Benchmarking Commercial and Open-Source Speech AI for Speaker Attribution in Real-World Clinical Conversations

C. Bruzinski, V. Palat, A. Cerezo, D. Schlosser, & S.P. Lord

mpathic, Bellevue, WA

The Problem

In clinical trials, capturing not only *what was said* but *who said it* is critical for fidelity, safety, and downstream analytics. Traditional evaluation methods for Automatic Speech Recognition (ASR) often stop at role identification—assigning broad labels like "clinician" or "patient" to speaker segments. But in real-world clinical settings, this coarse granularity is insufficient.

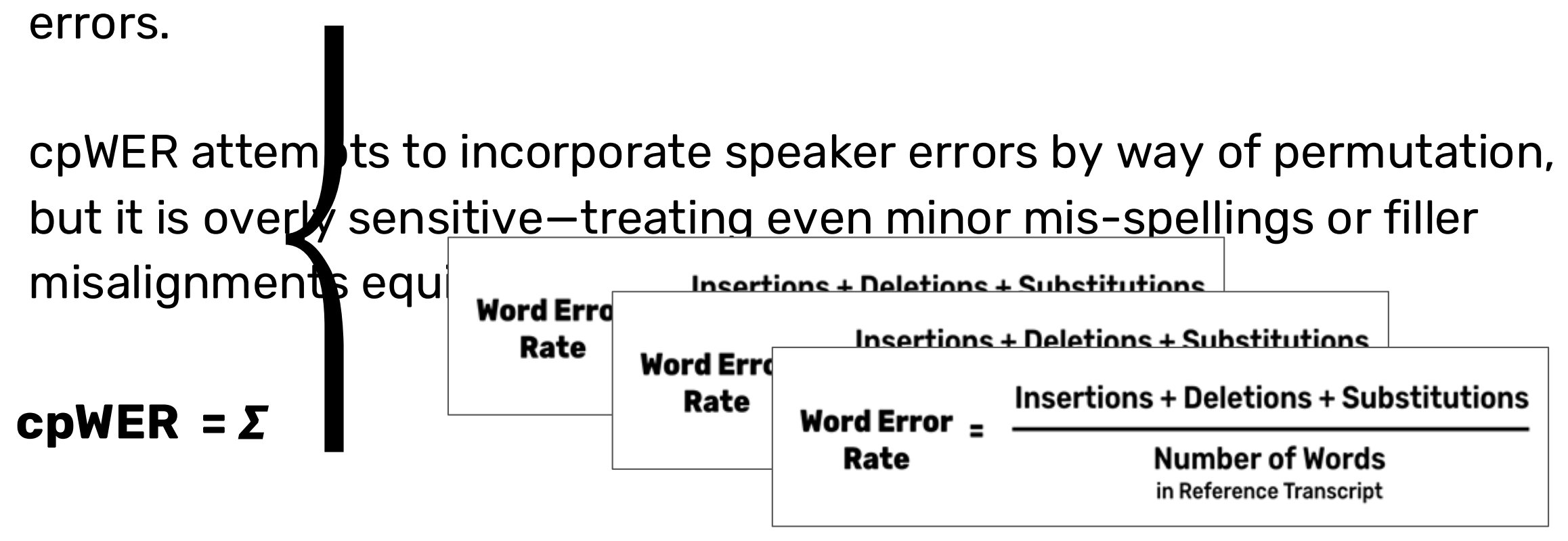
Our work advances beyond role identification to the more demanding, clinically essential task of speaker attribution, **wherein each utterance is both transcribed and correctly assigned to the proper speaker.** This allows precise behavioral and safety analytics at the utterance level, which is vital when even small attribution errors can mislead study outcomes.

To support this approach, we introduce cpHEWER (Clinician-Preferred Human-Evaluated Word Error Rate)—a novel evaluation metric that embeds speaker awareness into transcription accuracy. cpHEWER is designed to weight clinically meaningful errors (e.g., misattributed or misrecognized utterances) more heavily than common filler-word mistakes.

In this poster, we also address the compounding risk of dual errors—instances where both the transcription and its speaker attribution are incorrect. These "double faults" are especially dangerous in clinical contexts and should receive stronger penalization in benchmarking.

