Accuracy of automated scoring of word recall assessments

Rachel Kindellan [1], Sasha Sirotkin [1], Mengdan Xu [1], Celia Fidalgo [1], William Simpson [1], Jessica Robin [1]

(1) Winterlight Labs, Toronto, ON, Canada

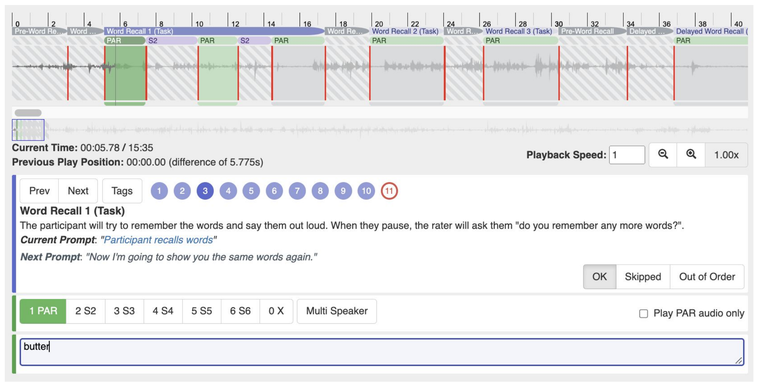
Background

- Natural language processing (NLP) tools can be used to automate and standardize the scoring of clinical assessments.
- Many cognitive assessments used as endpoints in Alzheimer's disease (AD) trials require manual scoring and review which can be costly and time consuming.
- Developments in natural language processing technology can be leveraged to develop automated and objective tools to generate text transcripts for cognitive assessments. These text transcripts can then be used to extract clinical scores for simple, quantifiable tasks, such as word recall.
- As a proof of concept, we tested an automated method to score the four word recall portions of the ADAS-Cog, a standard endpoint in AD research.
- The objective of this study was to evaluate the accuracy and feasibility of using this automated method in cognitive assessment scoring.

Methods

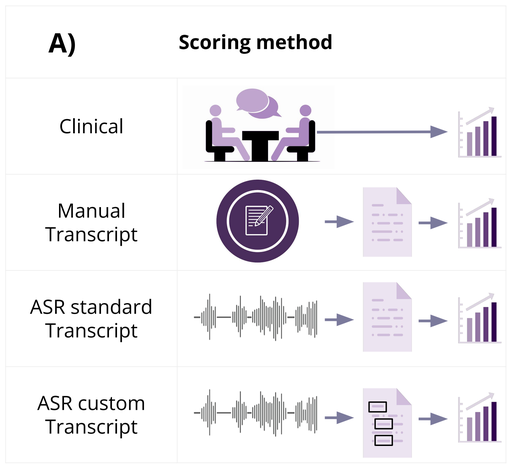
- 274 word recall trials from 70 audio recordings of the ADAS-Cog were collected from 54 older adult volunteers (four trials per ADAS-Cog).
- In the word recall subsection of the ADAS-Cog assessment, the participant is presented with a list of 10 words (of two possible lists), asked to read them aloud upon presentation on cards, then asked to recall the words (word recall). This is repeated for three trials. Later in the same assessment, the participant is asked to recall the words again without presentation (delayed recall).
- Each audio recording was manually split into the four word recall subsections. The non-word recall segments were not analysed.
- Three transcripts were generated for each word recall task. Two transcripts were generated using automatic speech recognition (ASR) software from Amazon Web Services (AWS) Transcribe. The first transcript was generated using a standard AWS ASR (ASR) model while the second was generated using an AWS ASR custom model (ASR custom) that was pre-configured based on the expected ADAS-Cog word list. The third transcript was generated manually by trained human raters.
- A word recall score was automatically calculated based on each transcript (ASR custom, ASR standard, and manual). These scores were computed by counting how many words per trial were correctly recalled (score out of 10 total). Word recall scores derived from both ASR transcripts, manual transcripts, and clinical scores were compared to evaluate agreement using intraclass correlations and by comparing mean values.
- The impact of scoring method and word list was assessed on the word recall score.

