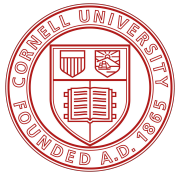
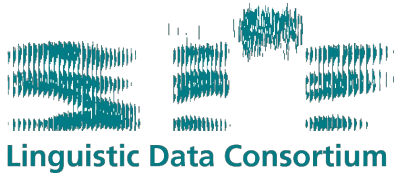


Comparative Evaluation of Acoustic Feature Extraction Tools for Clinical Speech Analysis

Interspeech 2025

Anna Seo Gyeong Choi, Alexander Richardson, Ryan Partlan, Sunny X. Tang*, Sunghye Cho*

The Clinical Context

Why handcrafted features still matter in clinical speech

- End-to-end models dominate ASR, but clinical needs are different
- Limited data: Patient populations, ethical constraints
- Interpretability: Medical decisions require explanations

The Clinical Context

- Current State:
  - Multiple acoustic toolkits widely used across research
  - Often used interchangeably without validation
  - Different research -> Different tools -> Different results?
- Research Question:
  - Do different acoustic feature extraction toolkits produce comparable results when applied to clinical speech data?

The Toolkits

OpenSmile openSMILE audio feature extraction

Standardized feature sets (eGeMAPS, ComParE)

Widely used for clinical usage, paralinguistic challenges

Statistical functionals over acoustic contours

Praat

Linguistically-motivated algorithms, manual verification

Specialized methods per feature type

Time-domain analysis, formant tracking

Librosa

Python ecosystem integration, recently joining clinical adoption

Spectral methods, probability estimation

Optimized for music, adapted for speech

The Toolkits

OpenSmile

Standardized feature sets (eGeMAPS, ComParE)

Widely used for clinical usage, paralinguistic challenges

Statistical functionals over acoustic contours

Praat

Linguistically-motivated algorithms, manual verification

Specialized methods per feature type

Time-domain analysis, formant tracking

Librosa

Python ecosystem integration, recently joining clinical adoption

Spectral methods, probability estimation

Optimized for music, adapted for speech

The Toolkits

OpenSmile

Standardized feature sets (eGeMAPS, ComParE)