While modern ASR systems can generate diarized transcripts, few are rigorously validated in multi-speaker clinical dialogues, especially in the presence of accents, varying speech rates, and domain-specific terminology. The field continues to rely on metrics such as WER (Word Error Rate) and cpWER (concatenated minimum permutation WER), which, despite their prevalence, overlook or distort the clinical impacts of attribution errors.

- WER measures raw transcription fidelity (i.e., the "edit distance" between system output and ground truth) but remains blind to speaker errors.
- cpWER attempts to incorporate speaker errors by way of permutation, but it is overly sensitive—treating even minor mis-spellings or filler misalignments equi



$$\text{cpWER} = \Sigma \left\{ \text{Word Error Rate} = \frac{\text{Insertions + Deletions + Substitutions}}{\text{Number of Words in Reference Transcript}} \right.$$

In our study, we benchmark several leading ASR systems using naturalistic clinical dialogues collected in trial-like settings, assessing how these systems handle clinician/patient audio under realistic conditions, including accent variation. Our goal is to inform clinical teams about which ASR systems offer the most robust performance when integrated into high-stakes clinical workflows—and to guide future tool selection.

Experimental Setup

Systems Evaluated

To assess real-world ASR performance in clinical contexts, we benchmarked **three state-of-the-art transcription systems**—two commercial and one open-source:

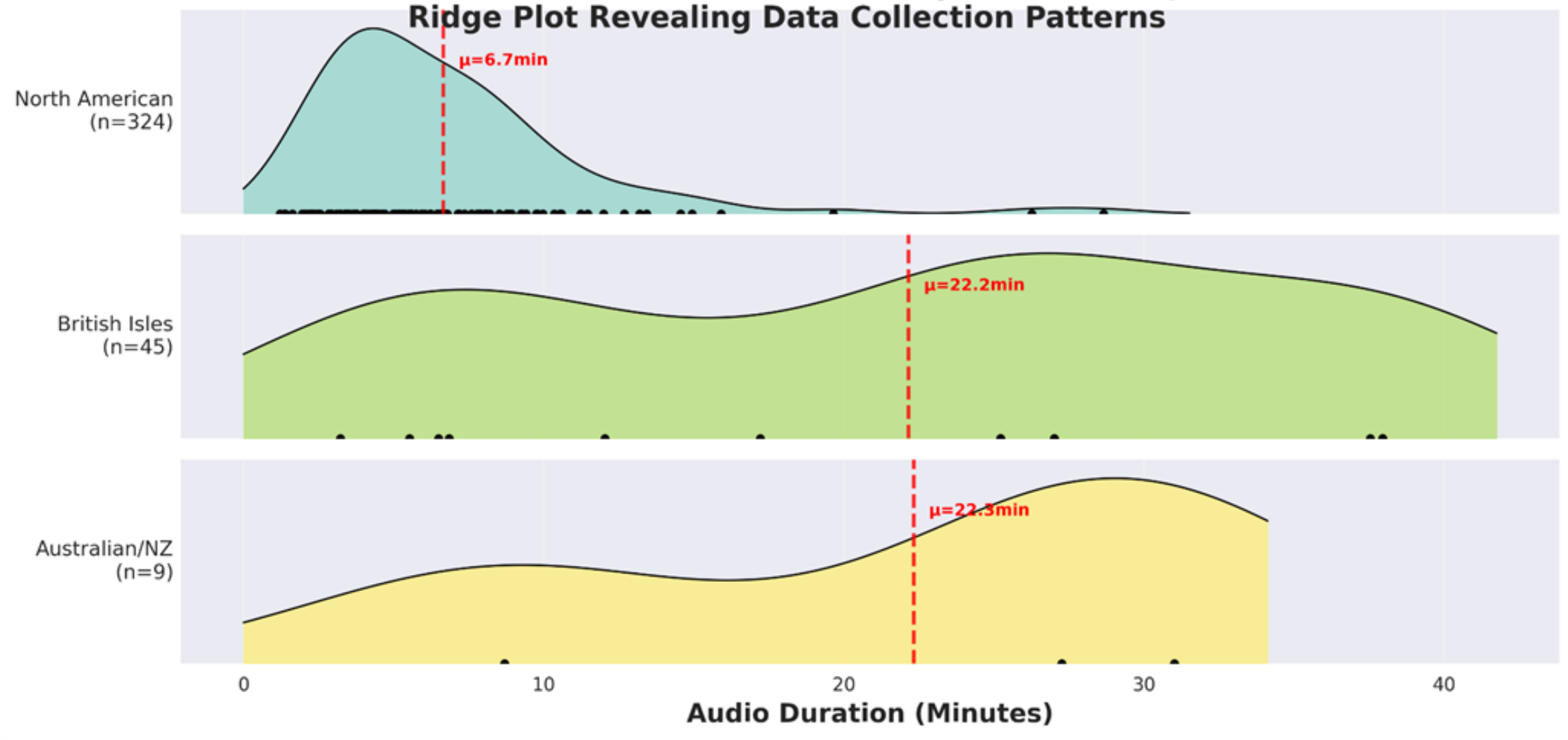
- **Rev AI** *(Commercial)*
- **Assembly AI** *(Commercial)*
- **Whisper Medium + Pyannote** *(Open Source; OpenAI Whisper for transcription, Pyannote for speaker diarization)*