Figure 1: Transcription platform



Transcription platform used to segment and transcribe the word recall segments of the ADAS-Cog

Figure 2: Scoring methods pictogram and word lists



B)

Word List A	Word List B
Butter	Bottle
Arm	Potato
Shore	Girl
Letter	Temple
Queen	Star
Cabin	Animal
Pole	Forest
Ticket	Lake
Engine	Clock
Grass	Office

Figure 3: Heat map of pairwise intraclass correlations on scores from ASR, manual transcripts and clinicians

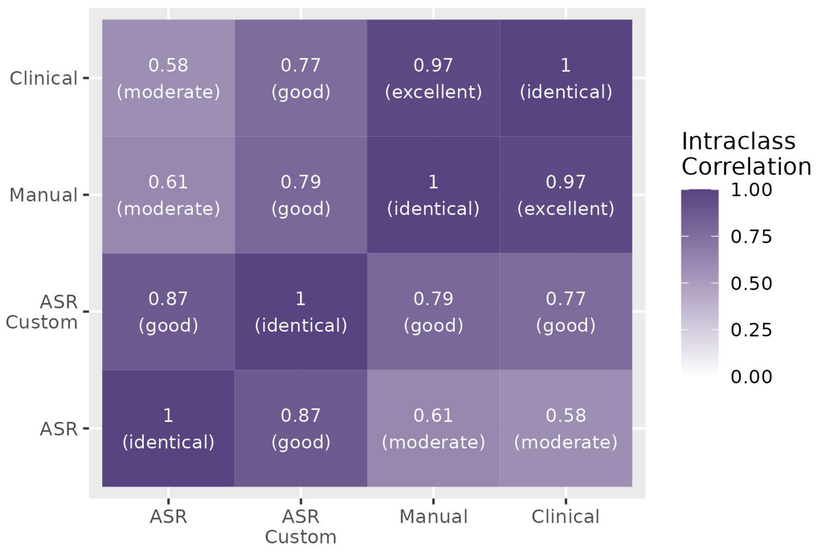


Figure 4: Boxplot of score distributions from ASR transcripts, manual transcripts, and clinicians

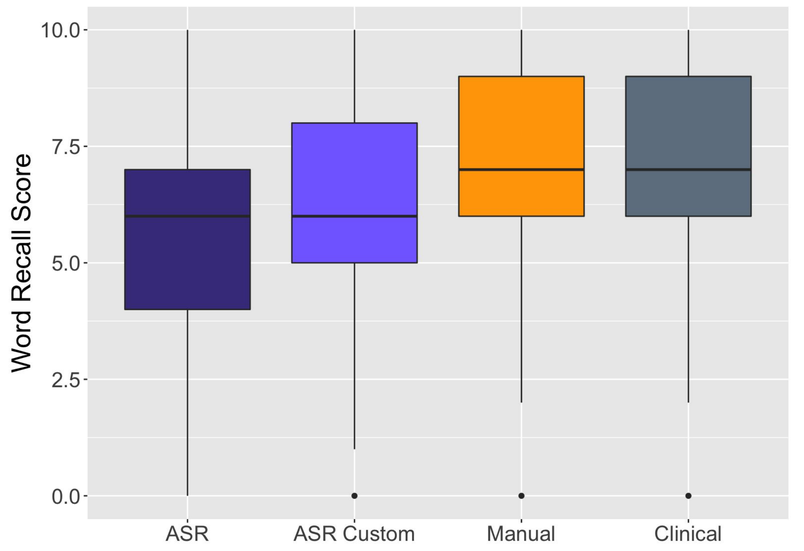
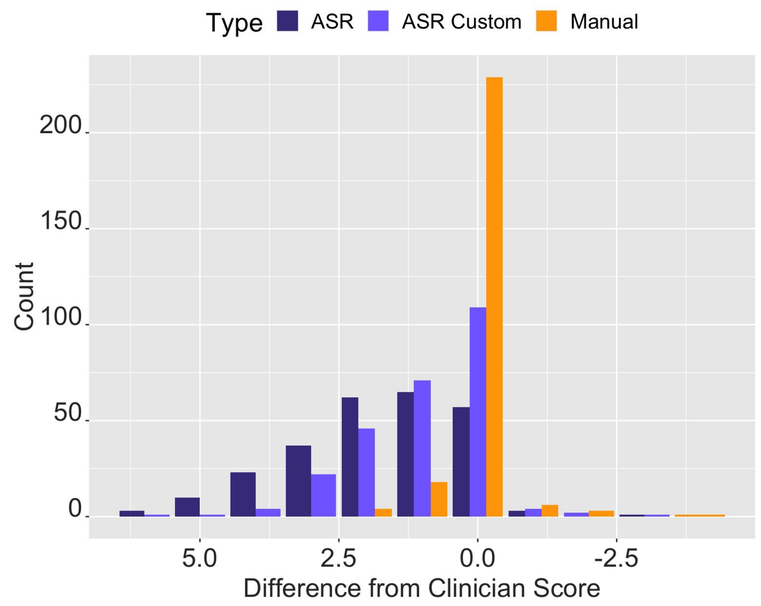