Widely used for clinical usage, paralinguistic challenges

Statistical functionals over acoustic contours

Praat

Linguistically-motivated algorithms, manual verification

Specialized methods per feature type

Time-domain analysis, formant tracking

Librosa

Python ecosystem integration, recently joining clinical adoption

Spectral methods, probability estimation

Optimized for music, adapted for speech

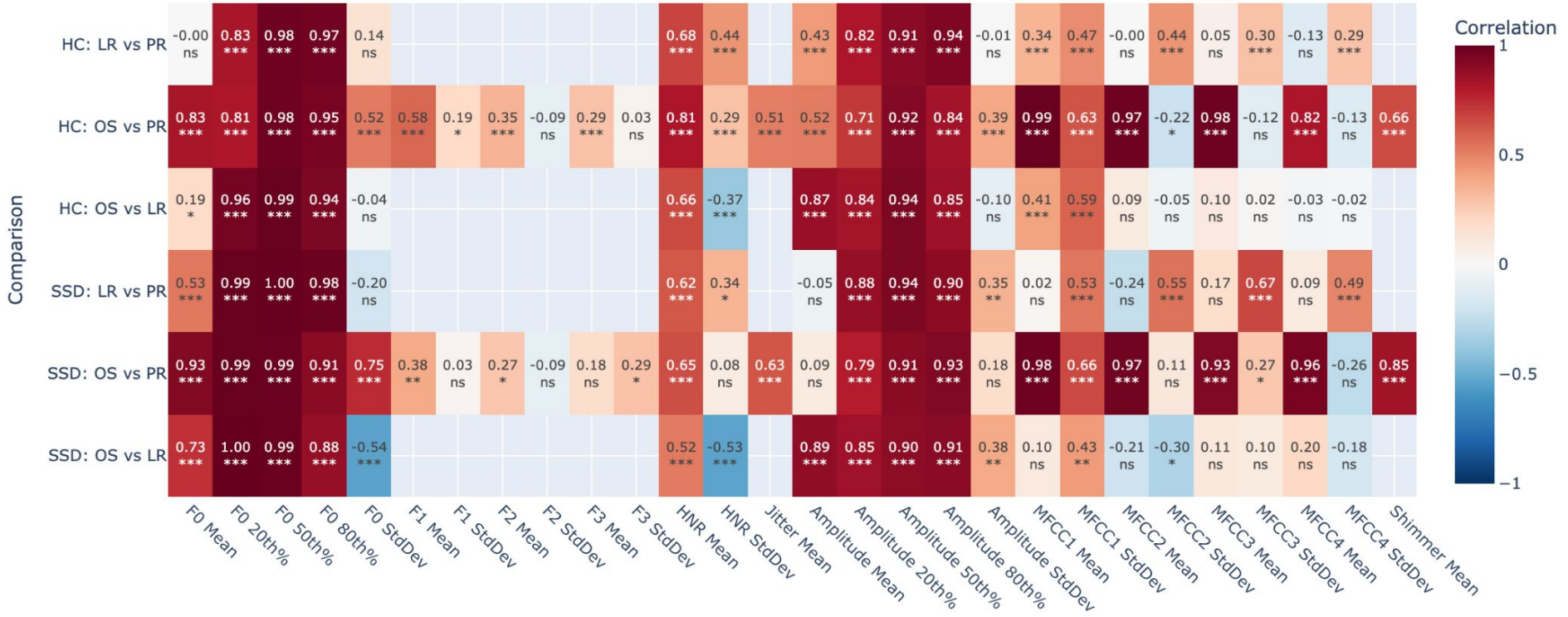
Clinical Acoustic Features

- Fundamental Frequency (F0)
  - Voice baseline, prosody patterns
- Formants (F1-F3)
  - Vocal tract resonances, articulation precision
- Voice Quality: Harmonics-to-Noise Ratio (HNR), Jitter, Shimmer, Amplitude
  - Vocal fold health indicators
- Mel-frequency cepstral coefficients (MFCCs)
  - Spectral shape, auditory-inspired representation

Dataset and Methodology

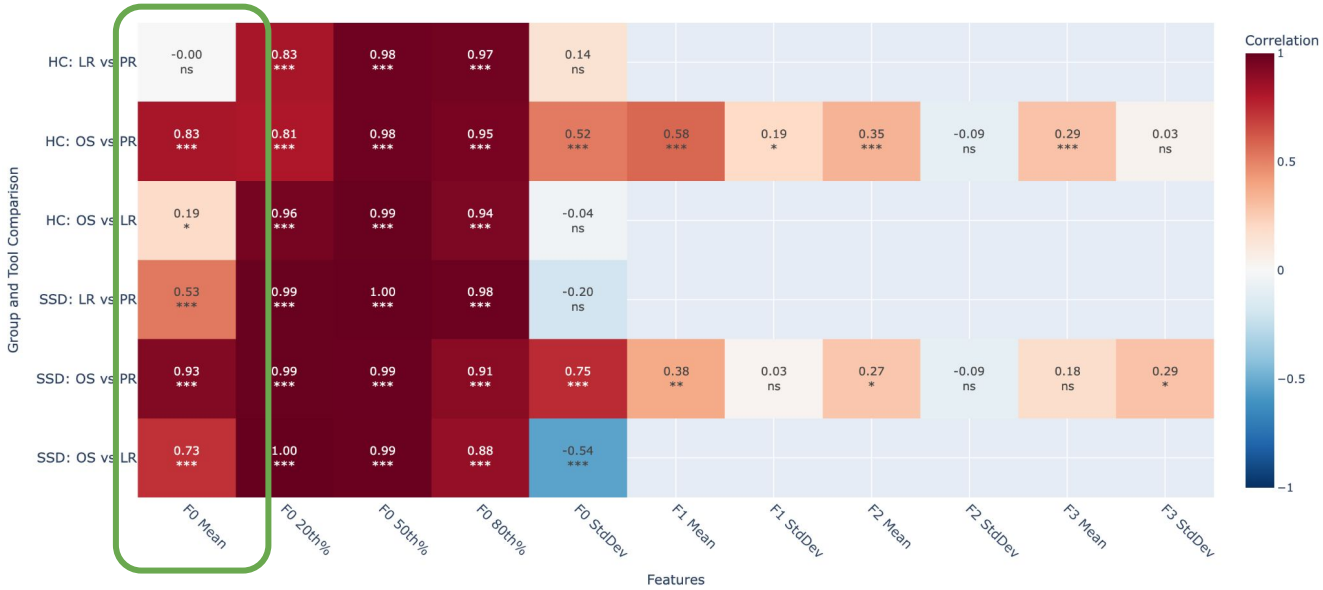
- Participants: 33 individuals with Schizophrenia Spectrum Disorder (SSD), 38 Healthy Controls (HC)
- Standardized extraction settings
  - Aligned parameters: Sample rate, frame size, window type
  - F0 range: 55-1000Hz (clinical populations)
  - Silence threshold: -60dB
  - Unavoidable differences: Core algorithmic approaches
    - F0: Cross-correlation (OpenSmile, Praat) vs. probabilistic methods (Librosa)
    - Formants: LPC vs. spectral peak tracking