Each system represents a leading approach in diarized transcription, and all are currently used or under consideration for clinical deployment. Our comparison focused on how well each system handled speaker attribution, transcription accuracy, and performance under noisy or accented conditions typical of clinical trial settings.

Dataset

We used the **AnnoMI Dataset**, a curated corpus of **134 naturalistic video conversations** in medical and therapy settings. Each recording features **two English-speaking participants** and includes a range of real-world acoustic challenges—such as **background noise**, **distant microphone placement**, **music overlays**, and **accent variation**—to simulate the diversity and complexity of clinical trial audio.

Of the 134 conversations, **89 were included in this analysis**; the remaining 45 were excluded due to **incomplete annotations or missing audio files**. This subset still provides a robust benchmark to evaluate ASR system performance under conditions reflective of actual clinical practice.



Better Metrics Aligned with Humans

To evaluate transcripts beyond raw error counts, we employed **Human-Evaluated Word Error Rate (HEWER)** and introduced a **speaker-aware variant: cpHEWER** (Clinician-Preferred HEWER). These metrics more accurately reflect **clinically meaningful transcription quality** by disregarding non-semantic errors (e.g., filler words like *uh*, *um*) that do not impact comprehension or care.

Unlike traditional WER, which penalizes all deviations equally—including regional spellings (*behaviour* vs. *behavior*) or misspellings of non-essential words—HEWER focuses only on **semantic deviations**. cpHEWER builds on this by also incorporating **speaker attribution**, a critical factor in clinical contexts.

Ground Truth:
Clinician: So, [uh] how have you been feeling this week, [okay]?
Patient: It's been [a-] a difficult one.

ASR Output:
Clinician: So, how have you been feeling, [ok]?
Clinician: It's been a difficult one.

red text indicates incorrect errors due to non-semantic words
blue text indicates incorrect speaker assignment

- **WER** -> how many words are wrong
- **cpWER** -> how many words are wrong OR attributed to the wrong speaker
- **HEWER** -> how many *meaningful words* are wrong
- **cpHEWER** -> how many *meaningful words* are wrong OR attributed to the wrong speaker

Clinical-Grade ASR Performance

Standard ASR evaluation metrics often over-penalize minor, non-impactful discrepancies while overlooking clinically significant speaker attribution errors. In contrast, our use of **HEWER** and **cpHEWER** enables a more clinically relevant evaluation of system performance by focusing on meaningful errors that affect real-world comprehension and accountability.

As shown below, our refined metrics demonstrate a **44% reduction in transcription errors** and a **48% reduction in speaker attribution errors** compared to traditional WER/cpWER assessments. These gains reflect more than statistical improvement—they indicate the precision required for **clinical-grade transcription** in trial documentation, therapeutic auditing, and compliance-sensitive workflows.

