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

Daniel V DeBonis[1], Suzanna Newton[1], Stephen Saber[1], Stephen Brannan[1,2], Philip D. Harvey[1,3]

1. EMA Wellness, Boston, MA, 2: CNS Clinical Consulting, Boston, MA, 3. University of Miami Miller School of Medicine

Methodological Issues 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.

Background:

- 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.

Study Over-View:

- 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.

Methods:

- Fifty audio-recorded MADRS interviews from late-phase studies were transcribed.
- All participants were evaluated at a screening assessment in a clinical Trial.
- All raters were trained site raters and not specialized central raters and interviews were performed in person.
- AI assessments were performed after the conclusion of the screening period, and no feedback was provided to raters.
- The model rated each item using detailed scoring rationales based on transcriptions.
- Individual item scores were rated, which were then summed into a total score.

Statistical Methods and Goals

Analysis strategies

- LLM scores were compared with interviewer scores.
- Total scores were examined with intra-class correlations (ICC).
- item-level analyses included statistics addressing absolute agreement.
- Additionally, rater agreement was examined according to the severity level of the items .
- Discrepancies were evaluated tin terms of directionality to determine whether the AI rater tended to score higher or lower than human raters.
- Correlational statistics were computed between rater and AI scores for each item.

Results:

- The ICC for MADRS total scores was .78 across site raters and AI.
- 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%, (Figure 1) with a mean of 70%.
- Overall, 91% of all ratings were within one point (Figure 2).
- In cases of disagreement, the LLM produced higher scores 66% of the time.
- Items with higher scores had higher levels of agreement between AI and rater.
- Correlations between AI and rater scores were consistently high (Figure 3), and affected as expected by item severity

MADRS Item Names and Codes

- 1. Apparent Sadness
- 2. Reported Sadness
- 3. Inner Tension
- 4. Reduced sleep
- 5. Reduced Appetite
- 6. Concentration problems
- 7. Lassitude
- 8. Inability To Feel
- 9. Pessimistic Thoughts
- 10.Suicidal Ideation

Figure 1
Item x item agreements: Rater vs AI

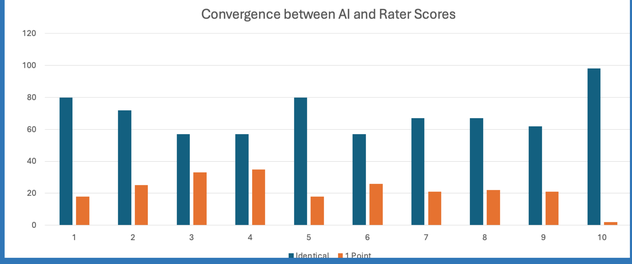


Figure 2 MADRS Item Mean Rater Scores and Differences with AI scores: All Differences are positive in Valence

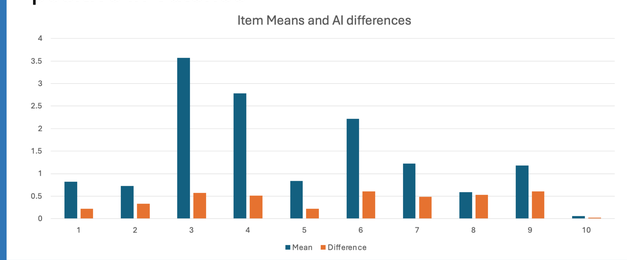
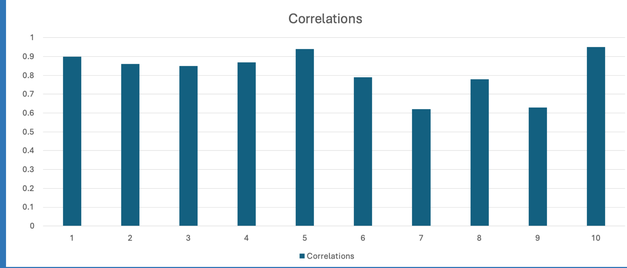


Figure 3: Pearson Correlations between AI and Rater Scores Item x Item



Conclusions:

- AI ratings are very convergent with site raters, with some differences in convergence associated with severity
- Higher-severity items are more convergent across ratings
- Our previous experience suggests that there is even greater convergence with highly trained central raters and the AI ratings
- These data suggest that ratings in acute treatment studies could be expected to similarly high convergence