Results Overview



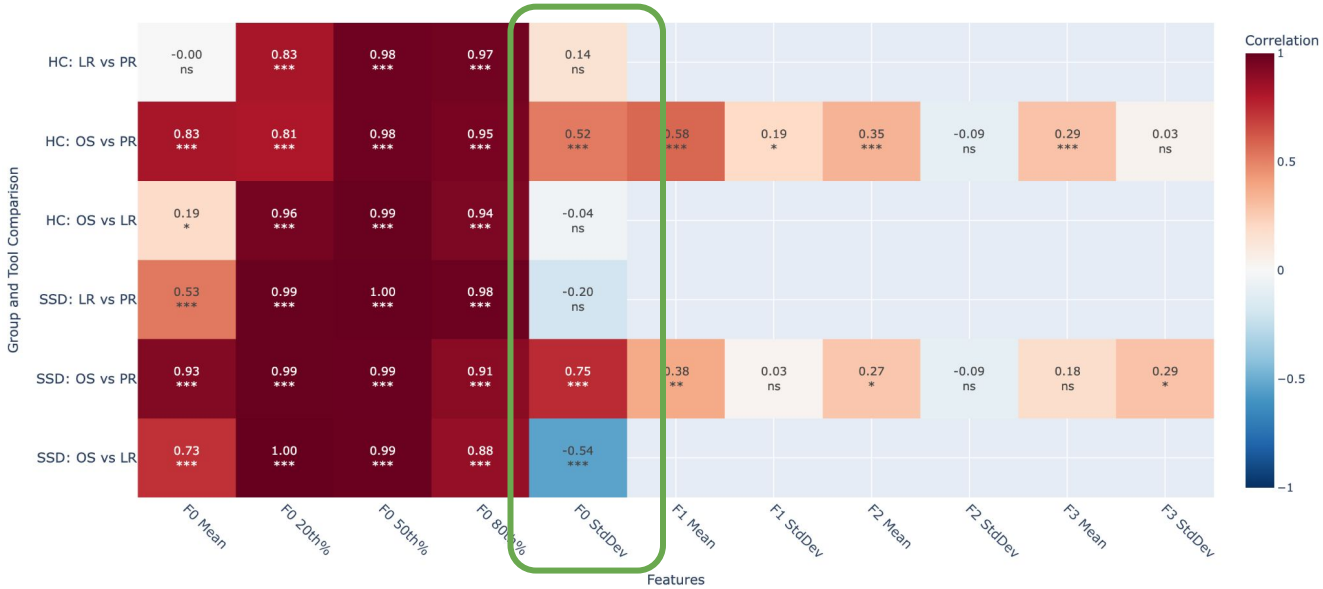
Results

- F0 mean moderate to poor agreement
- F0 standard deviation negative correlations
- F1, F2, F3 consistently low correlations across tools