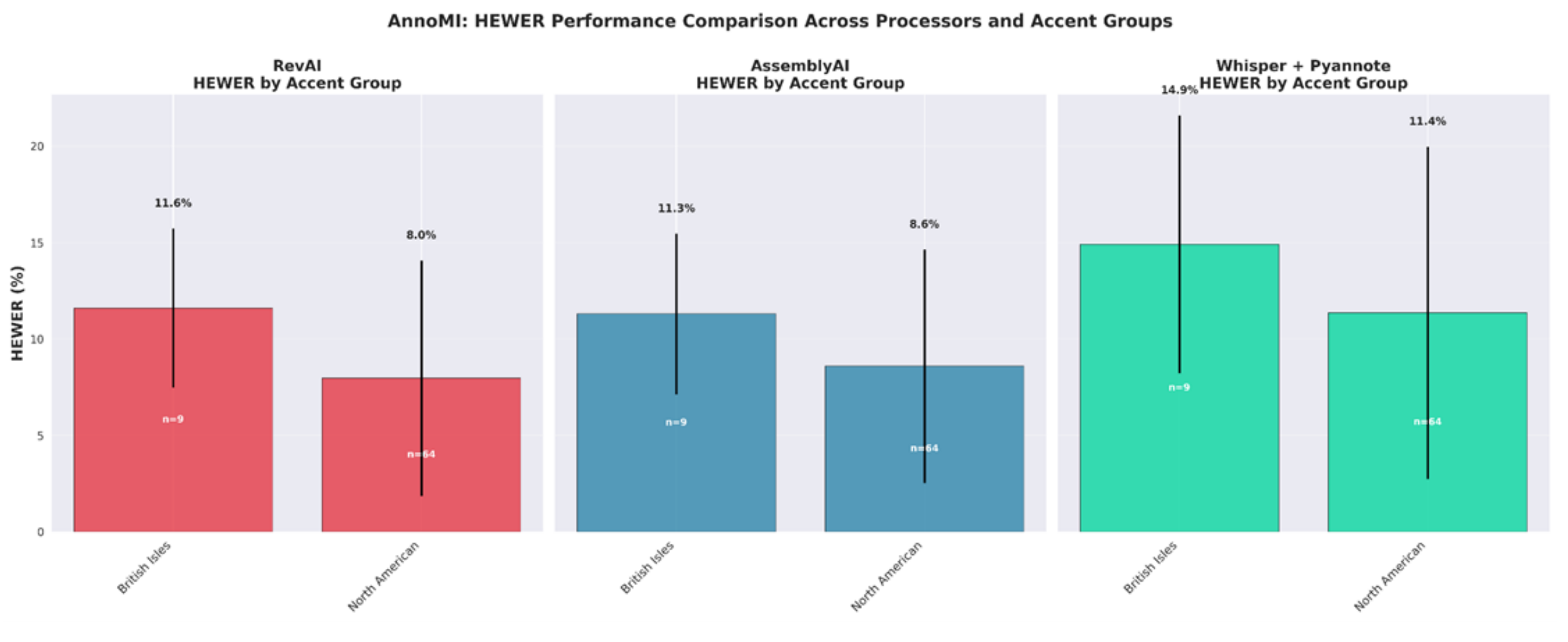
By adopting cpHEWER as the new gold standard, clinical teams gain actionable insight into **when ASR output is sufficiently accurate**, and when further refinement is needed. Crucially, this does **not** always mean defaulting to human oversight. Instead, these insights can inform **strategic model optimization**—including domain-specific fine-tuning, prompt engineering for transformer models, and contextual calibration—leading to smarter, more adaptive ASR pipelines that align with clinical risk profiles.

System	WER (%)	HEWER (%)	cpWER (%)	cpHEWER (%)
AssemblyAI	22.0	8.9	24.3	12.8
RevAI	19.2	8.4	23.2	11.2
Whisper + Pyannote	23.1	11.8	31.4	20.3

Accent-Sensitive ASR: Meeting the Needs of Global Trials

In global clinical trials, transcription accuracy must extend beyond American English. Clinical conversations involving speakers from the **British Isles**, **Australia**, and other English-speaking regions often feature distinct intonation, pacing, and idiomatic expressions—all of which pose challenges for ASR systems not trained on diverse linguistic inputs.

To assess ASR robustness across accents, we conducted a second benchmarking experiment comparing **Rev AI**, **Assembly AI**, and **Whisper + Pyannote** on audio data from North American, British Isles, and Australian speakers.



Accent-Sensitive ASR Cont.

Preliminary results showed:

- **Rev AI**: 11.6% error rate
- **Assembly AI**: 11.3% error rate
- **Whisper + Pyannote**: 14.9% error rate

While all systems performed reasonably well, **Whisper + Pyannote demonstrated a higher error rate**—underscoring the importance of **system selection** and **accent-aware calibration** when operating in multi-regional clinical settings.

