

UCSF

UC San Francisco Previously Published Works

Title

Storyteller in ADNI4: Application of an early Alzheimers disease screening tool using brief, remote, and speech-based testing.

Permalink

<https://escholarship.org/uc/item/4mk6b3qv>

Journal

Alzheimers & Dementia: The Journal of the Alzheimers Association, 20(10)

Authors

Skirrow, Caroline
Meepegama, Udeepa
Weston, Jack
et al.

Publication Date

2024-10-01


DOI

10.1002/alz.14206

Peer reviewed

RESEARCH ARTICLE

Storyteller in ADNI4: Application of an early Alzheimer's disease screening tool using brief, remote, and speech-based testing

Caroline Skirrow¹  | Udeepa Meepegama¹ | Jack Weston¹ | Melanie J. Miller^{2,3} | Rachel L. Nosheny^{2,4,5} | Bruce Albala^{6,7,8,9} | Michael W. Weiner^{2,3,5} | Emil Fristed¹ | for the Alzheimer's Disease Neuroimaging Initiative

¹Novoic Ltd, London, England

²Northern California Institute for Research and Education (NCIRE), San Francisco, California, USA

³VA Advanced Imaging Research Center, Department of Veterans Affairs Medical Center, San Francisco, California, USA

⁴Department of Psychiatry and Behavioral Sciences, University of California San Francisco, San Francisco, California, USA

⁵Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, California, USA

⁶Department of Environmental & Occupational Health, Public Health, University of California Irvine, Irvine, California, USA

⁷Department of Neurology, University of California Irvine School of Medicine, Irvine, California, USA

⁸Department of Pharmaceutical Sciences, University of California Irvine School of Pharmacy & Pharmaceutical Sciences, Irvine, California, USA

⁹Research Service, Veterans Administration Long Beach Healthcare System, Long Beach, California, USA

Correspondence

Emil Fristed, Novoic Ltd, 124 City Road, London EC1V 2NX, UK.
Email: emil@novoic.com

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Clinical trial registration:

AMY-PRED-US: NCT04928976;
AMY-PRED-UK: NCT04828122; ADNI4: NCT05617014.

Funding information

National Institute on Aging, Grant/Award Number: U19AG024904; National Institutes

Abstract

INTRODUCTION: Speech-based testing shows promise for sensitive and scalable objective screening for Alzheimer's disease (AD), but research to date offers limited evidence of generalizability.

METHODS: Data were taken from the AMYPRED (Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech) studies ($N = 101$, $N = 46$ mild cognitive impairment [MCI]) and Alzheimer's Disease Neuroimaging Initiative 4 (ADNI4) remote digital ($N = 426$, $N = 58$ self-reported MCI, mild AD or dementia) and in-clinic ($N = 57$, $N = 13$ MCI) cohorts, in which participants provided audio-recorded responses to automated remote story recall tasks in the Storyteller test battery. Text similarity, lexical, temporal, and acoustic speech feature sets were extracted. Models predicting early AD were developed in AMYPRED and tested out of sample in the demographically more diverse cohorts in ADNI4 (> 33% from historically underrepresented populations).

RESULTS: Speech models generalized well to unseen data in ADNI4 remote and in-clinic cohorts. The best-performing models evaluated text-based metrics (text similarity, lexical features: area under the curve 0.71–0.84 across cohorts).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 Novoic Ltd. *Alzheimer's & Dementia* published by Wiley Periodicals LLC on behalf of Alzheimer's Association.

of Health, Grant/Award Number: U19
AG024904

DISCUSSION: Speech-based predictions of early AD from Storyteller generalize across diverse samples.

KEYWORDS

Alzheimer's Disease Neuroimaging Initiative, digital recruitment, generalizability, mild cognitive impairment, speech-based testing

Highlights

- The Storyteller speech-based test is an objective digital prescreener for Alzheimer's Disease Neuroimaging Initiative 4 (ADNI4).
- Speech-based models predictive of Alzheimer's disease (AD) were developed in the AMYPRED (Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech) sample ($N = 101$).
- Models were tested out of sample in ADNI4 in-clinic ($N = 57$) and remote ($N = 426$) cohorts.
- Models showed good generalization out of sample.
- Models evaluating text matching and lexical features were most predictive of early AD.

1 | BACKGROUND

Alzheimer's disease (AD) is the most common cause of dementia worldwide, with a community point prevalence of $\approx 4\%$ in adults aged > 60 .¹ The health burden of AD is set to increase, with an aging worldwide population,² and rates of AD are projected to almost double every 20 years.³

There is increased pressure to find scalable methods for identifying patients at early stages of the disease. This is driven by new preventative AD treatments coming to market primarily developed for earlier disease stages,^{4,5} and a range of new disease-modifying drugs in the development pipeline.⁶

Neuropsychological testing, often administered in a question-and-answer format, has been extensively used to identify individuals at risk of AD. However, analysis of speech itself is emerging as a target for measuring subtle cognitive impairment and decline.⁷ Administration and analysis of speech-based testing can be fully automated,⁸ providing a scalable means for broader screening activities. This approach could be married with new, highly accurate blood-based tests for identifying disease-specific biomarkers,^{9,10} promising to further bring down the staffing burden and related costs of identifying AD-related pathological changes.

This approach is used in the Alzheimer's Disease Neuroimaging Initiative 4 (ADNI4), in which a large group of participants are recruited and assessed online before funneling those with cognitive impairment or self-reported cognitive decline for more in-depth clinical and biomarker evaluations.¹¹ Novocis's Storyteller, an automated story recall task, is being used as part of the online screening efforts in ADNI4.¹¹ Storyteller's generalized matching algorithm (G-Match) provides a fully automated method for eval-

uating proportional recall in story recall tasks sensitive to early AD.⁸

Beyond proportional recall, research has documented changes in speech and language occurring early in AD, including but not limited to changes in lexical variation, repetitions, the use of indefinite terms, noun and pronoun use, syntactic complexity, word finding difficulties, speech rate, pausing, and a range of acoustic measures,¹²⁻¹⁶ some of which have also been reported as differing in relation to AD biomarker (amyloid and tau) status.¹⁷⁻²⁰ Subtle signals like these have the potential to be embedded within larger machine learning or artificial intelligence (AI)-based models to improve overall predictiveness and sensitivity,^{21,22} with the availability of sufficient data.

However, studies of speech and language patterns in AD to date have typically been carried out within relatively small samples,²³⁻²⁶ with data collected on prespecified devices, and with limited evidence regarding the generalizability of the speech features analyzed, with some exceptions.²⁷⁻²⁹ When speech feature sets, from the thousands that are available, are curated and optimized, there is a risk of overfitting data and generating spurious results.³⁰ Additionally, aspects of speech and language also reflect the speaker's mother tongue, dialect, culture, social status, education, sex, race, and age.³¹ Identifying patterns in speech and language that are associated with early AD, and that generalize across different population groups, is key to accurate, equitable, and scalable prescreening.

ADNI collects high-quality standardized datasets across a variety of data modalities, and now also in a more diverse and representative participant sample,¹¹ to share with researchers to advance the AD field. Now, with Storyteller, speech data become one of these datasets. Analogous to how blood samples stored in the past can be reanalyzed using the latest technologies, speech can also be reanalyzed as

methodologies improve, for example, with advances in AI methods using newer large language models.

The current paper evaluates the feasibility of a large-scale device-agnostic speech-based data collection initiative, and the generalizability of speech-based prediction models when applied to this context. Using Storyteller's automated story recall tasks, we first develop models to predict clinical groups with and without cognitive impairment (mild cognitive impairment [MCI] or mild AD dementia) in the Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech (AMYSPRED) clinical study (NCT04828122, NCT04928976). The results are validated out of sample in ADNI4 (NCT05617014, data extraction date: April 19, 2024) in a smaller sample with clinical labels, and in a larger sample with self-reported diagnosis, allowing evaluation of generalizability across these studies and demographic groups.

2 | METHODS

The current study takes participant samples from three different cohorts, described in more detail below. This includes the AMYPRED cohort (Section 2.1), and ADNI4 in-clinic and remote cohorts (Section 2.2). Details on study-specific inclusion and exclusion criteria are provided in brief in the relevant sections, and a comparison of key sampling criteria is provided in Table S1 in supporting information, allowing a comparison of recruitment characteristics and inclusion and exclusion criteria. Participants across all groups were assessed in English, using English-language variants of test materials.

