Are we adequately evaluating and monitoring rater performance in clinical trials with dementia?

Richa Gaur, PhD[1], Roger Bullock, MD[2], Geetika Nath, MA[1], Susan De Santi, PhD[3]

[1]TCG, Newark, United States, [2]Kingshill Research Centre, Swindon, United Kingdom, [3]New York University School of Medicine, New York, United States

Abstract

Introduction

The Alzheimer's Disease Assessment Scale-Cognitive Scale (ADAS-cog) is widely used in clinical trials of Alzheimer's disease. A rater training program to train and certify raters on the ADAS-cog was examined to determine the efficacy of the program and to identify the items having the greatest scoring difference compared with the Gold Consensus Ratings (GCRs).

Methods

188 raters from 65 sites across the US participated in a rater training program. Raters were trained and certified on the ADAS-cog through an interactive website. Three patient video sessions demonstrating the ADAS-cog were used for training, certification I and certification II. Raters viewed the videos and provided their ratings, which were compared to the GCRs of three AD experts. Items and total scores were analyzed.

Results

For the training session, the highest concordance rate (88%) was observed for "delayed recall". "Remembering test instructions", "orientation" and "constructional praxis" were the most difficult to rate, with concordance rates of 46%, 48%, and 52%, respectively. At the first certification session, highest concordance was found for "spoken language" (98%), with "ideational praxis" and "word recognition task" achieving concordance rates of 52% and 60%, respectively. In the certification session II, "commands" showed the highest concordance rate (92%), whereas ideational praxis" and "spoken language ability" were the most difficult to rate. Rater agreement with the GCRs was slight (kappa = .08, p < .001) for training and certification I (kappa = .17 p < .001), and fair (Kappa = .33, p < .001) for certification II.

Conclusions

Inter-rater agreement was slight to fair. Some items were difficult to score, showing low concordance rates with the GCR. It is possible that, for difficult to rate items, raters continue to commit errors throughout the study. Accurately rating a scale is the most important component of a clinical trial. To ensure accurate rating of all items, we propose that the audio and video recordings of a patient interview be independently rated by an expert rater. Using such a technique would allow constant evaluation and monitoring of the site ratings to ensure administration of high quality, reliable and unbiased ratings.

Introduction

The Alzheimer's Disease Assessment Scale-Cognitive Scale (ADAS-cog) is the most widely used primary outcome measure for clinical trails in Alzheimer's Disease.

Accurately assessing the clinical symptoms of AD is essential for the administration of the ADAS-cog. Therefore training and certifying raters in administrating the ADAS-cog is critical before the start of any clinical trial.

A rater training program to train and certify raters on the ADAS-cog was examined to determine the efficacy of the program and to identify the items having the greatest scoring difference compared with the gold consensus ratings (GCRs).

The Cognition Group
254 Chapman Road, Topkis Building, Suite 100,
Newark, DE 19702 USA
Tel: +1-302-454-1265, Fax: +1-302-454-7868
Toll Free: 888-454-1265

Method

188 raters from 65 sites across the US participated in the online (web-based) rater training and certification program.

The program was divided into two sections;

Section I

Training raters via interactive, web-based interactive web-based presentations which included a review of the general guidelines for administering the ADAS-cog and a review of the concepts and scoring conventions for each item.

Section II

Raters watched the videos of an expert rater administering the ADAS-cog to an AD patient and provided their independent ratings . Patients portrayed in the videos were professional actors. Three patients videos were used.

The first session was a mandatory training session. After completion of the training session, raters were directed to another mandatory session (certification I). Raters failing to certify were directed to the last and final session (certification II).

Raters who failed certification session II were excluded from rating patients in the trial .

Individual raters performance was compared to the GCR set by the experts. Kappa statistics were used to assess Inter-Rater Reliability.

Figure 1 Rater viewing the ADAS-cog administration online

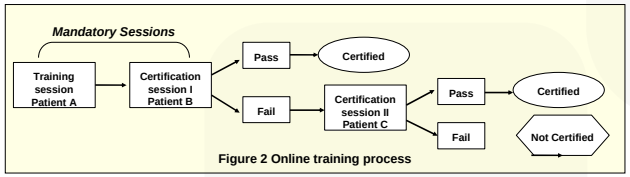

Figure 2 Online training process

Results

75% of raters passed the training session, 76% passed certification I, and 65% passed certification II.

The SD was highest for item 8, word recognition task (2.27, 1.70, 3.43) for all the three sessions indicating the variable scoring of this item.

The items with the greatest disagreement compared to the GCRs included "remembering test instructions" (46%) and "orientation" (49%) for the training session, "ideational praxis" (36%) for certification session I and "word finding difficulty" (18%),comprehension of spoken language (35%) and "remembering test instructions" (40%) for certification session II (Figure 1,2 & 3).

Rater agreement with the GCRs was slight (kappa = .08, p < .001) for training and certification I (kappa = .17 p < .001), and fair (Kappa = .33, p < .001) for certification II.

