

191

older adults
131 dementia · 60 controls

306

speech recordings
monologue + Cookie Theft

8

ASR systems
open-source + proprietary

2,754

transcripts
human + ASR

0.884

human AUC
reference transcripts

0.870

best ASR AUC
near human baseline

Background

ASR enables scalable dementia speech assessment, but transcription accuracy alone is not enough. The clinical question is whether ASR preserves diagnostic decisions, error balance and biomarker estimates.

Objective

Using human transcripts as the reference, test whether eight ASR systems alter classifier performance, diagnostic error profiles and biomarker validity.

Cohort and speech tasks

191 older adults: 131 with dementia and 60 healthy controls.
306 recordings: 193 dementia and 113 control.
Tasks: monologue and Cookie Theft picture description.
Total corpus: 2,754 transcripts from human + 8 ASR sources.

Analysis framework

Classifier trained on human transcripts only.
Participant-level 5-fold cross-validation prevented leakage.
The same model was applied to all ASR transcript sources.
Outputs: AUC, sensitivity, specificity, FNR/FPR, calibration and biomarker preservation.
Biomarkers: TTR, pronoun ratio, content-word ratio and mean word length.

Clinical question

Does ASR merely reduce transcription fidelity, or does ASR selection change the clinical behaviour of the entire AI-based dementia speech pipeline?

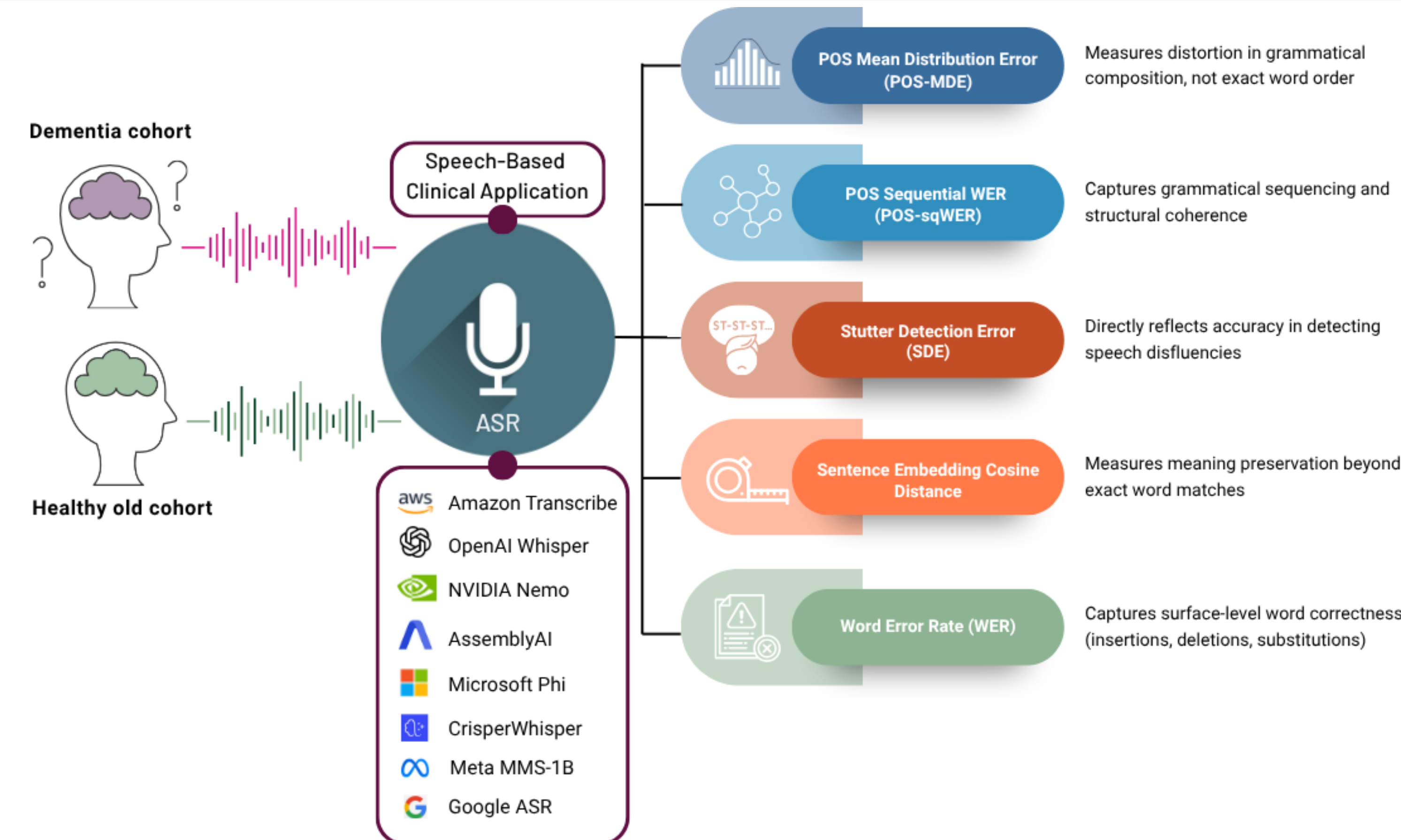
Conclusion

Speech-based dementia AI can retain strong aggregate discrimination with ASR transcripts, while still changing false-negative/false-positive balance and biasing linguistic biomarkers. Downstream clinical validity, not WER alone, should determine ASR selection.

Recommended validation minimum

Evaluate AUC plus FNR/FPR under the intended clinical decision threshold. Quantify biomarker preservation relative to human transcripts. Repeat by task, sample length, accent/dialect and impairment severity.