2.1 | AMYPRED

2.1.1 | Sample

Data were taken from the AMYPRED-UK and AMYPRED-US sister studies. Fuller details on sample characterization and methods are provided in Skirrow et al.⁸ and Fristed et al.^{21,22} Two hundred participants were recruited as a convenience sample from trial participant registries between November 2020 and August 2021. Only participants with confirmed amyloid beta biomarker status by positron emission tomography (PET) or cerebrospinal fluid test, and with established clinical diagnostic status (cognitively unimpaired [CU] or diagnosed with MCI or mild AD the previous 5 years) were approached. MCI due to AD and mild AD diagnoses were made following the 2011 National Institute on Aging–Alzheimer's Association core clinical criteria.³² Exclusion and inclusion criteria can be found in Table S1.

2.1.2 | Assessments

Participants completed clinical assessments via a secure Zoom link (UK) or in clinic (US), together with a trained psychometrician, during which a battery of clinical tests were administered and explored

RESEARCH IN CONTEXT

- 1. Systematic review:** The authors reviewed the literature using traditional (e.g., PubMed) sources. Speech and language changes are reported in Alzheimer's disease (AD), often in participants in the (more progressed) dementia stages of the disease, when speech is frequently evaluated in small samples and/or tested using cross-validation methods, limiting generalizability.
- 2. Interpretation:** Our results show out-of-sample generalization of speech-based models to predict cognitive impairment (both clinically determined and self-reported) as evaluated with the Novoic Storyteller test battery. Importantly the results generalize well across diverse samples, and in tests administered across a wider range of common devices and browsers, supporting scalability.
- 3. Future directions:** Research will continue to evaluate the generalizability and sensitivity of speech-based screening in the context of Alzheimer's Disease Neuroimaging Initiative 4's multi-tier screening and enrichment approach. Larger samples will allow the development and validation of more sophisticated speech-based models, including those detecting more subtle clinical presentations, and associated with AD biomarker status.

elsewhere.⁸ During the supervised clinical assessments, participants were supported with downloading the Novoic native application on their own mobile devices by study staff. The Novoic app was developed for self-administration of speech-based cognitive tests on participants' mobile smartphones, running on Android 7 and above or iOS 11 and above.

After their clinical assessments, participants were encouraged to engage in optional remote once-daily speech-based assessments using the app for up to 8 days. Remote assessments included the Automated Story Recall Task (ASRT) administered daily at the beginning of each assessment session. ASRTs were administered in UK English story variants for the UK study and in US English story variants for the US study.

When completing ASRTs, participants listen to prerecorded stories and are instructed to retell these in as much detail as they can remember, immediately after the presentation of each story and after a delay. Task responses are recorded and automatically uploaded to a secure server.

Data were taken from remotely administered ASRTs in AMYPRED. In AMYPRED, ASRTs were administered in threes, with an immediate recall of three different stories sequentially, followed by a delayed recall of each story after the completion of all immediate recalls.

For the current analyses, data were extracted from two immediate recalls of ASRT stories s1 and s2 (described in Skirrow et al.⁸) and

one delayed recall (story s1) administered in the same test session providing a cross-sectional dataset, to emulate a brief screening set-up.

2.2 | ADNI4

2.2.1 | Sample

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org.

The ADNI4 remote digital cohort is a large study cohort (target of 20,000 + participants), recruited and screened online, with the goal of enrolling 50% to 60% of new participants as individuals from historically underrepresented populations (URPs).¹¹ A detailed overview of the remote digital cohort can be found in Miller et al., this Special Issue. The study protocol and inclusion and exclusion criteria are detailed on the ADNI website (<https://adni.loni.usc.edu/methods/documents/>), and a summary can also be found in Table S1.

Out of the 20,000 participants recruited into the digital cohort, a subset of \approx 4000 will be selected to provide a blood sample for plasma biomarker analysis, of which a further subset will be referred to in-depth characterization in clinic at ADNI clinical sites to enroll \approx 500 new participants into the core study.¹¹ The newly recruited cohort will be joined by an estimated 500 participants who completed prior ADNI studies and will roll over into the ADNI4 study. Participants in the CU and MCI diagnostic arms of the in-clinic ADNI4 study are invited to complete remote assessments, including the Novoic Storyteller test, at baseline and 6-month intervals via the ADNI online study platform.

Although data collection in ADNI4 is ongoing across both in-clinic and cohorts, in the current study, cross-sectional data were taken from a subsample of participants completing their first, baseline Storyteller assessments up to April 19, 2024, at which time data was locked and exported.

2.2.2 | In-clinic assessments

In-clinic assessments were completed by roll-over participants from prior ADNI studies, including a clinical workup, in which their clinical status was evaluated, with participants designated as either being CU or having a diagnosis of MCI or AD dementia according to study protocols (<https://adni.loni.usc.edu/methods/documents/>). Evaluations included in-clinic administration of the Clinical Dementia Rating (CDR) scale³³ and the Mini-Mental State Examination (MMSE).³⁴

In-clinic participants also completed remote assessments and data collection procedures, as described below in more detail in Section 2.2.3, with the exception that self-reported diagnostic or medication information was not collected for in-clinic participants using

the ADNI online platform. The in-clinic ADNI4 participants join the ADNI online study to answer basic demographics questions, self-report memory concern and memory decline questions, and complete the Everyday Cognition 12-item scale (ECog-12)^{35,36} and the Novoic Storyteller test.

2.2.3 | Remote assessments

ADNI4 remote digital cohort participants provided data from remote digital assessment through an online portal based on the Brain Health Registry infrastructure,³⁷ which collects a range of information, including participant demographics, medical history related to study exclusionary criteria, self-report questions, and a measure of subjective cognitive/functional decline (ECog-12^{35,36}), with updated language to improve relevance to older more diverse adults.

Self-report questions, evaluated only in the remote digital sample as a proxy for diagnosis, included: self-reported clinical diagnosis of MCI, AD, or dementia, and having received a prescribed medication for memory problems or cognitive impairment.

One of the final tasks administered in the ADNI online portal is the Novoic Storyteller test. Storyteller is a remotely administered speech-based cognitive test battery, leveraging the ASRTs, validated in the AMYPRED studies.⁸ In ADNI4, the Storyteller battery comprises the following tasks: immediate recall of two different stories sequentially (ASRT story s1 and story s2), a category fluency distractor task (animals), followed by the delayed recall of the first story presented (story s1). The current study evaluates data from story recall tasks only, in US English ASRT story variants. Fluency tasks will be evaluated in more detail in separate analyses.

Compared to the legacy native app, the application visuals in Storyteller have been redesigned to further improve ease of use, and the application can now be accessed from a wide range of common devices via URL or Weblink (see [Supplementary Materials](#), Section 2 in supporting information, for device and browser compatibility), or integrated into online portals such as those in ADNI4 via a Software Development Kit (SDK), reducing burden on participants in terms of navigating downloads and access permissions on their devices. A visual representation of the Storyteller screens can be found in Figure S1 in supporting information.

Information on which devices and browsers were used to access Storyteller was collected from user agent strings. Information on the usability of Storyteller was collected via a participant satisfaction survey at the end of the task, evaluated using a 5-point scale customer satisfaction question format ("Did you enjoy Storyteller?"), with frowning to smiley emoji response options (1 = not at all, 5 = very much).

2.3 | Speech data preprocessing

Only participants completing Storyteller assessments were included in the current analysis. An overview of the fuller remote digital cohort,

including those who did not complete Storyteller, is reported in Miller et al., this Special Issue.

Participants who completed the Storyteller battery, but with no audible speech data or with transcription or feature extraction failure were excluded. The speech was analyzed using the following key approaches, described below, with resulting data available for researchers to access via the Laboratory of Neuro Imaging (LONI) system.

2.3.1 | Automated transcription

ASRT speech data was automatically transcribed using Google's speech-to-text automatic speech recognition (ASR) system,³⁸ and analyzed with Novoc's proprietary speech analysis software. In AMYPRED, data were also transcribed manually following a standardized procedure, including specified verbatim transcription of commentary, filled pauses, and partial words.

