Standardizing Audio and Video Capture to assess biomarkers in Clinical Trials – Avoiding Garbage in/Garbage out

Suzanna Newton[1], Patrick Harrington[1], Andrew Cutler,[1] Philip D. Harvey,[1,2] Daniel DeBonis[1]

1. EMA Wellness, Boston, MA, 2: University of Miami Miller School of Medicine

Methodological Issues Addressed:

- Clinical interviews are often recorded so experts and now LLMs can analyze for scoring and biomarkers. Recording quality is variable. **We wanted a simple, scalable way to measure how usable these recordings really are.**
- Use cases for audio recordings are centralized rating, quality control and digital phenotyping in psychiatric research. Recording quality varies substantially across environments (site vs teleconference, site-by-site, rater-by-rater), potentially affecting interpretability. **Objective, scalable methods for evaluating recording usability are needed.**
- Poor audio quality is a limitation of the reliance of large language models (LLMs) that utilize audio transcription for review/scoring, and for the effective use cases for biomarker analysis. **Can audio recordings be used reliably for evaluation and analysis?**

Background:

- At the ISCTM Biomarker Working Group (Amsterdam, 2025), the lack of unified standards for capturing audio and video during CNS clinical interviews was noted. This applies to biomarker collection and the growing use case of LLM scoring analysis for standardized rating scales.
- Developing a standardized protocol for collecting audio and video data across clinical sites, quantitative biomarker analyses, including voice, speech, and non-verbal cue would benefit the industry.
- The use cases and publications to date were in controlled environments, not real-world multi-center (and multi-lingual) clinical research studies.

While video is an important consideration, this poster focuses on audio

Methods, Part 1:

Common causes of recording failure (e.g., distortion, framing issues, and variable signal-to-noise ratios) were considered alongside requirements for reliable speech analysis and transcription*. For Part 1, a standardized audio setup was used:

- Microphones with flat frequency response (±3 dB, 50–5000 Hz) to preserve clinically relevant vocal features (speech rate, prosody, pauses)
- Commercial tabletop microphones under $100 that meet these specs and are compatible with eCOA devices
- A brief pre-interview verification check, including frequency response confirmation

These steps can be integrated into routine visit workflows.

*Prior work from the Bridge2AI-Voice Consortium demonstrated that voice recordings meeting basic acoustic criteria are sufficient for biomarker analysis. However, some recommendations from that research were not practical for clinical research (i.e., using headsets and lavalier mics with study participants)

Methods:

Objective 2: A rubric was developed to mirror the 0 – 1.0 'confidence' scores assigned by LLMs for interpretability of audio and scoring of structured interviews such as the MADRS, HAM-A and PANSS. Recordings from four sources were analyzed:

- Control: reference-grade structured interviews for rater training (n=3)
- Test: recordings using a standardized audio/video collection equipment with variables, such as participant distance from microphone (n=10)
- Site: site interviews; prior analysis set (n=20)
- Remote: web conference interviews; prior analysis set (n=10)

The site and teleconference interviews had already been part of the analysis in an ISCTM Fall 2025 poster. A weighted audio confidence score (0–1.0) was computed from speech presence, signal stability, and technical viability, as assessed by the same LLM that was utilized for scoring analysis in previous research presented at ISCTM Fall 2025. A box plot distribution of recording confidence was generated.

Rubric Overview:

