The development and validation of an automated secondary review model for clinical endpoint interviews

Submitter Michelle Worthington

Affiliation Brooklyn Health

SUBMISSION DETAILS

I agree to provide poster pdf for attendee download. Yes

What is the Methodological Issue Being Addressed? Traditional secondary review methods in CNS clinical trials, including manual audio reviews, are standardly used to mitigate rater drift and scoring subjectivity. However, these approaches are resource-intensive, time-delayed, expensive, and structurally unscalable, which often prevents comprehensive, real-time quality control across an entire study. There is an operational need for a framework that provides objective, independent interview evaluations directly from clinical interview recordings without introducing additional data collection burden.

Introduction Rater variability and subjective scoring compromise data quality and can inflate placebo response rates in CNS trials. While traditional secondary review offers a partial check against these errors, it cannot scale to evaluate every recorded interview. LLM-based AI models can bridge this gap by reading through transcripts to understand the context of the conversation and evaluate the participant's responses using standard scoring guidelines. Here, we introduce and aim to validate an automated LLM-based secondary review model for reviewing and scoring endpoint interviews in CNS clinical trials. Validation for the score prediction model was conducted in an open-label observational study in depression. We evaluated the automated scoring pipeline by comparing AI-generated scores against human ratings and identifying discrepancies to detect potential scoring anomalies.

Methods Audio recordings from 609 Montgomery-Åsberg Depression Rating Scale (MADRS) MADRS assessments from an open-label study in depression were processed with an automated pipeline that conducts speech transcription and speaker diarization. Transcripts were then segmented into item-level interview segments based on the Structured Interview Guide for the MADRS (SIGMA). Each segment was then passed to an LLM (Claude Sonnet 4.5), which scored each item from 0 to 6 based on the interview content and scale guidelines; a total score was then calculated by summing these individual item scores. Agreement with human ratings was evaluated using Pearson's R, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Bland-Altman plots. A subset of interviews were also reviewed by a human secondary reviewer to verify score alignment. Finally, score discrepancies were analyzed across visits and raters to detect systematic scoring discrepancies.

Results Rater-assigned and AI total scores demonstrated strong aggregate correlation globally (Pearson's $R = 0.93$, $p < 0.001$, $RMSE = 4.55$). This pattern of results was also observed longitudinally across visits and when stratified by individual clinical raters. Strong agreement was

found at the item level; however, specific items requiring clinical inference were slightly less aligned. WillisAI and human secondary reviewer scores also showed high alignment with each other and with rater scores in a subset of interviews (n = 10, WillisAI MAE = 4.6, human secondary reviewer MAE = 3.8, W = 18.5, p = 18.5).

Conclusion Automated secondary review using LLM-based models provides a scalable, low-burden method to independently score clinical interviews from existing audio records. By demonstrating strong concurrent alignment with human scores while highlighting individual item-level and rater-level deviations, this pipeline establishes the feasibility of using LLMs as continuous, automated data-auditing mechanisms to safeguard data quality in CNS trials.

Co-Authors

Michelle Worthington[1], Jacob Ouyang[1], Kieran McVeigh[1], Miriam Evans[2], Vijay Yadav[1], Anzar Abbas[1]

[1] Brooklyn Health

[2] Adams Clinical Site Network

Keywords

Keywords
AI in clinical trials
secondary review
data quality monitoring

Guidelines I have read and understand the Poster Guidelines

Disclosures MW, JO, KM, VY, and AA are all current or former employees of and hold equity in Brooklyn Health, a private company that developed and indirectly benefits from the model described in this abstract. ME is an employee at Adams Clinical Site Network, where data was collected for the validation study.

Related Tables and Supporting Materials <blank>