Transcription accuracy in AMYPRED data was evaluated with word error rate (WER), calculated using the HuggingFace Evaluate package³⁹ as the average number of ASR errors per manually transcribed word. This was calculated after removing punctuation, setting all text characters to lowercase, and removing filled pauses and partial words from transcripts before comparison.

2.3.2 | Text similarity analysis

Text similarity analysis was completed using a generalized matching score (referred to here for brevity as "G-match"). G-match was computed in Python as the weighted sum of the cosine similarity between the embeddings of original ASRT text and the transcribed retellings,⁸ based on a pretrained large multilingual language model. G-match quantifies the similarity across the two texts, with potential scores ranging from 0 to 100 (best performance score). G-match scores were generated separately for each story recall on Storyteller and averaged across the three-story recalls.

2.3.3 | Feature extraction

More than 50 prespecified speech features based on the research literature, and showing evidence of sensitivity to early-stage AD or other neurological and psychiatric conditions,¹²⁻¹⁶ were extracted using Surfboard⁴⁰ and BlaBla⁴¹ feature extraction packages. An abbreviated list of extracted features is provided in Table 1, with a fuller list in Table S2 in supporting information. These were analyzed together and separately according to feature domain, with lexical features including information relating to the language and the types of words used, temporal relating to timing-related features in speech (pauses, speech rate, and duration), and spectral features relating to the audio characteristics of the voice itself.

Individual features were extracted for each story recall task. Features were normalized for each task within training data folds, whereby z scores for each task-feature dyad were derived for each participant.

TABLE 1 Overview of speech metrics and domains extracted for analysis.

Text similarity	Feature domains		
	Lexical features	Temporal features	Spectral features
G-match	Number of words	Speech duration	F0 (mean and SD)
	Idea density	Speech rate	Harmonics-to-noise
	Pronoun/noun ratio	Total number of long pauses (> 200 ms)	ratio
	Noun rate	Total number of pauses	MFCC 1 to MFCC 13 (mean and SD)
	Unique word ratio	Total pause time	Jitter (5 features)
			Shimmer (5 features)

Abbreviations: F0, fundamental frequency; G-match, generalized matching algorithm; MFCC, mel-frequency cepstral coefficients; SD, standard deviation.

Adjusting the feature distributions in this way helps models to train more robustly. Test folds are similarly normalized according to training data means and standard deviations.

Normalized features were then averaged across tasks to improve the robustness of speech features, with each story recall serving as a repeated administration. Normalized, averaged features were then finally concatenated, resulting in a vector of dimension (number of features in the group) for each feature group. In addition, a combined feature group was evaluated, combining features across all feature domains (text similarity, lexical, temporal, and spectral features).

2.4 | Model development and analysis

2.4.1 | Defining clinical groups

Clinical groups were defined by the available data in each cohort:

1. Participants in AMYPRED and ADNI4 in-clinic cohorts were evaluated clinically and had clinically confirmed diagnostic labels (CU and MCI), which were used as diagnostic labels in predictive models.
2. In the ADNI4 digital cohort sample, only self-reported diagnosis of AD, MCI, or dementia was available. This was used as a proxy for clinical diagnostic labels in predictive models. Secondary, exploratory analysis was completed in relation to a simple (yes/no/I don't know/prefer not to say) self-reported question on prescription of medications for cognitive impairment or memory problems ("Have you ever been prescribed a medication for cognitive impairment or memory problems by a health-care provider?"). No additional information on medication types was provided by participants.

2.4.2 | G-match

G-match score was used to predict clinical group directly using receiver operating characteristic (ROC) analysis, by evaluating sensitivity and specificity at each G-match score value. This process was

carried out separately across the three cohorts (AMYSPRED, ADNI4 in-clinic, ADNI4 remote digital cohort). Note that for G-match (a single extracted feature) no training was carried out to optimize performance in the AMYSPRED data or other cohorts.

Area under the ROC curve (AUC) is reported. Ninety-five percent confidence intervals for AUCs were computed as the margin of error between 2.5th and 97.5th centile from 1000 randomly sampled bootstrap samples with replacement from the original dataset. Positive predictive value (PPV) and negative predictive value (NPV) are reported for a range of sensitivities (target 0.7, 0.8, and 0.9) and associated specificities. To investigate generalizability of G-match thresholds, a set of thresholds determined using the same target sensitivities in AMYSPRED were applied to the two ADNI4 cohorts, and the resulting sensitivities, specificities, PPVs, and NPVs in ADNI4 cohorts are reported.

2.4.3 | Within-sample generation and cross-validation of speech biomarker models in AMYSPRED

The predictiveness of speech-based models was evaluated within AMYSPRED. Feature vectors for lexical, temporal, acoustic, and combined features were used to train logistic regression models predicting clinical groups (MCI/mild AD and CU), evaluated with 5-fold cross-validation to generate ROC curves. For G-match, a single extracted feature, no training was carried out to optimize performance. Rather, the ROC curve was generated as described in Section 2.5.2 but on the five test folds to provide a direct comparison to the other models. Ninety-five percent confidence intervals were computed as the margin of error between the 2.5th and 97.5th centile by bootstrapping with replacement 1000 mean AUCs from the 5-fold cross-validation. The statistical significance of differences between AUCs was computed using the Wilcoxon signed-rank test.³⁶

2.4.4 | Out-of-sample generalization of speech biomarkers in ADNI-4

Out-of-sample predictions from AMYSPRED were produced using methods equivalent to those described above in Section 2.5.3, with the exception that for multi-feature domains a single predictive model was generated for each feature set in the AMYSPRED data. These models developed in AMYSPRED were used to predict previously unseen data in ADNI-4. The statistical significance of differences between AUCs was computed using permutation testing. Ninety-five percent confidence intervals for AUCs were computed in the same way as described in Section 2.5.2.

2.4.5 | Demographic comparison

All models were evaluated relative to a demographic comparison, combining age, sex, and years of education as input to a logistic regression

model analysis using an identical set-up to the feature-based models described above.

3 | RESULTS

3.1 | Participants

3.1.1 | AMYSPRED participants

Out of 200 participants in the AMYSPRED study, 101 participants completed the prespecified optional remote assessment session, including 46 individuals with a clinical diagnosis of MCI, and 55 who were CU. Characteristics of participants completing remote clinical assessments versus those who did not complete remote assessments have already been reported in detail previously.⁸

3.1.2 | ADNI4 participants

From the remote digital cohort, out of 914 who joined the study online and completed at least one remote assessment consented to complete remote assessments (demographics questionnaire is the first study task), 447 completed Storyteller (49% of consenting participants). An analysis of completion rates in the full remote digital cohort is provided by Miller et al. (this issue).

From the in-clinic cohort, out of 113 invited and of the 80 that joined and consented to complete remote assessments, 60 completed Novoc Storyteller (54% of invited and 75% of consenting participants, respectively).

Overall, 503 participants in ADNI4 provided responses to Storyteller, full demographic information (age, sex, years of education), and had either confirmed diagnosis via in-clinic assessment ($N = 58$), or self-reported diagnosis in the remote digital cohort ($N = 444$). Consort diagrams are provided in Figures S2 and S3 in supporting information.

Data from 20 participants (4%) with transcription or feature extraction failure from audio were excluded. The final usable ADNI4 sample overall comprised 483 participants from ADNI4, including 57 with complete in-clinic diagnostic evaluations (44 CU, 13 MCI), and 426 from the remote digital cohort (368 self-reporting as CU or diagnosis not known, and 58 with self-report of MCI, AD, or dementia). Of those self-reporting a diagnosis, 50% (29/58) also reported currently being prescribed medication for memory problems or cognitive impairment by their health-care provider, compared to 0.8% who did not report a diagnosis (3/368).

The ADNI4 cohorts had a better representation of individuals from historically URPs than AMYSPRED. In the included sample that completed Storyteller, 38.5% of the ADNI4 digital cohort and 33.3% of the in-clinic ADNI sample self-reported an ethnocultural URP background, compared to just under 2% in AMYSPRED. Demographic information for all three samples is provided in Table 2.

TABLE 2 Sample characteristics of AMYPRED, ADNI4 in-clinic participants, and ADNI4 digital cohort.

