SPEECH-BASED SCREENING OF DIFFERENT PSYCHIATRIC CONDITIONS-AND CHALLANGES OF CLINICAL IMPLEMENTATION

AUTHORS
Julianna Olah, Raheem Chaudhry, Edwin Wong, Aeron Laffare, Armin Sinka, Omar Mena, George Moses

AFFILIATIONS
Psyrin Inc.,
Allwell Behavioral Health Center.

INTRODUCTION

Speech pattern analysis shows promise as a diagnostic tool in mental health, but most studies are limited by unmatched demographics, single-condition focus and lack of clinical validation. We developed speech-based diagnostic models across five psychiatric conditions using demographically matched populations and started to validate their application in a behavioral health center in Ohio.

OBJECTIVES

- **Step 1: Training on self-reported data**: Develop machine learning models based on automatedly extracted speech characteristics collected from demographically matched populations with self-reported mental health diagnoses of OCD, ASD, MDD, ADHD, GAD
- **Step 2: Testing and implementation in clinical environment** in two two sites of a Allwell Behavioral Health Center (Zanesville, Guernsey)

STEP 1: MODEL TRAINING

- Data collection: 5-minute prompt-based speech recordings from participants with self-reported mental health diagnoses and healthy controls (N=900) via online research
- Conditions studied: Major Depressive Disorder, Attention-Deficit/Hyperactivity Disorder, Autism Spectrum Disorder, Generalized Anxiety Disorder, and Obsessive-Compulsive Disorder
- Feature extraction: Acoustic and linguistic features (automated transcription)
- Model development: Machine learning models with ensemble learning
- Demographic controls: Training data (80%) were controlled for demographic differences (age, English as first language, gender, ethnicity) between groups
- Multi-label approach: Participants could have multiple diagnostic labels, allowing models to capture disorder-specific speech characteristics
- Models were trained on binary classification of condition present/non-present, including potentially co-morbid conditions in the non-control group

STEP 1: MODEL TRAINING RESULTS (20% TEST SET)

	ADHD	ASD	OCD
Balanced accuracy:	88%	84%	86%
AUC:	0.9	0.92	0.85

	MDD	GAD
Balanced accuracy:	73%	89%
AUC:	0.74	0.91

THE PLANNED CLINICAL IMPLEMENTATION AND TESTING FLOW

1. consenting while waiting for appointment
2. 8-minutes speech task (5-minutes for testing, 3-mins warm up)
3. recorded in a dedicated room while waiting
4. speech quality test applied and patients asked to rerecord the task if they do not meet speech quality, duration or English speaking citeria
5. clinician review results before seeing the patient
6. clinician adds diagnosis after assessment
7. patient completes a feedback form at home

To test clinical feasibility, participation was voluntary and not monetarily compensated for either clinicians or patients

STEP 2: TIMELINE OF A SO FAR FAILED CLINICAL IMPLEMENTATION AND VALIDATION

- Training of clinical staff completed, with 6 clinicians volunteering to participate. Estimated sample size: 30-60 patients over 3 months. First 3 patients consented and completed the study.
- No new participants enrolled for 2 weeks. After in-person investigation, we discovered that the dedicated study laptop had been stolen from the recording room. A new laptop was secured and fixed to the table.
- Patients failed to create accounts independently, so we updated the account creation flow.
- 4 participants completed the speech task, while 5 discontinued participation due to frustration with failed speech quality tests.
- A new speech quality test design was deployed, with duration and quality criteria lowered. Signs were added to clinics highlighting the recording room location, as some patients had gotten "lost on the way."
- 3 patients completed the speech assessment, but another 2 failed the speech quality test. UX enhancements were added. Clinicians reported that they supervised patients while they completed the assessments, which wasted their time, but they did not trust that patients could complete the tasks independently.
- After an in-person visit, we realized that speech quality tests were failing because there was an air conditioner running in the room and music playing in the corridors. Clinics refused to turn off the music as it was designed to relax patients as part of a new quality improvement project.
- We turned off the speech quality test and informed clinicians that AI results with low quality scores were most likely to be incorrect. We temporarily stopped testing for accuracy. We are currently developing ideas on how to overcome the speech quality problems and how to test different solutions.

POTENTIAL WAYS TO OVERCOME IMPLEMENTATION CHALLANGES

- Streamline login and authentication process
- Improve UX and provide more insight into failed speech quality checks, with actionable feedback
- Add number-of-speakers detection to speech quality test
- Explore different prompting approaches
- Incentivize receptionists/case managers to help patients
- Deliver instructions to patients via video
- Apply enhanced filtering and voice enhancement
- Simplify clinician interface
- Collect voice data through an interactive, agent-based structure, instead of task-based assessments
- Increase clinical trust in at-home testing and transition to this model
- Implement gamification elements in the patient interface
- Add features to reduce clinician burden (e.g., generating notes, collecting voice via conversation settings or intake)
- Shorten speech assessment to 1-2 minutes and retrain models on shorter speech data
- Conduct speech testing via automated, agent-led voice calls

HAVE AN IDEA? SHARE IT!

CONCLUSION

Despite potential technical feasibility and observed success in research settings, the clinical application of speech-based screening for psychiatric conditions proved challenging (in two rural behavioral health clinic sites). The lack of research experience among clinical staff and low technical literacy of patients posed various difficulties in application. Clinicians' lack of trust in remote assessment completion enforced at-clinic assessment; however, both sites lacked sufficient human and physical infrastructure for successful implementation. Controlling the speech quality of recordings was particularly difficult. For successful adoption, intensive study of clinical workflows, patient behavior, and alternative user experiences is required. The authors further conclude that doing research is sometimes really hard.