Clinical AI evaluation pipeline

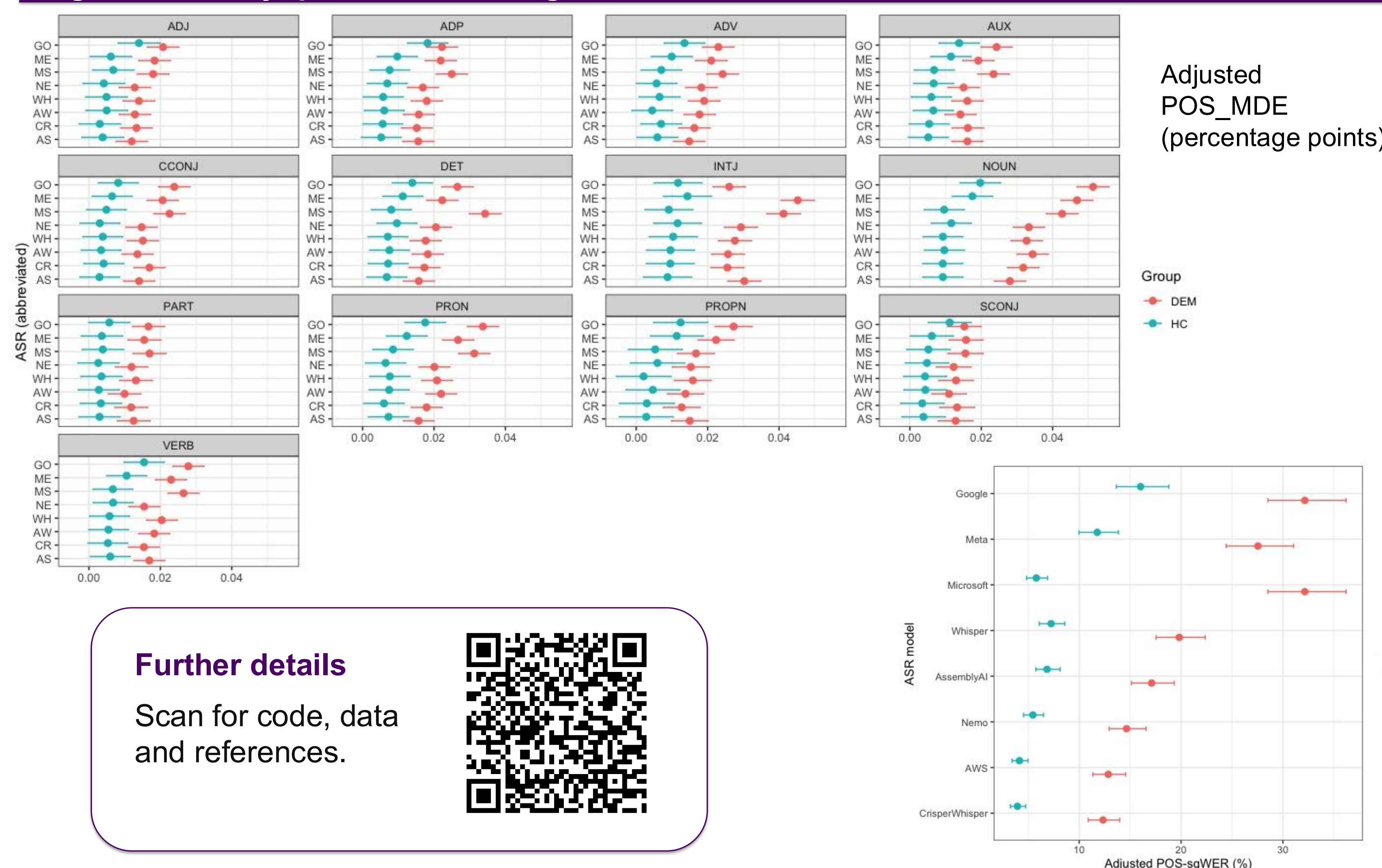


Experiment 1: aggregate discrimination

Source	AUC	ΔAUC
Human	0.884	—
Whisper (large)	0.870	-0.014
CrisperWhisper	0.870	-0.015
AWS Transcribe	0.867	-0.017
AssemblyAI	0.865	-0.019
Google ASR	0.862	-0.022
NVIDIA NeMo	0.843	-0.041
Meta MMS	0.835	-0.049
Microsoft Phi-4	0.829	-0.055

AUC stayed close to human baseline, but error balance and biomarker magnitude changed.

Figure summary: performance, diagnostic error and biomarker distortion



Further details

Scan for code, data and references.



Experiment 2: diagnostic error profiles

Google ASR
highest sensitivity
0.969
lowest specificity 0.433
FPR +0.053

Meta MMS
missed-case risk
FNR +0.093
sensitivity 0.865

Whisper
fewer false positives
FPR -0.106
but FNR +0.062

ASR shifted missed-case versus false-alarm balance, despite similar AUC.

Experiment 3: biomarker preservation

All four biomarkers showed significant group × transcription-source interactions. ASR changed the magnitude of dementia-control differences, not just transcript fidelity. Largest distortions included Meta MMS for TTR and Whisper for mean word length.

Where distortion was greatest

Monologues 12 FDR-significant tests vs 1 in picture description

Short recordings Q1 ≤57 words no significant tests in Q4 >123 words

Most impaired speakers highest vulnerability shorter / less fluent samples

Interpretation and deployment implication

Automated transcription does not merely add random noise. It can systematically change diagnostic operating characteristics and distort clinically meaningful language features. ASR choice should be specified, validated and monitored for each intended context of use.