	Cohort, group					
	AMYPRED (N = 101)		ADNI4 in-clinic cohort (N = 57)		ADNI4 remote digital cohort (N = 426)	
	CU	MCI/mild AD	CU	MCI	CU or no known diagnosis (self-report)	MCI, AD, or dementia (self-report)
Total (N)	55	46	44	13	368	58
Female, N (%) / male, N (%)	34 (61.8%) / 21 (38.2%)	22 (47.8%) / 24 (52.2%)	27 (61.4%) / 17 (38.6%)	5 (38.5%) / 8 (61.5%)	284 (77.2%) / 84 (22.8%)	37 (63.8%) / 21 (36.2%)
Age, mean (SD)	69.89 (4.12)	68.93 (7.47)	73.70 (7.78)	75.00 (7.54)	66.48 (6.91)	69.71 (7.29)
Years of education, mean (SD)	15.12 (3.58)	15.26 (2.82)	17.16 (2.15)	15.62 (2.50)	16.26 (2.47)	15.72 (2.40)
MMSE, mean (SD)	29.02 (1.05)	27.29 (1.9)	29.32 (0.69) ^a	27.83 (1.70) ^b	-	-
CDR-G, mean (SD)	0.10 (0.2)	0.51 (0.13)	0.06 (0.17) ^c	0.50 (0.00) ^d	-	-
Race/ethnicity, N (%)						
Black or African American	1 (1.8%)	-	11 (25%)	2 (15.3%)	113 (30.7%)	5 (8.6%)
Asian	1 (1.8%)	-	2 (4.5%)	-	8 (2.2%)	2 (3.4%)
Native American	-	-	-	-	1 (0.3%)	-
Pacific Islander	-	-	-	-	2 (0.5%)	-
Mixed race	-	-	1 (2.3%)	-	10 (2.7%)	1 (1.7%)
Latino/a	-	-	3 (6.8%)	-	25 (6.8%)	3 (5.2%)

Abbreviations: AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; AMYPRED, Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech; CDR-G, Clinical Dementia Rating, Global score; CU, cognitively unimpaired; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination; N, number; SD, standard deviation.

^aData available in a subsample of N = 26.

^bData available in N = 12.

^cData available in N = 25.

^dData available in N = 12.

3.2 | Devices and browsers in ADNI4

In ADNI4, participants accessed Storyteller on a wide range of devices, as revealed through user agent strings, collected during task completion. This included Windows devices (26.7%), MacOS X devices (18.8%), iPhones (33.3%), Android phones (16.6%), and iPads (3.5%). Just under half of participants accessed Storyteller with Safari (44.7%), followed by Chrome (38.9%), with fewer on Edge (9.5%), Firefox (3.7%), and Samsung Internet (1.9%).

3.3 | Usability

The usability of the Novaic native app in AMYPRED has been reported before.⁸ Within the Storyteller test battery in ADNI4, responses were broadly positive for 58% of respondents, and neutral or positive for 88% (Figure 1). A breakdown of distributions by diagnostic and self-reported diagnostic groups is provided in Figure S4 in supporting information.

Comparing participant satisfaction with the test ("Did you enjoy Storyteller?" on a 5-point scale) between groups, in the in-clinic sample ratings for the Storyteller test experience were similar for MCI and CU groups (mean of 3.45 and 3.20, respectively, $P = 0.63$); in the remote

digital sample ratings differed between those who reported a diagnosis of MCI, AD, or dementia and those who did not (mean of 3.37 and 3.88, respectively, $P = 0.0003$).

Participant-reported enjoyment of the test did not differ by sex (mean 3.73 for men, 3.76 for women, $P = 0.92$), but participants from URP groups reported moderately better satisfaction with the test than White non-Hispanic participants (URP mean 3.88, White non-Hispanic mean 3.67, $P = 0.03$). Enjoyment of Storyteller did not correlate with years in education ($\rho = 0.03$, $P = 0.43$), but did correlate with age, with a weak but significant correlation showing lower levels of enjoyment with older age ($\rho = -0.18$, $P < 0.0001$).

3.4 | Transcription accuracy

In AMYPRED average WER across participant recordings for automatic transcripts compared to manual transcripts was 0.10. Equivalent information for ADNI4 is not yet available.

3.5 | Speech features

An overview of extracted speech features (means and standard deviations) by clinical group and cohort is provided in Table S3 in supporting information.



FIGURE 1 Participant feedback after completing the Storyteller test battery.

3.6 | Text similarity analysis: G-match

A comparison of G-match score distributions across cohorts is shown in density plots in Figure 2, which plots the relative frequency of scores according to score intervals and shows the separation between the clinical groups.

G-match shows good predictive performance in all three cohorts (Figure 3), with AUCs of 0.82 (95% confidence interval [CI] = 0.74–0.89) in the AMYPRED cohort, 0.73 (95% CI = 0.58–0.86) in the ADNI4 in-clinic cohort, and 0.75 (95% CI = 0.67–0.82) in the ADNI4 remote digital cohort.

A selection of sensitivities, specificities, NPVs, PPVs, and their corresponding G-match thresholds are given in Table 3, representing target sensitivities of 0.7, 0.8, and 0.9. Thresholds for comparable sensitivity levels across the cohorts were similar for the two ADNI cohorts and higher for the AMYPRED (also seen in distributional differences in Figure 2; for example, a sensitivity of ≈ 0.7 was seen at a threshold of 66.9 for AMYPRED, and 57.7 and 57.8, respectively, for ADNI4 in-clinic and remote digital cohorts). To illustrate out-of-sample generalizability, G-match thresholds taken from AMYPRED and tested on the ADNI4 cohorts are given in Table S4 in supporting information.

3.7 | Within-sample generation and cross-validation of speech biomarker models in AMYPRED

Results from the comparison of the speech biomarker models within AMYPRED are presented in Figure 4A. The results show good performance of the G-match metric (AUC = 0.84, 95% CI = 0.77–0.90), albeit with a subtly different AUC to that shown in Section 3.4 due to the different analysis methodology, and also good predictiveness of lexical (AUC = 0.79, 95% CI = 0.71–0.88) and temporal features (AUC = 0.70,

95% CI = 0.56–0.78). More modest predictions are seen for spectral features (AUC = 0.61, 95% CI = 0.59–0.63). Combining all features into one predictive model did not improve predictiveness beyond separate feature domains (AUC = 0.70, 95% CI = 0.66–0.73). The demographic comparison performs at the chance level. The statistical significance of differences between AUCs was not apparent (minimum $P = 0.06$, see Table S5 in supporting information).

3.8 | Out-of-sample generalization of speech biomarker models in ADNI4

Overall, the speech biomarkers generated show good out-of-sample generalization to previously unseen data in ADNI4. Figure 4B shows predictions of clinical diagnostic labels of MCI in the ADNI4 in-clinic data (Figure 4B). The G-match metric (AUC = 0.73, 95% CI = 0.58–0.85), was moderately outperformed by lexical features (AUC = 0.79, 95% CI = 0.62–0.92). Temporal features (AUC = 0.68, 95% CI = 0.55–0.82) and spectral features (AUC = 0.63, 95% CI = 0.46–0.78) showed similarly lower strengths of prediction as shown within the AMYPRED sample. The statistical significance of differences between AUCs was not apparent, except comparison between models evaluating lexical features and demographics ($P = 0.05$, see Table S5).

