Integrating synchronous AI-based fidelity adherence monitoring of facilitators in psychedelic clinical trials

Kelman, A[1], Field, M[1], Philp, W[1], Schlosser, D[2], Lord, S[2], Jolley-Paige, A[2], Greenlaw, M[2], Coyle, M[2], Mahableshwarkar, A[1], Inamdar, A[1]

[1]Cybin IRL Limited, 1 Spencer Dock Quay, Dublin 1, DO1 X9R7, Ireland
[2]mpathic, 14655 Bel-Red Road Suite 203. Bellevue, WA

History of Fidelity Adherence Monitoring

What is the methodological issue being addressed?
Synchronous mechanisms exist to monitor clinical assessors and other parties involved in psychedelic clinical trial conduct; however, tools are lacking to provide scalable, unbiased, consistent and near real-time monitoring of facilitators who provide psychological support for participants in studies with psychedelics. The below poster highlights this notable gap and provides context on how an AI-based tool may consistently monitor psychological support model fidelity adherence in near real-time. This AI-assisted monitoring may also allow for needed in-trial facilitator remediation training and potentially improve future adherence by facilitators who are not meeting fidelity adherence expectations.

Fidelity Adherence Monitoring: Why And Historically How

How and When Has Fidelity Adherence Been Tracked Historically?

- Fidelity monitoring is a form of qualitative coding or annotation where human raters agree on a behavior or construct (e.g., adverse events, unblinding, consent) to identify in a recording.
- Adherence has been traditionally tracked with central raters (i.e., offsite medical monitors) or in-person monitors.
- These monitors include teams of human reviewers that ensure trial compliance by listening to recordings or observing in real-time.
- Before monitors are released to ensure compliance in a trial, they must achieve interrater reliability (IRR) so that all raters have the same interpretation of the fidelity manual.
- The standard for IRR for qualitative coding or annotation is Cichetti (1995) with interrater agreement ranging from 0 to 1 with above .7 as acceptable, .8 as good and .9 as excellent.
- There is variation in the literature on the statistics and methods used to establish reliability when it comes to behavioral rating at the level of the sentence or utterance (see Hallgren, 2012 and Lord, 2014 for review).
- Krippendorf's alpha is the most appropriate statistic for IRR at the utterance level; Kappa and ICCs are sometimes used despite distortions with multiple raters and labels.
- While standards for agreement are established, the reality of fidelity monitoring is quite variable as reported in the literature with few constructs achieving acceptable agreement in trials.
- The below table summarizes the highest and lowest agreement reported in the literature for basic constructs used in common factors counseling (see review in Lord 2014).
- The lowest agreement reported in the literature, of peer-reviewed studies in clinical trials for human IRR, is 0.
- Reasons for not attaining reliability may range from poor construct definition to data scarcity; but more commonly the problems stem from disagreements in the interpretation of the construct leading to human variability.
- Thus, pre-training raters in agreement prior to evaluating clinical fidelity is extremely important, though difficult.
- Attained IRR may drift over time between raters or even within a rater themselves as time passes. Therefore, AI models (which are not subject to such drift) may serve as an augmentation for humans and thus address many of these issues and promotes quality oversight.
- AI may be trained on gold standard human raters and has less chance of drift and an overall higher chance of consistency, especially when exposed to training data of many different raters.

Label (Behavior)	Lowest in Literature	Highest in Literature
Closed Ended Questions	.55	.94
Open Ended Questions	.65	.89
Giving Information	0	.98
Direct	0	.57
Structuring Comments	0	.28

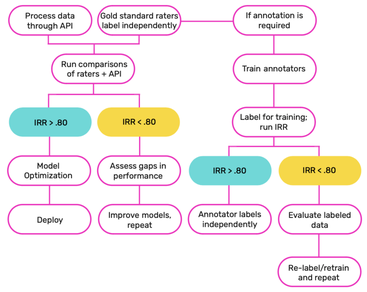
Evolution to Synchronous Adherence Monitoring

LLMs and AI to Conduct Synchronous Monitoring

- In our Phase III CYB003 trials, we will validate (using Phase I/II data) and employ an AI-based fidelity adherence monitoring system that is synchronous and scalable.
- AI systems may be:
  (1) trained to detect fidelity adherence criteria and provide synchronous monitoring in clinical trials at or above the performance of human raters for key adherence criteria, and
  (2) increase quality oversight by automating review and identifying moments that may require human review, and
  (3) be able to efficiently review 100% of multi-hour sessions.
- AI, when developed and used ethically and responsibly, may also improve outcomes by automating detection of potential trial risks such as detecting when key adherence criteria are missing.