Table 1. Frequency distribution of rating responses in training session of ADAS-cog

	Item Label	GCR	Frequency distribution of ratings												%raters
			0	1	2	3	4	5	6	7	8	9	10	11	12	
1	Word recall task	7	0	0	0	5	2	0	6	86	10	1	2	-	-	76.79
2	Naming objects and fingers	2	1	27	75	9	0	0	-	-	-	-	-	-	-	66.96
3	Delayed recall	8	0	0	4	1	0	1	5	2	99	0	0	-	-	88.39
4	Commands	3	1	4	12	90	4	1	-	-	-	-	-	-	-	80.36
5	Constructional Praxis	2	5	44	58	04	1	0	-	-	-	-	-	-	-	51.79
6	Ideational praxis	2	1	7	94	10	0	0	-	-	-	-	-	-	-	83.93
7	**Orientation**	**4**	**0**	**4**	**5**	**19**	**54**	**26**	**04**	**0**	**0**	**-**	**-**	**-**	**-**	**48.21**
8	Word recognition task	12	0	1	0	1	1	2	1	9	4	3	4	12	74	66.07
9	Spoken language ability	0	70	28	11	2	1	0	-	-	-	-	-	-	-	62.50
10	Comprehension of spoken language	0	63	30	16	3	0	0	-	-	-	-	-	-	-	56.25
11	Word-finding difficulty in spontaneous speech	0	92	16	3	1	0	0	-	-	-	-	-	-	-	82.14
12	**Remembering test instructions**	**0**	**51**	**42**	**11**	**8**	**0**	**0**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**45.54**

Table 2. Frequency distribution of rating responses in certification session I of ADAS-cog

	Item Label	GCR	Frequency distribution of ratings												%raters
			0	1	2	3	4	5	6	7	8	9	10	11	12	
1	Word recall task	6	0	0	1	0	2	18	84	01	0	0	1	-	-	78.50
2	Naming objects and fingers	1	10	92	1	3	1	0	-	-	-	-	-	-	-	85.98
3	Delayed recall	6	0	0	0	0	2	1	102	0	0	0	0	-	-	95.33
4	Commands	1	5	97	2	0	3	0	-	-	-	-	-	-	-	90.65
5	Constructional Praxis	0	91	14	0	1	1	0	-	-	-	-	-	-	-	85.05
6	**Ideational praxis**	**2**	**12**	**55**	**38**	**1**	**0**	**1**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**35.51**
7	Orientation	1	3	75	26	0	0	0	1	1	1	-	-	-	-	70.09
8	Word recognition task	7	2	0	0	0	1	2	27	64	3	1	1	0	6	59.81
9	Spoken language ability	0	105	1	0	0	0	1	-	-	-	-	-	-	-	98.13
10	Comprehension of spoken language	0	95	11	0	0	0	1	-	-	-	-	-	-	-	88.79
11	Word-finding difficulty in spontaneous speech	0	101	5	0	0	0	1	-	-	-	-	-	-	-	94.39
12	Remembering test instructions	0	80	25	1	0	0	1	-	-	-	-	-	-	-	74.77

Table 3. Frequency distribution of rating responses in certification session II of ADAS-cog

	Item Label	GCR	Frequency distribution of ratings												%raters
			0	1	2	3	4	5	6	7	8	9	10	11	12	
1	Word recall task	9	0	0	0	0	0	0	0	1	6	14	2	-	-	60.87
2	Naming objects and fingers	3	0	0	3	19	0	1	-	-	-	-	-	-	-	82.61
3	Delayed recall	9	0	1	0	0	0	0	0	0	0	21	1	-	-	91.30
4	Commands	3	0	1	0	20	1	1	-	-	-	-	-	-	-	86.96
5	Constructional Praxis	3	0	1	3	19	0	0	-	-	-	-	-	-	-	82.61
6	Ideational praxis	2	0	0	14	3	7	0	-	-	-	-	-	-	-	60.87
7	Orientation	4	0	0	2	4	14	2	0	0	1	-	-	-	-	60.87
8	Word recognition task	12	0	0	0	3	0	2	0	0	0	0	1	2	15	65.22
9	Spoken language ability	1	0	17	2	1	2	1	-	-	-	-	-	-	-	73.91
10	**Comprehension of spoken language**	**1**	**9**	**8**	**1**	**3**	**2**	**0**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**34.13**
11	**Word-finding difficulty in spontaneous speech**	**1**	**14**	**4**	**2**	**2**	**0**	**0**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**17.39**
12	**Remembering test instructions**	**0**	**3**	**2**	**9**	**7**	**1**	**1**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**39.13**

Conclusions

In certification II the pass/fail rate was lower suggesting that raters who fail to certify on their 1st attempt, may require more extensive training to learn the ADAS-cog.

Word recognition obtained highest variability across all three sessions.

Concordance with the expert raters across the three sessions for individual items varied suggesting that ratings may be inconsistent during the clinical trial.

Only online training appeared to be insufficient to enable accurate ratings at trial initiation.

Other training models (i.e. live training at IM, rater administering the ADAS-cog) and methods during the clinical trial (i.e. repeated training) have been successful in other populations.

Further detailed assessment of administration skills prior to the start of the study, as well as continuous centralized monitoring of audio & video capture of patient interviews by independent raters, is likely to reduce the variance in CNS clinical trials.