Similar results were also shown for self-reported diagnostic labels in the larger remote digital cohort (Figure 4C), with out-of-sample generalization to self-reported diagnosis (MCI, AD, or dementia) with an AUC of 0.75 (95% CI = 0.68–0.82) for G-match, and consistent performance of lexical (AUC = 0.71, 95% CI = 0.64–0.79), temporal (AUC = 0.69, 95% CI = 0.61–0.76), and spectral features (AUC = 0.56, 95% CI = 0.48–0.64) to those shown in the prior models. With the larger sample size for the remote digital ADNI4 cohort with self-reported diagnostic labels, statistically significant differences were seen between AUCs, with G-match outperforming all other models

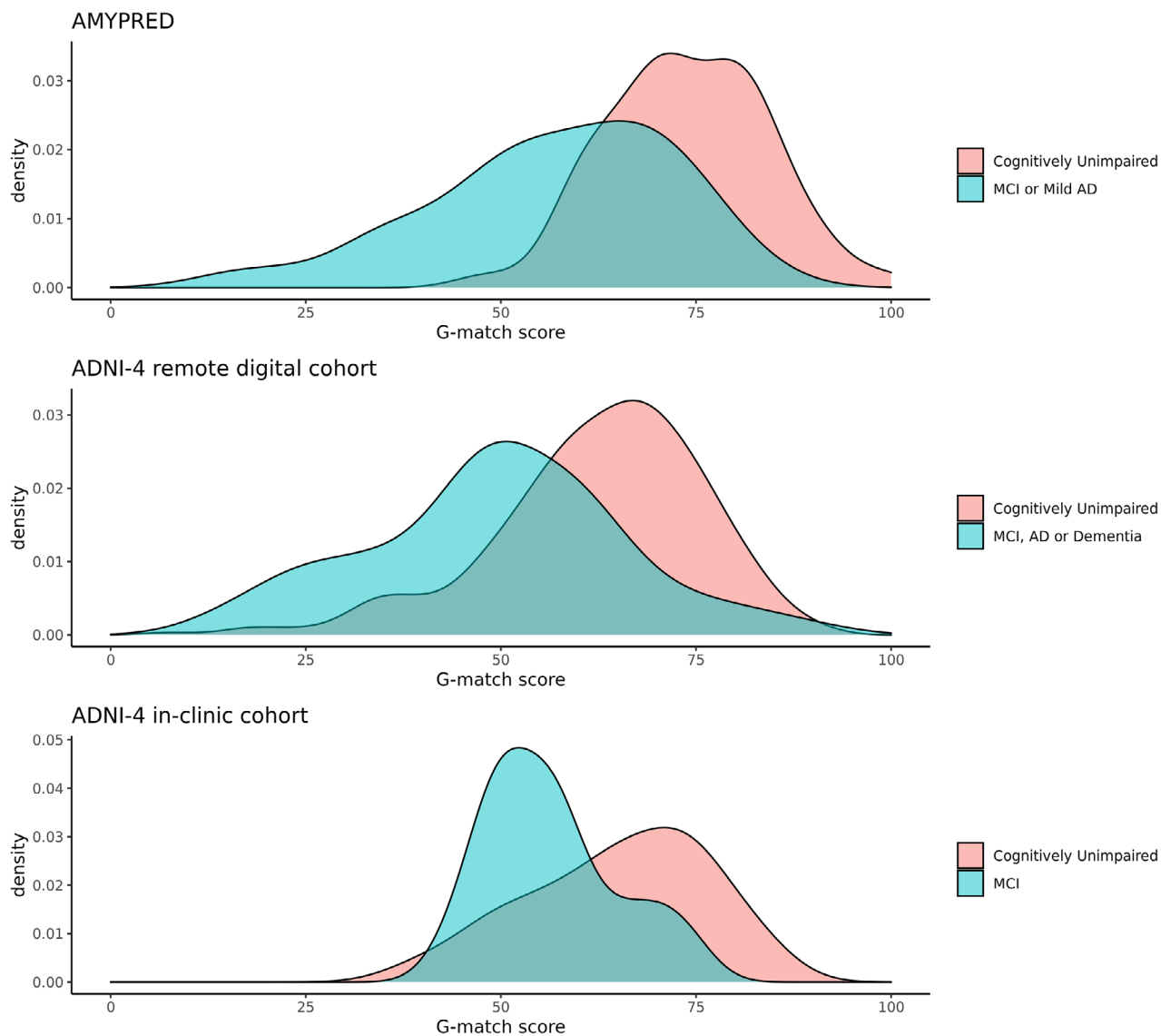


FIGURE 2 Density plot of G-match, an automatically derived text similarity metric comparing source text and the participant's retelling, for groups in each of the three cohorts. AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; AMYPRED, Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech; G-match, generalized matching algorithm; MCI, mild cognitive impairment

(min $P < 0.05$) except lexical features ($P = 0.24$), and lexical features outperforming the spectral features model ($P < 0.001$). All speech-feature models performed significantly better than the demographic comparison (min $P < 0.01$), except the spectral features model ($P = 0.28$). A full overview of AUC comparison statistics is provided in Table S5.

Results were similar for predicting participant reports of having been in receipt of a prescription for medication for cognitive impairment or memory problems by a health-care provider (Figure S5 in supporting information). The strongest predictions were seen for G-match and lexical features (both AUC = 0.76; 95% CI = 0.67–0.84, and 0.68–0.84, respectively), followed by temporal features (AUC = 0.68, 95% CI = 0.58–0.78), all features combined (AUC = 0.65, 95% CI = 0.54–0.75), demographics and spectral features (AUC = 0.59,

95% CI = 0.48–0.70, and AUC = 0.55, 95% CI = 0.45–0.66, respectively). Overall, models combining features across all feature domains did not afford additional predictive power, with the AUC sitting somewhere between the most and least predictive feature sets.

4 | DISCUSSION

4.1 | Summary

Overall, the results show excellent generalizability of simple speech-based testing prediction models of early AD across diverse samples, recruitment and assessment strategies, and assessment devices.

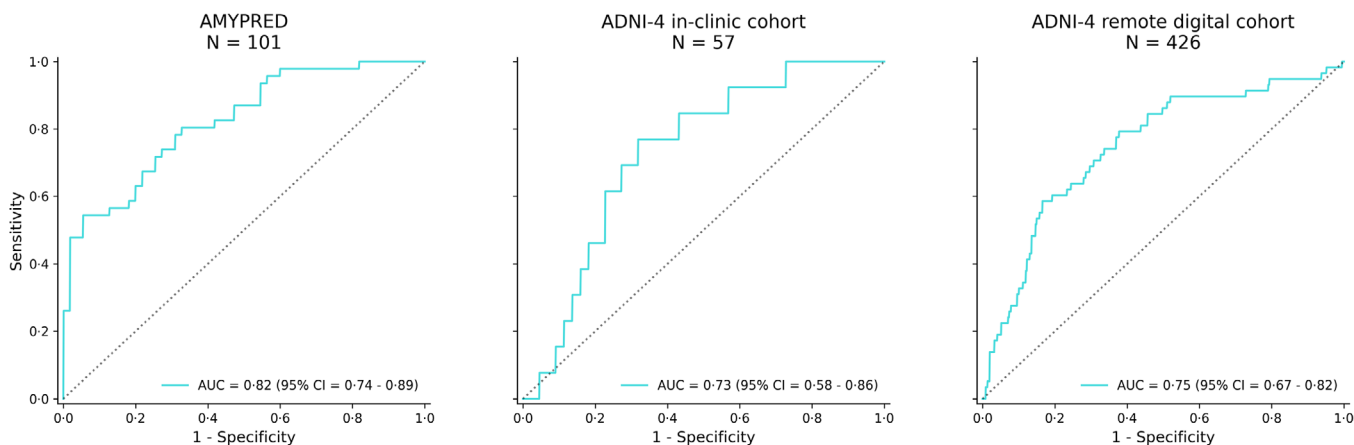


FIGURE 3 ROC curves and AUCs for the prediction of MCI, AD, or dementia diagnoses using the G-match metric (an automatically derived text similarity metric comparing source text and retelling) as a predictor in each of the three cohorts. AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; AMYPRED, Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech; AUC, area under the curve; CI, confidence interval; G-match, generalized matching algorithm; MCI, mild cognitive impairment; ROC, receiver operating characteristic

TABLE 3 Selected sensitivities (at 0.70, 0.80, 0.90), and associated specificities, PPVs, and NPVs for predicting diagnosis (clinical or self-reported), as well as their corresponding G-match thresholds in each of the three cohorts.

Cohort	Target sensitivity	Actual sensitivity	Specificity	PPV	NPV	G-match score threshold
AMYPRED	0.70	0.72	0.75	0.70	0.76	66.9
	0.80	0.80	0.67	0.67	0.80	69.9
	0.90	0.87	0.53	0.61	0.83	71.8
ADNI4 in-clinic cohort	0.70	0.69	0.73	0.43	0.89	57.7
	0.80	0.77	0.68	0.42	0.91	59.3
	0.90	0.92	0.43	0.32	0.95	69.2
ADNI4 remote digital cohort	0.70	0.71	0.69	0.27	0.94	57.8
	0.80	0.79	0.62	0.25	0.95	60.1
	0.90	0.90	0.48	0.21	0.97	64.7

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; AMYPRED, Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech; G-match, generalized matching algorithm; NPV, negative predictive values; PPV, positive predictive values.