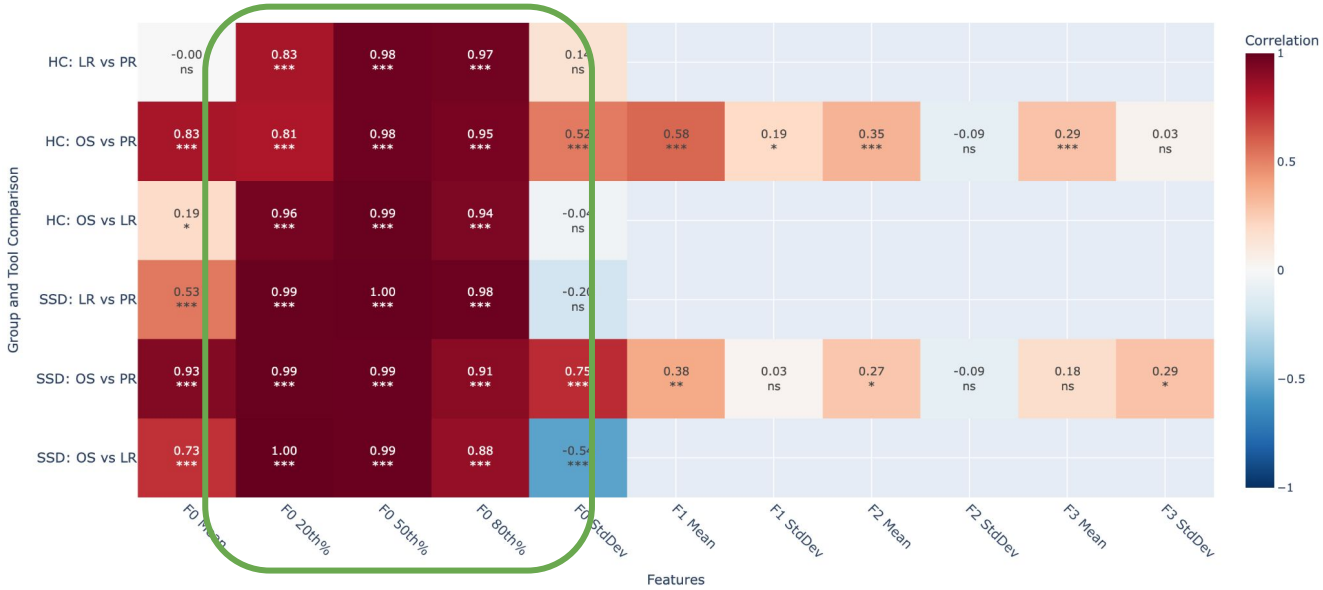
Results

- F0 mean moderate to poor agreement
- F0 standard deviation negative correlations
- F1, F2, F3 consistently low correlations across tools



