Analytical validation of a novel quality assurance approach for COAs in Alzheimer's Disease clinical trials

The Methodological Issue

- CNS clinical trials rely heavily on rater-administered assessments as primary endpoints. Existing standards used in registrational trials, including the CDR, ADAS-Cog and others, comprise subjective clinical judgments and are complex to administer.
- These factors can introduce inconsistencies and errors in clinical ratings, which can impact signal detection.
- Current gold standard quality assurance practices for Clinical Outcome Assessments (COAs) are typically reliant on expert reviews, which are expensive, time consuming and practically limited to a small subset of COAs.
- There is a need for a more cost-effective, scalable, and reliable alternative to reduce measurement variance and ensure high quality administration.

Aims

This study aims to tackle methodological challenges in clinical trial assessments by validating a novel quality metric and COA quality review process using retrospective audio recordings of Clinical Dementia Rating (CDR) assessment.

Table 1. Major Quality Index Calculation

Metric	Description	Weighting
Rater interjection	Does the clinician or other speaker interfere with the task? 0 - None, 1 - Minor, 2 - Major	0.75
Rater clarity	Is the clinician's voice hard to hear or understand? 0 - Excellent, 1 - Somewhat unclear, 2 - Often unclear	0.5
Speaker number	Are there more speakers than required? 0 - No, 1 - Yes	1
Skipped prompt	Is the prompt skipped? 0 - No, 1 - Yes	0.75
Out of order prompt	Is the prompt out of order? 0 - No, 1 - Yes	0.25
Repeated prompt	Is the prompt repeated? 0 - No, 1 - Yes	0.25
Prompt deviation	What is the deviation level of the prompt? 0 - No deviation from script, 1 - Minor deviation, 2 - Major deviation	0.75
Reworded prompt	What is the level of the repetition rewording? 0 - No repetition, 1 - Minor rewording, 2 - Major rewording	0.5
Task skipped	Is the task skipped? 0 - No, 1 - Yes	1
Out of order task	Is the task out of order? 0 - No, 1 - Yes	0.5
Disjointed task	Is the task administration disjointed, i.e., task was partially completed and revisited after another task was started? 0 - No, 1 - Yes	0.5

Table 1 shows the different metric definitions, descriptions, and weighting. These values were summed to create a single Major Quality value.

Methods

73 recordings of the CDR were reviewed by skilled clinical experts. Experts were asked to flag instances where the rater deviated from the COA guidelines of administration, including rater interjection, rater clarity, skipped tasks, number of speakers, and deviations from prompts. Together, quality flags were summated into a weighted composite score called the Major Quality (MQ) index. See **Table 1** for MQ calculations.

The same 73 recordings were then assessed by our novel proprietary quality review pipeline, AQUA, which utilizes a team of non-expert, human reviewers who had received training to identify the same COA guideline deviations flagged by experts. These human reviewers compared the text transcript of the audio recording with a highly defined schema of the CDR instructions, along with all of the core features comprising the MQ index. Previous work with other COAs had shown a high degree of concordance between experts and our pipeline in flagging of administration variances[1,2]. MQ scores were generated for both expert raters and AQUA. The degree of concordance between the two were assessed using correlational analyses.