Although all samples evaluated were tested remotely using the same tasks, the samples differed across their demographic characteristics (Table 2), country and methods of recruitment, inclusion and exclusion criteria (Table S1), application type, and breadth of devices used for testing (Section 3.2). The best-performing speech models were for G-match, currently in use in ADNI4 prescreening recommendations (to categorize participants as cognitively impaired or not based on their story recall performance), and lexical features. Generally, the pattern and strength of association of the different speech-based metrics were consistent across out-of-sample validation analyses, supporting the generalizability of the findings.

Although all models performed consistently in ROC analyses across the different cohorts evaluated, what was seen for G-match in particular was a shift in the absolute scores associated with specific sensitivity thresholds, and the overall shift in the distribution of scores between

AMYPRED and ADNI4. There may have been several contributing factors here, including practice effects as AMYPRED participants were well versed in the task during data collection. The results indicate that although the methods generalize well, thresholds in the original sample may not be broadly representative and may need to be re-evaluated and adjusted in new populations or use cases.

4.2 | Usability of Storyteller test

Most participants (88%) reported a neutral or positive experience with Storyteller, with 66% reporting some enjoyment of the test. However, user feedback was collected at the end of the assessment, which limits our understanding of the overall user experience for those who dropped out earlier. Additional efforts are now required to evaluate the

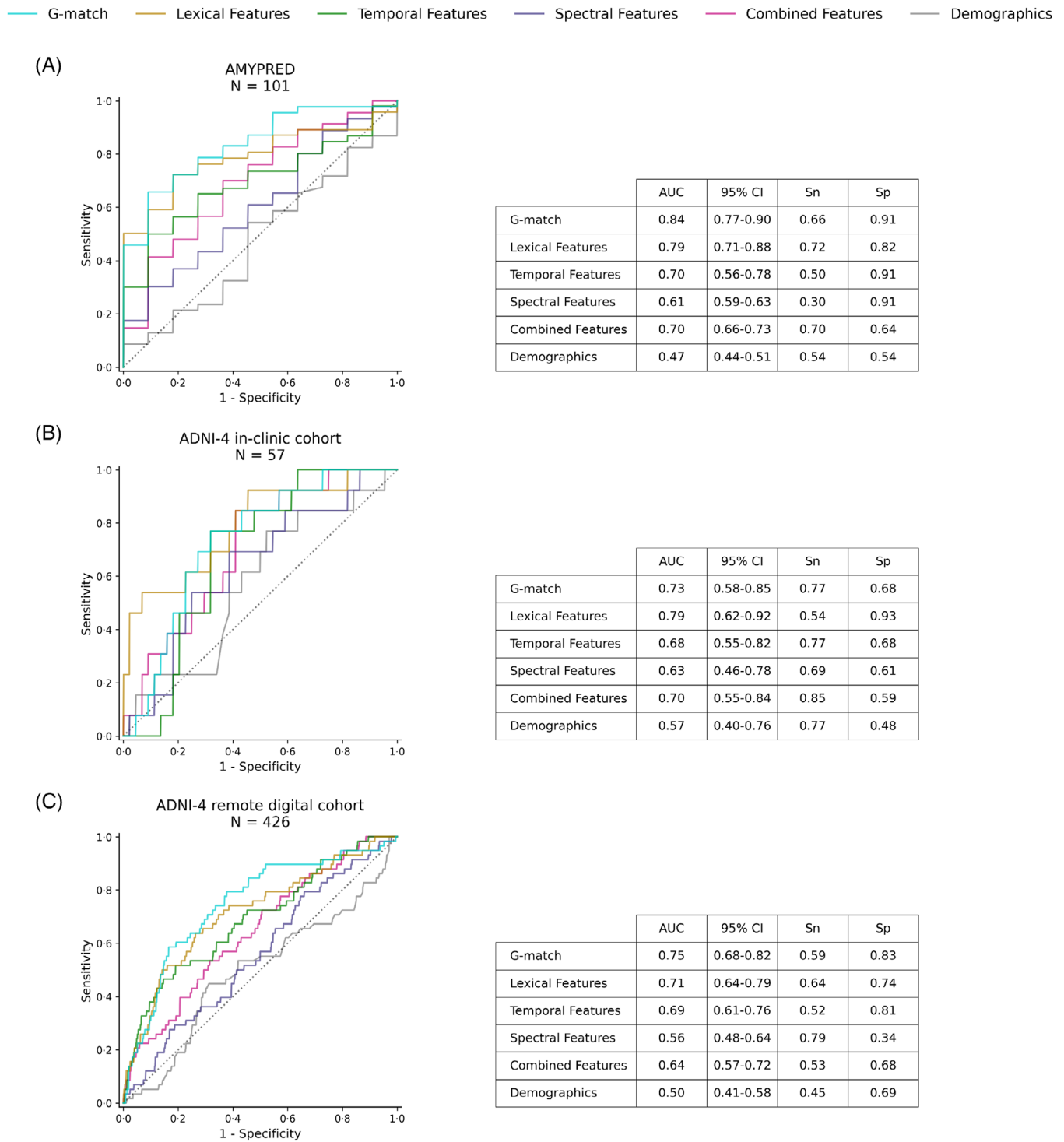


FIGURE 4 Prediction of MCI across cohorts evaluating a range of speech-based feature models, ROC curves on the left, with AUCs, 95% confidence intervals, and Sensitivity (Sn) and Specificity (Sp) at Youden index, tabulated on the right: (A) within-sample prediction of clinical diagnostic status within the AMYPRED sample (5-fold validation analysis); (B) out-of-sample prediction of MCI clinical status in the ADNI4 in-clinic cohort; (C) out-of-sample prediction of self-reported diagnosis (MCI, AD, or dementia) in the ADNI4 remote digital cohort. G-match: automatically derived text similarity metric comparing source text and retelling. ADNI, Alzheimer's Disease Neuroimaging Initiative; AMYPRED, Amyloid Prediction in Early Stage Alzheimer's Disease from Acoustic and Linguistic Patterns of Speech; AUC, area under the curve; CI, confidence interval; G-match, generalized matching algorithm; MCI, mild cognitive impairment; ROC, receiver operating characteristic

broader user experience, and reasons for dropout for participants who do not complete the test.

Modestly lower overall satisfaction was reported by older participants and those with a self-reported diagnosis. This may be related to greater difficulty in completing the tasks or navigation of the web application, due to cognitive impairment or lower familiarity with technology. Importantly, there was no evidence of greater or lesser satisfaction in relation to sex or education level, and satisfaction with the test was modestly higher in URPs than in non-URP populations, indicating that the test is generally well received across a diverse range of individuals.

4.3 | Comparison of speech models

The most strongly predictive speech-based feature models were text-matching (G-match) and lexical features, followed by temporal features and finally spectral features. The results and ranking of model performance were remarkably robust across all the analyses completed. This is generally in keeping with the AD literature, in which some of the most consistently reported speech features associated with AD may be captured in transcription,¹²⁻¹⁶ including, for example, word production and complexity, semantic content, lexical diversity, content density of speech, and the over- and under-use of certain word classes (e.g., nouns, verbs, pronouns).

Overall, a lesser signal was seen for temporal features (pauses, speech rate, and speech duration) in isolation, in keeping with some prior research,⁴² although evidence suggests that pause information may augment signal in large language models for detecting AD signatures.⁴³ Vocal acoustic changes in AD are not as commonly reported in the early stages of the disease (e.g., MCI) compared to the more progressed AD dementia stages.¹⁶ However, the more modest sensitivity of spectral and temporal features must be considered in the context of the methodological limitations for data control, caused by the remote, uncontrolled setting of assessments and variations in microphone across devices (discussed further in Section 4.6).