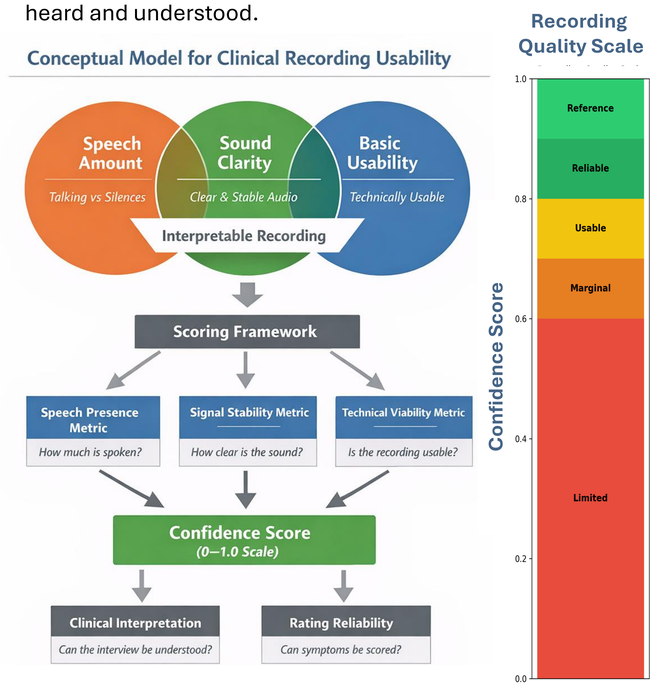
A Clinical Audio Confidence Framework, designed to focus on clinical interpretability rather than raw signal analysis.

Component	Weight	Description of factors
Speech Presence	50%	**Overall quality of speech:** Speech could be clearly heard above background noise. Criterion: Signal-to-Noise Ratio, with speech at least 1.5x louder than background noise
Signal Stability	30%	How clear and steady the sound is (not distorted or uneven, doesn't jump or cut out); responses are audible and complete. Criterion: Stability pattern based on dB variation (±3–6 dB as acceptable)
Technical Viability	20%	**Usability of Recording:** Clear enough to be reviewed and interpreted. Recording is complete, with limited gaps, missing words; speech is audible relative to environment. Criterion: 50% above average signal level in recording

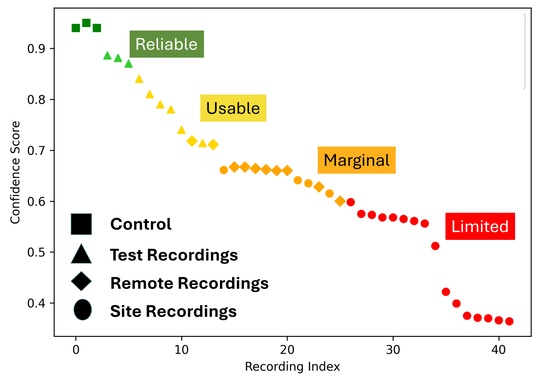
.

Rubric Model:

The conceptual model of the rubric is shown on the left figure. A recording quality scale was created on the rubric output, The ranges were set to represent practical levels of recording usability, based on how clearly speech could be heard and understood.

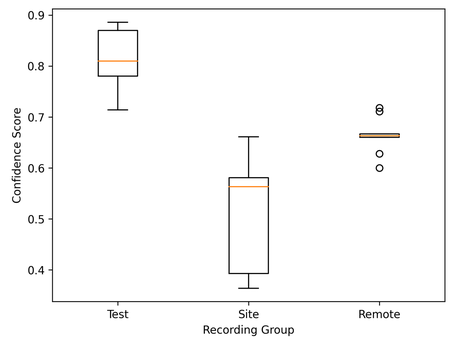


Recording Confidence by Group



- The test recordings using standardized audio equipment had 9 of the top 10 confidence scores (.81 mean)
- Site recordings had the lowest overall confidence score (.51), with all but two 'limited'
- Remote interview recordings averaged as marginal (.66)

Distribution of Recording Confidence



- Site and remote groups show different variability patterns, possibly indicating different sources of recording inconsistency.
- Site recordings were the largest group and had the most variability; not surprising given the higher number of potential variables at sites.

Conclusions:

- There are inherent challenges in a multi-site trial, with variations in raters, participants and environments.
- A rubric to create a measurable scale can be an essential tool to establish reliability and interpretability of LLM scoring and biomarker analysis.
- The test controls with standardized equipment and training, such as positioning microphones, showed higher recording confidence. This approach would increase reliability and confidence in results.

This research was funded by EMA Wellness.