Figure 5: Histogram of score agreement by transcription method compared to clinical score



Results

- Overall, all scoring methods had moderate to excellent agreement with one another. Manual scoring had excellent agreement with clinical scores (ICC = 0.97), while ASR customized for the expected word list had good agreement (ICC = 0.77) and standard ASR had moderate agreement (ICC = 0.58).
- A repeated-measures ANOVA indicated a significant difference between scoring methods (clinical, manual, ASR, ASR custom; $F = 229$, $p < 0.0001$). Follow up pairwise comparisons indicated that there were significant differences between both ASR scoring methods and clinical and manual scoring (all p's < 0.0001).
- There was a significant difference between custom and standard ASR scoring (mean difference = 0.80, $p < 0.0001$).
- There was no significant difference between clinical and manual scoring methods (mean difference = 0.04, $p = 0.62$).
- Automated scoring tended to underestimate scores compared to the human raters, with average scores of 5.6 (SD = 2.3) for standard ASR and 6.4 (SD = 2.2) for custom ASR, compared to average scores of 7.3 (SD = 2.1) for manual scoring and 7.3 (SD = 2.1) for clinical scoring.
- Additionally, there was a significant interaction with word list, with lower overall scores on word list A (mean score = 6.0) compared to word list B (mean score = 7.6), with automated methods in particular having significantly lower scores for word list A (mean score ASR = 4.7, ASR custom = 5.7) than word list B (mean score ASR = 7.0, ASR custom = 7.6).

Conclusions

- In this study, we found that automated scoring methods reached moderate to good agreement with standard clinical scoring of a word recall task, although they tended to underestimate scores due to transcription errors. Comparatively, manual scoring methods reached high accuracy compared to clinical scores.
- Notably, customization of the ASR scoring method led to improved performance and better agreement with manual scoring.
- Automated NLP tools show promise for quantitative scoring in cognitive assessments thereby increasing the efficiency of quality assurance and review of clinical assessments conducted as part of trials.
- Future work to refine the use of ASR to evaluate clinical endpoints includes: optimizing ASR accuracy by improving noise filtering and diarization, further customizing language models, and exploring the accuracy of ASR in different contexts, such as different tasks or word lists, to understand the effect of speech content.
- Further research and development of ASR and NLP tools will improve the accuracy of automated scoring and quality assessment methods, making them more efficient and scalable for use in clinical research.

Presented at Clinical Trials for Alzheimer's Disease (CTAD), San Francisco, CA, USA. Nov 29 - Dec 2 2022

Contact: jessica@winterlightlabs.com