4.4 | Generalizability and diversity

Generalizability is of key importance when considered in the context of scalability. Speech-based screening and diagnostic testing will only be truly useful if it works well in a broad range of demographic groups. Because aspects of speech and language also reflect a range of demographic features of the speaker (their dialect, culture, social status, sex, race, and age),³¹ identifying patterns in speech and language that are consistently associated with disease signatures across different population groups is key to accurate, equitable, unbiased, and scalable prescreening.

In the context of the availability of thousands of speech features, which are curated and optimized into feature sets, there is a risk of generating spurious results.³⁰ This problem is exacerbated in the context of data sparsity, in which due to small sample sizes, validation analysis is frequently carried out using internal cross-validation techniques

(e.g., k-fold cross-validation), where there is no held-out dataset to verify the generalizability of the findings beyond the source sample in which predictive models are generated.^{25,26,44,45} Problems with internal cross-validation are common in the speech research literature, even in some of the authors' own publications.^{21,22}

There are, however, some notable exceptions to this with studies now looking across different speaker sets or languages, or holding out small subsamples for validation.²⁷⁻²⁹ Taken together with a body of work systematically mapping out patterns in the data,¹²⁻¹⁶ these studies are starting to show that speech-based testing is both consistent and generalizable.

Going forward, the field requires larger, and more diverse, samples to evaluate generalizability and equity of speech-based screening models. This may be particularly important for models of speech detecting more subtle variations in language use, that have been reported as sensitive to AD biomarker status.^{21,22,46} Larger cohorts and studies such as ADNI4 are key to validating and developing the next generation of speech biomarkers.

4.5 | Generalizability across diverse samples, testing methods, and devices

The consistency of the results from speech-based models across the three different samples supports the broad generalizability of the findings to unseen data and the sensitivity of the underlying speech features evaluated.

One of the strengths of the current study is the diversity of the methods used in the three different samples evaluated (tabulated in Table S1). Model development was carried out exclusively in AMYPRED, a racially homogenous but mixed US- and UK-English dialect-speaking sample.²¹ The participants were established research volunteers and were familiar with, and well supported in, the use of the native application that delivered their remote testing.

By contrast, in ADNI4 there are different recruitment strategies currently predominating in identifying in-clinic and remotely recruited participants. The remote digital cohort for ADNI4 is a newly recruited sample that aims specifically to improve engagement and enrollment of URP participants,¹¹ and has seen success in this approach (see Miller et al., this issue; Rivera Mindt, Arenoff et al., this issue). To date, the in-clinic cohort comprises roll-over participants from prior ADNI phases. In both ADNI4 cohorts, at the initial time of completing Storyteller, participants are unfamiliar with the task design and test set-up, and the assessment is completed fully without support or supervision. The new Storyteller application has also been developed to be administered online and is accessed via a much broader range of devices than before.

4.6 | Limitations

Although the current results are promising, there are several key limitations. The two ADNI4 cohorts have different strengths and

drawbacks: diagnostic labels in the ADNI4 in-clinic sample are externally, clinically evaluated, but the sample is currently small; by contrast, self-reported diagnostic labels in the remote digital cohort are less robust, but the sample is much larger. As recruitment continues, future analysis of the ADNI4 cohorts will allow for the evaluation of model performance in a much larger clinically defined sample. Furthermore, it will be possible to track the convergence between self-reported diagnosis and clinical diagnosis in the sample referred for in-clinic evaluations.

Although transcription error rates were modest in AMYPRED (WER = 0.10), further evaluation of transcription error rates is required in ADNI4. It is possible that some of the distributional differences in text-based analytic outputs may be affected by transcription accuracy. This area, in addition to transcription equity across demographic groups, requires much further research.

Participants completed assessments remotely and unsupervised, and unprocessed speech data were collected on a range of devices in both AMYPRED and ADNI4 studies. This will have influenced the recorded audio quality, due to variable speaker distance from the microphone, and different microphone hardware used. Analyses of spectral features were therefore restricted to those less likely to be influenced by the amplitude and intensity of the audio signal, which is most strongly affected by speaker distance and types of microphones.

Additionally, there is limited control over the presence of additional, non-task-related audio that may have been recorded (secondary speakers, laughing, coughing, throat clearing). Improvements in sensitivity may be expected under more controlled environments, particularly for spectral and temporal features which may have been more strongly influenced by these uncontrolled factors.

4.7 | Future directions

The larger data set of speech data in ADNI4 holds promise for further evaluating equity of speech-based screeners and biomarkers, allowing for investigation of the contributions of sex, race, ethnicity, and education, evaluating and improving model performance in different demographic segments.

Future challenges and opportunities lie in the context of multilingual testing. The current data presented evaluates participants in English language only, and ADNI4 will soon begin testing US-Spanish native speakers, including Storyteller in US Spanish. The multilingual transcription and analysis models used here provide a good basis for comparable results across different languages, and there is fruitful future research in evaluating cross-language speech screeners and biomarkers to predict clinical diagnostic status.

As the size of clinical speech datasets increases, the use of more advanced models that can learn complex relationships from data, such as deep learning models, becomes possible. Outside of the clinical domain, such approaches have long displaced manual feature engineering with vastly superior results, for example in natural language processing and speech recognition applications. These

models will be able to leverage not just what is already known about speech changes in disease, but may also uncover previously unknown relationships.

ACKNOWLEDGMENTS

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

CONFLICT OF INTEREST STATEMENT

C.S., U.M., E.F., and J.W. are employees of Novoic Ltd and are Novoic option holders or shareholders; E.F. and J.W. are directors of the company. M.J.M. reports no disclosures. B.A.'s only disclosure is ADNI4 grant support. R.L.N. reports grants to institutions from NIH, California Department of Health, and Genentech, Inc. M.W.W. serves on editorial boards for *Alzheimer's & Dementia*, *MRI*, and *TMRI*. He has served on advisory boards for Acumen Pharmaceutical, ADNI, Alzheon, Inc., Biogen, Brain Health Registry, Cerecin, Dolby Family Ventures, Eli Lilly, Merck Sharp & Dohme Corp., National Institute on Aging (NIA), Nestle/Nestec, PCORI/PPRN, Roche, University of Southern California (USC), NervGen. He has provided consulting to Baird Equity Capital, BioClinica, Cerecin, Inc., Cytox, Dolby Family Ventures, Duke University, Eisai, FUJIFILM-Toyama Chemical (Japan), Garfield Weston, Genentech, Guidepoint Global, Indiana University, Japanese Organization for Medical Device Development, Inc. (JOMDD), Medscape, Nestle/Nestec, NIH, Peerview Internal Medicine, Roche, T3D Therapeutics, University of Southern California (USC), and Vida Ventures. He has acted as a speaker/lecturer to The Buck Institute for Research on Aging, the China Association for Alzheimer's Disease (CAAD), the Japan Society for Dementia Research, and the Korean Dementia Society. He holds stock options with Alzheon, Inc., Alzeca, and Anven. The following entities have provided funding for academic travel: the University of Southern California (USC), NervGen, ASFNR, and CTAD Congress. Author disclosures are available in the [supporting information](#).

ETHICS STATEMENT

The AMYPRED studies were approved by institutional review boards in the relevant research authorities (UK REC reference: 20/WM/0116; US IRB reference: 8460-JGDuffy). Informed consent was taken at the study site (US) or electronically in accordance with HRA guidelines (UK). ADNI4 was approved by the Advarra Institutional Review Boards for the US (IRB reference: Pro00064250). Informed consent for the remote digital cohort was taken electronically via the online platform; consent for the in-clinic cohort was taken at the study site.

CONSENT STATEMENT

All human subjects provided informed consent.