Table 2. Correlation Values of Expert Quality Ratings and AQUA

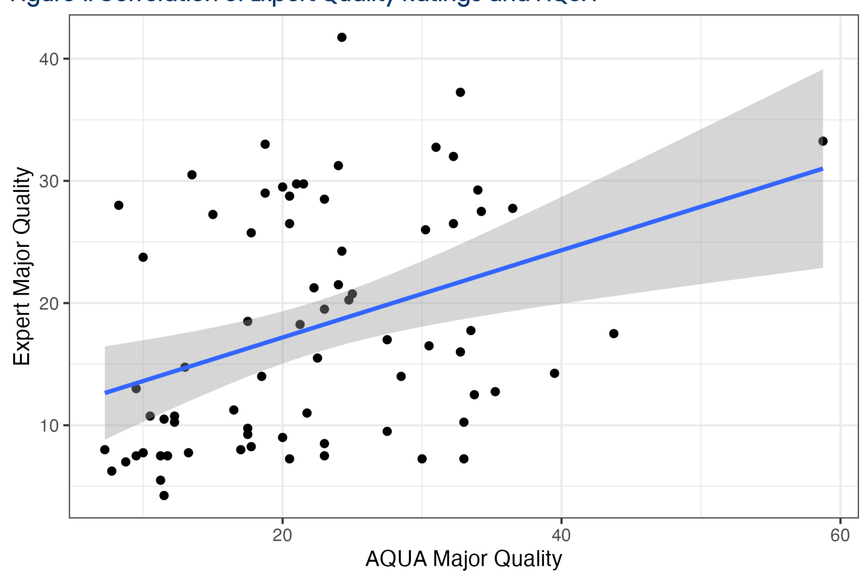
Metric	Spearman's Rho
Rater interjection	0.22
Rater clarity	**0.044**
Speaker number	1
Skipped prompt	1
Out of order prompt	1
Repeated prompt	0.55
Prompt deviation	**0.059**
Reworded prompt	0.235
Task skipped	1
Out of order task	1
Disjointed task	1

Table 2 shows correlation values between expert raters and AQUA on major quality metrics.

Table 3 shows example descriptions from real assessments flagged by AQUA and expert raters

Figure 1 shows a correlation between Expert Major Quality and AQUA Major Quality, r = 0.40, p<0.001.

Figure 1. Correlation of Expert Quality Ratings and AQUA



Results

A Spearman's correlation was run to determine the relationship between expert reviewers and non-experts using our pipeline to identify quality issues. There was a positive correlation (r= 0.40, p<0.001) of MQ index scores, suggesting that assessments flagged by the system for major quality concerns have a degree of overlap with those highlighted by expert raters. Notably, the pipeline successfully flagged instances where a rater inappropriately provided answers to a patient and identified deviations such as unsolicited hints and deviant administration. Follow-up analyses found moderate to high positive correlations (r's=0.23 - 1, **Table 2**) for all quality flags, with the exception of prompt deviation (r=0.06) and rater clarity (r = 0.04).

Table 3. Example Flags

Flag	Description
Rater gave unsolicited hints	In recent memory task, rater prompted with four different hints
Rater gave the answer	In name/address task, participant asked for name and rater gave it
Rater encouraged participant to guess	In orientation task, after participant expressed they didn't know, rater prodded 9 times for guesses

Conclusions

Our quality review pipeline showed good alignment with expert reviewers for several quality indicators. By leveraging non-expert reviewers and technology, via an in depth review platform, the process of reviewing COAs can scale more effectively. Natural Language Processing (NLP) methods can automate the detection of nuanced speech patterns. We have also previously demonstrated[3] that speech features, particularly those related to speech complexity, like graph features, correlate strongly with MQ index. By flagging quality issues through our review pipeline along with automated speech processing, NLP methods can significantly reduce the time and cost associated with manual reviews, while ensuring a higher level of consistency and accuracy in identifying clinical trial quality issues. Further development and refinement of these quality review practices is ongoing.

References

1. Kindellan, Newsome, Fidalgo, Robin (2023). ISCTM. Assessments of ADAS-Cog administration quality are comparable across expert and non-expert reviewers
2. Kindellan, Sirotkin, Xu, Fidalgo, Simpson, Robin (2022). CTAD. Accuracy of automated scoring of word recall assessments
3. Newsome, Kindellan, Fidalgo, Simpson (2024). ISCTM. Towards fully automated rating scale review: Identifying speech feature signatures for administration variances in CDR interviews during Alzheimer's Disease clinical trials

Authors

Newsome, RN, Kindellan, R, Fidalgo, C, Simpson, W.

rachel.newsome@camcog.com
All authors are employees of Cambridge Cognition

CAMBRIDGE COGNITION