From Clinician Training to Scale-Specific Conventions: A Novel Approach for LLM-Based Rater and Data Quality Control

Submitter Daniel DeBonis

Affiliation EMA Wellness

SUBMISSION DETAILS

I agree to provide poster pdf for attendee download. Yes

I have used the poster abstract template to develop my abstract. Yes

Methodological Issue Being Addressed A common data quality control strategy in CNS trials involves recording primary outcome interviews and generating a second set of scores from an independent rater. This strategy has expanded to include Artificial Intelligence (AI) and Large Language Models (LLMs) as independent raters, forming a core component of EMA Wellness' eCOA and data-surveillance platforms. The objective is to train models to produce ratings that exceed human accuracy and to support immediate, targeted training. Convergence statistics index the success of this effort. A key challenge is that many scale items have low mean scores—often below 1.0—and represent non-central illness features, complicating convergence analyses. This presentation examines AI-human convergence using item-by-item absolute agreement and compares these findings with traditional ICC models. We also directionally quantified discrepant scores to determine whether the AI rater detected symptoms missed by human raters.

Introduction Independent interview scoring has long been used to improve data quality in CNS trials. The use of AI and LLMs extends this practice by providing standardized, reproducible second ratings capable of flagging scoring or administration concerns. LLM-generated scores can highlight potential inconsistencies and offer real-time insights into rater performance. EMA Wellness' data-surveillance framework integrates trained LLMs that apply scoring conventions across CNS scales. Despite high overall convergence, items with low scores or limited clinical salience present methodological challenges. This study focuses on refining analytic approaches for evaluating agreement between LLM-generated ratings and human raters.

Methods Most existing LLM models have been trained using scale-specific prompts and structured inputs. The current data were generated while developing an LLM capable of producing symptom ratings across therapeutic areas and applying standardized scoring conventions. A private GPT-4o model hosted on Microsoft Azure was trained using interview-skills and placebo-response methodologies similar to those used for human raters. Additional expert clinical inputs on assessing severity and frequency in interviews were incorporated. Training materials included a SIGMA rating guide and a MADRS program featuring the SIGMA. Fifty audio-recorded MADRS interviews from late-phase studies were transcribed. The model rated each item using detailed scoring rationales based on transcriptions. LLM scores were compared with interviewer scores using intra-class correlations (ICC) for total scores and item-level absolute agreement statistics. Discrepancies were directionally evaluated to determine whether the AI rater tended to score higher or lower than

human raters.

Results Paired rater and LLM scores were analyzed. The ICC for MADRS total scores was .78. The mean total score for human raters was 14.2 versus 15.8 for the LLM. Item-level absolute agreement ranged from 57% to 98%, with a mean of 70%. Overall, 91% of all ratings were within one point. In cases of disagreement, the LLM produced higher scores 66% of the time.

Conclusion Validated LLM models enable real-time monitoring of rater biases, interview quality, and anomalous scoring patterns. Agreement between AI and human raters varies substantially by item, particularly for low-scoring or non-central features (e.g., Inner Tension: 57% agreement vs. Suicidal Thoughts: 98%). Identifying such patterns allows for targeted escalation and training to improve scoring consistency across all items. As trained LLMs incorporate additional scales and develop greater familiarity with difficult-to-rate items, their clinical interpretive accuracy should continue to advance.

Co-Authors

Daniel DeBonis[1], Suzanna Newton[1], Andrew Culter[1], Stephen Brannan[2], Phil Harvey[3]

[1] EMA Wellness

[2] CNS Clinical Consulting, Boston, MA

[3] University of Miami Miller School of Medicine, Miami, FL

Keywords

Keywords
Rating quality
AI in Clinical Trials
Analytic approaches

Guidelines I have read and understand the Poster Guidelines

Disclosures DeBonis, Edman, Cutler are full time employees of EMA Wellness. Other authors are affiliated with EMA Wellness and the institutions listed