ORCID

Caroline Skirrow  <https://orcid.org/0000-0001-8692-7787>

REFERENCES

- Fiest KM, Roberts JI, Maxwell CJ, et al. The prevalence and incidence of dementia due to Alzheimer's disease: a systematic review and meta-analysis. *Can J Neurol Sci.* 2016;43(Suppl 1):S51-S82.
- He W, Goodkind D, Kowal P. An aging world: 2015. international population reports. United States Census Bureau; 2016. Accessed April 20, 2021. Available from: <https://www.census.gov/content/dam/Census/library/publications/2016/demo/p95-16-1.pdf>
- Alzheimer's Disease International. World alzheimer report 2015. 2015 Accessed April 26, 2021. Available from: <https://www.alzint.org/resource/world-alzheimer-report-2015/>
- Sims JR, Zimmer JA, Evans CD, et al. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. *JAMA.* 2023;330(6):512-527.
- van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in early Alzheimer's disease. *N Engl J Med.* 2023;388(1):9-21.
- Cummings J, Zhou Y, Lee G, Zhong K, Fonseca J, Cheng F. Alzheimer's disease drug development pipeline: 2023. *Alzheimers Dement.* 2023;9(2):e12385.
- Mueller KD, Kosciak RL, Hermann BP, Johnson SC, Turkstra LS. Declines in connected language are associated with very early mild cognitive impairment: results from the Wisconsin registry for Alzheimer's prevention. *Front Aging Neurosci.* 2018;9:437.
- Skirrow C, Meszaros M, Meepegama U, et al. Validation of a remote and fully automated story recall task to assess for early cognitive impairment in older adults: longitudinal case-control observational study. *JMIR Aging.* 2022;5(3):e37090.
- Teunissen CE, Verberk IMW, Thijssen EH, et al. Blood-based biomarkers for Alzheimer's disease: towards clinical implementation. *Lancet Neurol.* 2022;21(1):66-77.
- Barthélemy NR, Salvadó G, Schindler SE, et al. Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. *Nat Med.* 2024;30:1085-1095.
- Weiner MW, Veitch DP, Miller MJ, et al. Increasing participant diversity in AD research: plans for digital screening, blood testing, and a community-engaged approach in the Alzheimer's disease neuroimaging initiative 4. *Alzheimers Dement.* 2023;19(1):307-317.
- Kavé G, Goral M. Word retrieval in connected speech in Alzheimer's disease: a review with meta-analyses. *Aphasiology.* 2018;32(1):4-26.
- Slegers A, Filiou R-P, Montembeault M, Brambati SM. Connected speech features from picture description in Alzheimer's disease: a systematic review. *J Alzheimers Dis.* 2018;65(2):519-542.
- Hecker P, Steckhan N, Eyben F, Schuller BW, Arnrich B. Voice analysis for neurological disorder recognition-a systematic review and perspective on emerging trends. *Front Digit Health.* 2022;4:842301.
- Mueller KD, Hermann B, Mecollari J, Turkstra LS. Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J Clin Exp Neuropsychol.* 2018;40(9):917-939.
- Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JGG. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. *Front Psychol.* 2021;12:620251.
- Cho S, Cousins K, Shellikeri S, et al. Lexical and acoustic speech features relating to Alzheimer disease pathology. *Neurology.* 2022;99:e313-e322.
- Mueller KD, Kosciak RL, Du L, et al. Proper names from story recall are associated with beta-amyloid in cognitively unimpaired adults at risk for Alzheimer's disease. *Cortex.* 2020;131:137-150.
- Hale MR, Kosciak RL, Du L, et al. Associations between semantic memory for proper names in story recall and CSF amyloid and tau in a cognitively unimpaired sample. *Alzheimers Dement.* 2022;18(S7):e059439.
- García-Gutiérrez F, Marquié M, Muñoz N, et al. Harnessing acoustic speech parameters to decipher amyloid status in individuals with mild cognitive impairment. *Front Neurosci.* 2023;17:1221401.
- Fristed E, Skirrow C, Meszaros M, et al. Leveraging speech and artificial intelligence to screen for early Alzheimer's disease and amyloid beta positivity. *Brain Commun.* 2022;4(5):fcac231.
- Fristed E, Skirrow C, Meszaros M, et al. A remote speech-based AI system to screen for early Alzheimer's disease via smartphones. *Alzheimers Dement.* 2022;14:e12366.
- Robin J, Xu M, Kaufman LD, Simpson W. Using digital speech assessments to detect early signs of cognitive impairment. *Front Digit Health.* 2021;3:749758.
- Yeung A, Iaboni A, Rochon E, et al. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimers Res Ther.* 2021;13(1):109.
- Santander-Cruz Y, Salazar-Colores S, Paredes-García WJ, Guendulain-Arenas H, Tovar-Arriaga S. Semantic feature extraction using SBERT for dementia detection. *Brain Sci.* 2022;12:270.
- Ter Huurne D, Ramakers I, Possemis N, et al. The accuracy of speech and linguistic analysis in early diagnostics of neurocognitive disorders in a memory clinic setting. *Arch Clin Neuropsychol.* 2023;38(5):667-676.
- Robin J, Xu M, Balagopalan A, et al. Automated detection of progressive speech changes in early Alzheimer's disease. *Alzheimers Dement.* 2023;15(2):e12445.
- Lindsay H, Tröger J, König A. Language impairment in Alzheimer's disease-robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Front Aging Neurosci.* 2021;13:642033.
- Fraser KC, Linz N, Li B, et al. Multilingual prediction of Alzheimer's Disease through domain adaptation and concept-based language modelling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA. Association for Computational Linguistics; 2019: 3659-3670.
- Wadle L-M, Ebner-Priemer UW. Smart digital phenotyping. *Eur Neuropsychopharmacol.* 2023;76:1-2.
- Bent T, Holt RF. Representation of speech variability. *Wiley Interdiscip Rev Cogn Sci.* 2017;8(4), e1434.
- Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 2011;7(3):270-279.
- Morris JC. The clinical dementia rating (CDR): current version and scoring rules. *Neurology.* 1993;43(11):2412-2424.

34. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state" A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12:189-198.
35. Tomaszewski Farias S, Mungas D, Harvey DJ, Simmons A, Reed BR, Decarli C. The measurement of everyday cognition: development and validation of a short form of the everyday cognition scales. *Alzheimer's Dement.* 2011;7(6):593-601.
36. Farias ST, Weakley A, Harvey D, Chandler J, Huss O, Mungas D. The measurement of everyday cognition (ecog): revisions and updates. *Alzheimer Dis Assoc Disord.* 2021;35(3):258-264.
37. Weiner M, Nosheny R, Camacho M, et al. The brain health registry: an internet-based platform for recruitment, assessment, and longitudinal monitoring of participants for neuroscience studies. *Alzheimer's and Dementia.* 2018;14(8):1063-1076.
38. Google Speech-to-text. Accessed June 25, 2024. Available from <https://cloud.google.com/speech-to-text>
39. Huggingface Evaluate. Accessed July 5, 2024. Available from <https://github.com/huggingface/evaluate>
40. Lenain R, Weston J, Shivkumar A, Fristed E. *Surfboard: Audio Feature Extraction for Modern Machine Learning.* In: Proceeding of Interspeech, Shanghai, China; 2020, 2917-2921. doi:10.21437/Interspeech.2020-2879
41. Shivkumar A, Weston J, Lenain R, Fristed E. BlaBla: linguistic feature extraction for clinical analysis in multiple languages. In: Proceedings of Interspeech 2020, Shanghai, China; 2020; 2542-2546, doi:10.21437/Interspeech.2020-2880
42. Fraser KC, Lundholm Fors K, Eckerström M, Öhman F, Kokkinakis D. Predicting MCI status from multimodal language data using cascaded classifiers. *Front Aging Neurosci.* 2019;11:205.
43. Yuan J, Cai X, Bian Y, Ye Z, Church K. Pauses for detection of Alzheimer's disease. *Front Comput Sci.* 2021;2:624488.
44. Balagopalan A, Eyre B, Robin J, Rudzicz F, Novikova J. Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. *Front Aging Neurosci.* 2021;13:635945.
45. Clarke N, Barrick TR, Garrard P. A comparison of connected speech tasks for detecting early Alzheimer's disease and mild cognitive impairment using natural language processing and machine learning. *Front Comput Sci.* 2021;3: 634360.
46. Weston J, Lenain R, Meepegama U, Fristed E. Generative pretraining for paraphrase evaluation. Generative pretraining for paraphrase evaluation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland. Association for Computational Linguistics; 2020, 1:4052-4073.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Skirrow C, Meepegama U, Weston J, et al. Storyteller in ADNI4: Application of an early Alzheimer's disease screening tool using brief, remote, and speech-based testing. *Alzheimer's Dement.* 2024;20:7248-7262. <https://doi.org/10.1002/alz.14206>