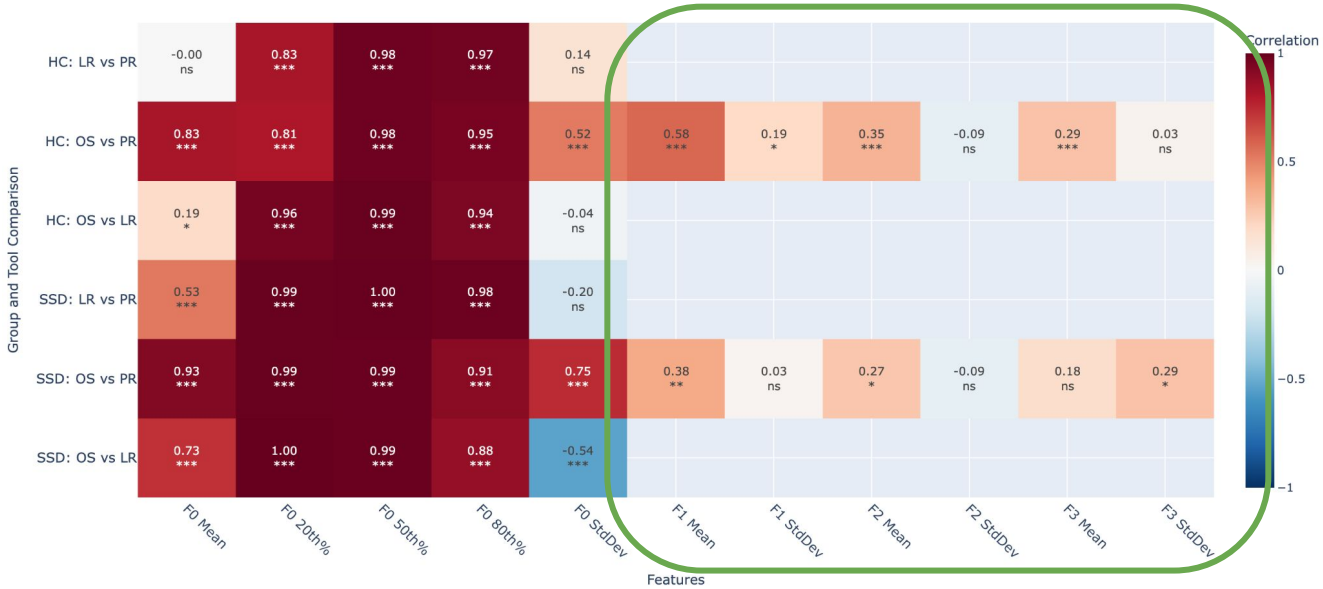
Results

- F0 mean moderate to poor agreement
- F0 standard deviation negative correlations
- F1, F2, F3 consistently low correlations across tools



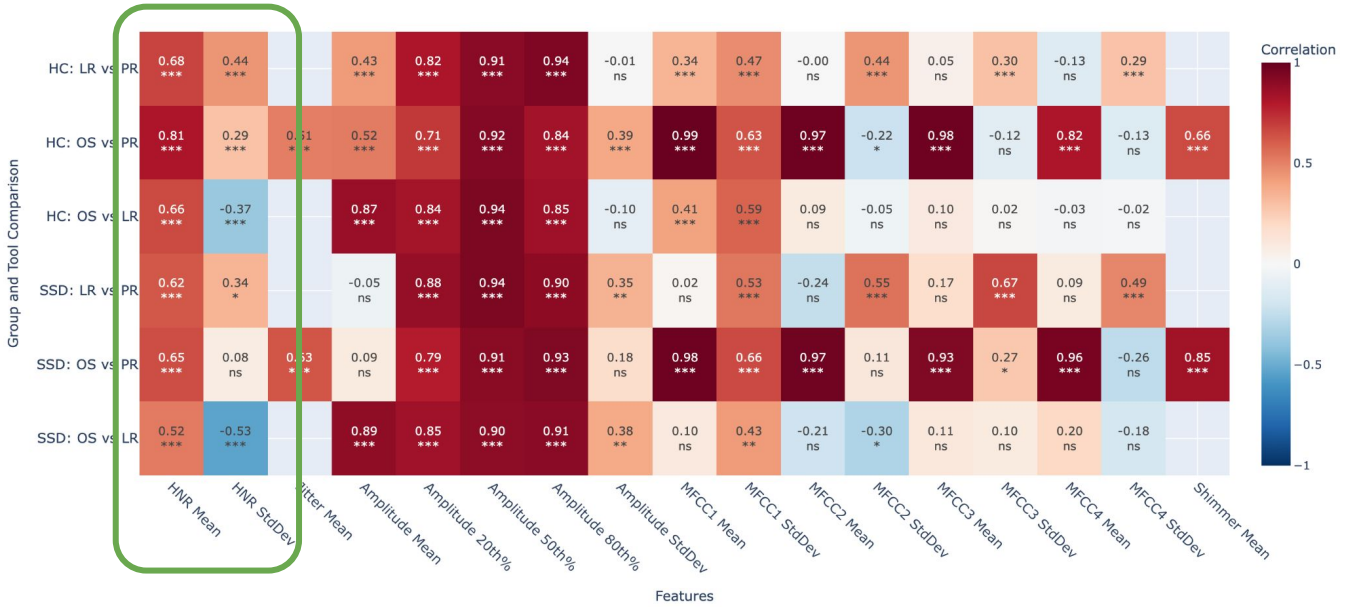
Results

- F0 mean moderate to poor agreement
- F0 standard deviation negative correlations
- F1, F2, F3 consistently low correlations across tools