These findings reinforce a core message: high-quality ASR in clinical trials or healthcare requires not only technical precision but also **linguistic inclusivity**. Accent variance must be considered in both **model selection** and **training data strategy**, especially when deploying in global trials or health systems serving diverse populations.

Recommendations: Driving Clinical-Grade ASR Forward

To ensure ASR systems can be safely and effectively deployed in global clinical trials, we propose the following innovations—balancing technical advancement with immediate utility for clinicians and sponsors.

1. Build Compound Error Metrics for Real-World Risk Scenarios

Current metrics like cpWER and cpHEWER treat transcription and speaker attribution errors independently. Yet in clinical settings, errors that are both semantically and speaker-inaccurate (e.g., hallucinated speech assigned to the wrong person) can have outsized impact on safety and documentation accuracy. Future work should define new compound metrics that penalize dual errors more heavily—better reflecting the real-world stakes of compounded ASR failure.

2. Prioritize Speaker Count Detection Accuracy

Before word- or speaker-level accuracy can be assessed, systems must detect the correct number of speakers. Misidentifying a two-speaker conversation as one voice renders all downstream analysis unreliable. We propose a standalone "speaker count accuracy" metric to flag systems that struggle with multi-speaker diarization—critical for quality control in trials and medical teams.

3. Close the Accent Gap to Advance Equity and Global Readiness

Our evaluation revealed higher error rates for non-North American English speakers. This "AI accent gap" risks reinforcing inequities in clinical trial participation and care delivery. To address this, we recommend targeted fine-tuning, inclusive dataset development, and accent-specific benchmarking for future ASR deployments. This is especially urgent for trials in the UK, Australia, and global sites where linguistic variation is the norm.

4. Expand Diverse, High-Quality Clinical Datasets

Our benchmarking relied on the AnnoMI dataset, which contains rich naturalistic audio but lacks full coverage across speaker labels and conversation types. To advance ASR for clinical applications, more open, ethically collected, multi-speaker datasets are needed—with diverse accents, background noise, and emotional tone. We invite partners to collaborate on dataset sharing and co-creation.

5. Enhance Audio Context with AI + Human Collaboration

Precision can also be improved through context-aware enhancements. Audio pre-processing tools (e.g., noise suppression, speaker separation) and prompt-based AI models can reduce transcription errors at the source. Just as importantly, clinician interfaces should surface confidence scores or risk alerts—indicating when human review may be warranted. This hybrid approach ensures safe deployment in even the most sensitive trial contexts.

Conclusion

This work contributes to the growing body of research aimed at aligning ASR evaluation methods with the nuanced demands of clinical trial communication. By advancing from role identification to utterance-level speaker attribution, and by introducing speaker-sensitive metrics such as cpHEWER, we emphasize the importance of evaluation frameworks that reflect real-world clinical risk. Our findings demonstrate that current standard metrics may underestimate clinically meaningful transcription and attribution errors—especially in the presence of accent variation or conversational overlap. As ASR systems become increasingly embedded in clinical research and care delivery, the need for rigorous, context-aware validation will only grow. Future work will continue to refine benchmarking approaches and explore methods to reduce both human and model-based transcription errors, with the goal of supporting high-integrity, human-centered communication in healthcare.

To connect with our team:
danielle@mpathic.ai

Accent types were automatically labeled using the Hugging Face model "Jzuluaga/accent-id-commonaccent_ecapa", an ECAPA-TDNN system fine-tuned on the CommonAccent dataset (16 English accents) [1].

References:
[1] Zuluaga-Gomez J, Ahmed S, Visockas D, Subakan C. *CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification.* Interspeech 2023. arXiv:2305.18283
[2] Ueda Y, Masumura R, Ueda Y, Hayashi T, Aso T, Ueda Y, Watanabe S. cpwer: A Concatenated Minimum-Permutation Word Error Rate for Automatic Speech Recognition. ICASSP 2023. arXiv:2211.08657
[3] Omar M, Prassad PS, Haghani P, et al. Humanizing Automatic Speech Recognition. Apple Machine Learning Research. 2023. Available: https://machinelearning.apple.com/research/humanizing-wer. Accessed: September 26, 2025.