Establishing Interrater Reliability in Our Phase III Trial

- To assess performance, the model should be trained on a ground truth dataset representing a set of utterances annotated with each construct the model needs to monitor fidelity adherence.
- These examples (e.g., how well it fits the fidelity standard) show the acceptable variability in the quality of the utterance as well as a range in both syntax and structure.
- The data should be a representative sample of the larger population it will be used with.
- The AI model ratings should be compared against expert rater evaluations to assess areas of misalignment across all constructs, and model performance should be evaluated by having two gold standard raters (i.e., expert raters) annotate a sample of data and compare it against the automated assessment.
- The aim is for less than 20% variance between raters and the AI for constructs, measured with an F1 score. When this score is not achieved, expert reviewers will assess the reason behind the variance.
- Prior to rolling out in our Phase III trial, we will confirm sufficient IRR to deploy the AI system by testing IRR and detections in our Phase I/II CYB003 audio video recordings.



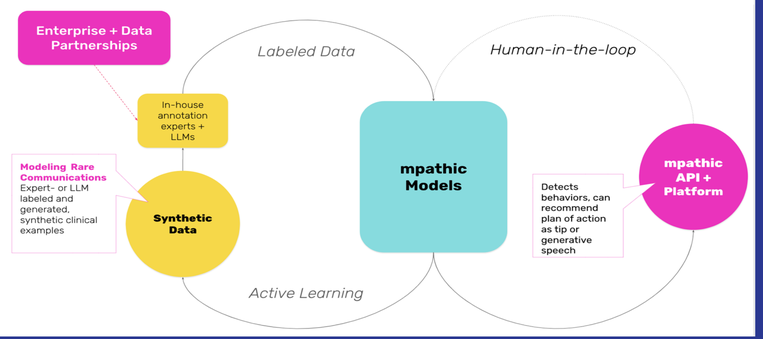
Training the Model

- Training AI on sufficient quality, traceable, accurate medical information will lead to better predictive accuracy and more robust performance.
- Datasets used to train Large Language Models (LLMs) may contain harmful or toxic data from questionable sources (e.g., Reddit, social media) which may contribute to inaccurate, harmful responses from publicly available LLM-powered products, and may also lead to mistakes/failure to detect clinical risks.
- AI should be trained on clinically/medically robust and bias-free data as much as possible.
- The AI model that will be used in the Phase III trial employs a mix of data including public and proprietary data to train models for behavior detection.
- Multiple tools and techniques including natural language processing, conversational patterns, and embeddings will be applied to create behavior detections. Post training, the model's performance will be assessed on unseen human-labeled testing data, using metrics like accuracy, precision, recall, and F1 score to determine readiness for use.

Within Trial Functionality

Deploying the Model

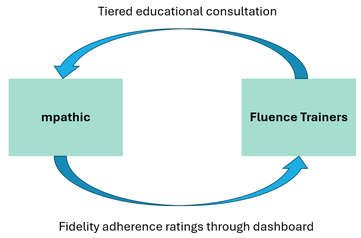
- Once IRR is established and the model is trained, expert monitors should review model performance to identify potential areas of drift as the Phase III study needs evolve and will update as needed with new labeled data.
- A stratification protocol should be employed to determine the frequency and variety of sessions to sample to verify model quality (e.g., samples across languages, locations, and demographic categories).

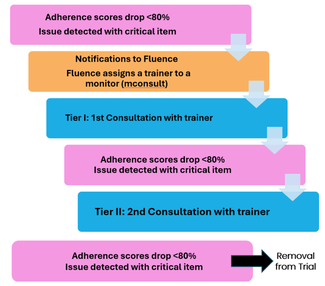


Implications on Facilitator Oversight During Trial

Ongoing Facilitator Monitoring and Governance During Our Phase III Trial Conduct

- Historically, it was not feasible to synchronously generate fidelity adherence output for 100% of sessions, contingent on those sessions being recorded in line with AV procedures at the site level.
- Reviewing 100% of sessions asynchronously is not practical in psychedelic trials because sessions are long, often 6-10 hours.
- With this synchronous adherence monitoring, trainers who qualify facilitators prior to study start of our Phase III program will have access to a portal where they may review timestamped, de-identified recordings and initiate tiered remediation consultation (see below Figures) with facilitators based on determined criteria (e.g., if they dip below 80% fidelity adherence and/or missing critical adherence items).
- This approach allows for continued governance and oversight and provides for more standardized facilitation with the intention to provide better participant outcomes.





Conclusions

Conclusions
Synchronous fidelity adherence monitoring using AI may allow for more robust (up to 100%) and timely adherence monitoring allowing for more consistent and predictable facilitation, which would reduce facilitator-related bias impacting trial outcomes. Other trials which include a psychological support component could benefit from exploring this near real-time fidelity adherence monitoring.

One or more authors report potential conflicts which are described in the program (paid employees of Cybin and mpathic, respectively).