aging and Alzheimer's disease should be: improved interview standardization and quality, resulting in improved inter-rater reliability and decreased rater drift over the course of a trial.

References

1. Conner DJ, Sabbagh MN. Administration and scoring variance on the ADAS-Cog. J Alzheimers Dis. 2008 Nov;15(3):461-4.
2. Jeglic E, Kobak KA, Engelhardt N, Williams JB, Lipsitz JD, Salvucci D, Byson H, Bellew K. A novel approach to rater training and certification in multinational trials. Int Clin Psychopharmacol. 2007 Jul;22(4):187-91.
3. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. Am J Psychiatry. 2005; 162(3):628.

MARQ-AD – MedAvante Assessment of Rating Quality – Alzheimer's Disease

Author Disclosure Information: C. Weber, MedAvante, Tangcept; Worldwide Clinical Trials, Part 1; MedAvante, Part 2; MedAvante, Worldwide Clinical Trials, Part 3; MedAvante, Part 5; J. Williams, MedAvante, Part 1; MedAvante, Part 2; MedAvante, Part 3; MedAvante, Part 5; L. Kirby, Neurtus Pharmaceuticals, Part 1; Phoenix Neurology, Part 2; Neurtus Pharmaceuticals, Part 3