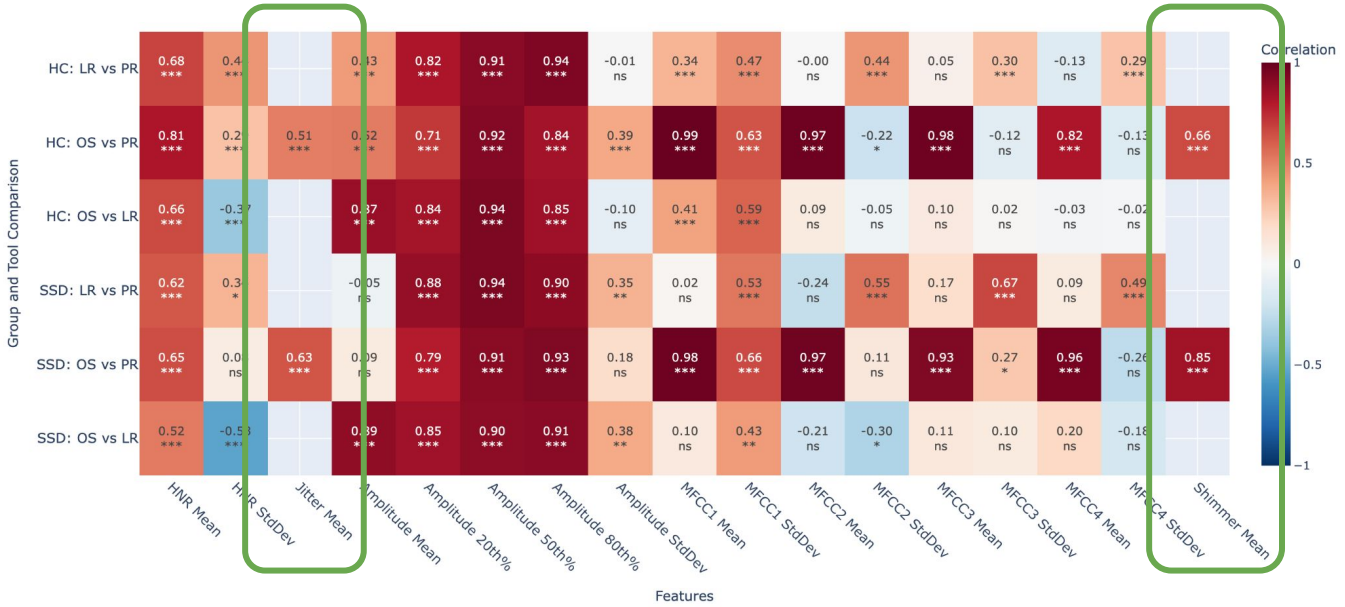
Results

- HNR moderate correlation
- Jitter/Shimmer reasonable agreement
- MFCCs patterns vary across coefficients



Results

- HNR moderate correlation
- Jitter/Shimmer reasonable agreement
- MFCCs patterns vary across coefficients



Results

- HNR moderate correlation
- Jitter/Shimmer reasonable agreement
- MFCCs patterns vary across coefficients

Key Findings

- Feature reliability is highly variable across toolkits
  - F0 percentiles: Excellent agreement ($r > 0.90$)
  - Formants: Systematic disagreement across all tools
  - Voice quality: Moderate to good reliability
- Robust vs. sensitive features identified
  - Stable extractions: F0 percentiles, jitter, shimmer
  - Sensitive to algorithms: F0 std dev, formants, some MFCCs
- Clinical implications are significant
  - Toolkit selection cannot be overlooked as methodological detail

Responsible AI Framework

Building Trustworthy Clinical Speech AI

- Transparency: Report tools, versions, parameters
- Cross-validation: Multiple toolkits, consistent findings
- Uncertainty quantification: Confidence in feature reliability
- Standardized protocols: Validated extraction pipelines

Future Directions

- **Same audio + Different tools = Different “biomarkers”**
- Deep embeddings gaining popularity, but same validation needed
- Balance innovation with clinical transparency requirements
- Patient safety requires reproducible methods

Thank you!

Read our full paper here:

Funding for this study was received from NIH K23 MH130750 and the Brain and Behavior Research Foundation Young Investigator Grant (SXT)

Special thank you to my collaborators!