Received: 2 March 2023 | Revised: 21 April 2023 | Accepted: 27 April 2023

DOI: 10.1002/dad2.12445

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Automated detection of progressive speech changes in early Alzheimer's disease

Jessica Robin[1] | **Mengdan Xu**[1] | **Aparna Balagopalan**[1,2] | **Jekaterina Novikova**[1] | **Laura Kahn**[3,4] | **Abdi Oday**[3] | **Mohsen Hejrati**[3] | **Somaye Hashemifar**[3] | **Mohammadreza Negahdar**[3] | **William Simpson**[1] | **Edmond Teng**[3]

[1]Winterlight Labs Inc., Toronto, Ontario, Canada

[2]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[3]Genentech, Inc., South San Francisco, California, USA

[4]ReCode Therapeutics, Menlo Park, California, USA

Correspondence
Jessica Robin, Winterlight Labs, Inc., 46 Hayden St, Suite 400, Toronto, ON M4Y 1V8, Canada.
Email: jessica@winterlightlabs.com

Present address
Aparna Balagopalan, Genentech, Inc., South San Francisco, California, USA
Laura Kahn, ReCode Therapeutics, Menlo Park, California, USA

Funding information
Genentech Inc.; South San Francisco

Abstract

Speech and language changes occur in Alzheimer's disease (AD), but few studies have characterized their longitudinal course. We analyzed open-ended speech samples from a prodromal-to-mild AD cohort to develop a novel composite score to characterize progressive speech changes. Participant speech from the Clinical Dementia Rating (CDR) interview was analyzed to compute metrics reflecting speech and language characteristics. We determined the aspects of speech and language that exhibited significant longitudinal change over 18 months. Nine acoustic and linguistic measures were combined to create a novel composite score. The speech composite exhibited significant correlations with primary and secondary clinical endpoints and a similar effect size for detecting longitudinal change. Our results demonstrate the feasibility of using automated speech processing to characterize longitudinal change in early AD. Speech-based composite scores could be used to monitor change and detect response to treatment in future research.

KEYWORDS
Alzheimer's disease, digital biomarkers, language, mild cognitive impairment, natural language processing, speech

HIGHLIGHTS

- Longitudinal speech samples were analyzed to characterize speech changes in early AD.
- Acoustic and linguistic measures showed significant change over 18 months.
- A novel speech composite score was computed to characterize longitudinal change.
- The speech composite correlated with primary and secondary trial endpoints.
- Automated speech analysis could facilitate remote, high frequency monitoring in AD.

1 | BACKGROUND

Changes to speech and language have long been observed in Alzheimer's disease (AD), [1–3] and recent advancements in natural language processing have facilitated the measurement of these changes. Digital speech assessment may lead to novel and more sensitive markers of AD symptoms and progression, enhancing current methods for detecting and monitoring disease.[4–6] Most recent research using natural language processing methods to analyze the speech and language changes that occur in AD has focused on cross-sectional differentiation of participants with AD versus healthy controls,[7–12] with fewer studies on longitudinal changes.[13,14]

Cross-sectional studies validate that digital approaches can identify changes in speech and language occuring in AD. A next step is the identification of longitudinal changes that align with disease progression, which is essential for the development of endpoints suitable for monitoring change over time. Existing assessments of language abilities are often subjective in nature, limited in sensitivity, and burdensome in administration and scoring. Additionally, many existing assessments rely on structured speech tasks, such as verbal fluency, which may be less ecologically valid than analyzing spontaneous natural speech.[15–17] As such, current studies and clinical trials may be insensitive to decline with disease progression and/or potential improvements following effective therapies. Leveraging digital technologies to develop a novel speech-based endpoint could produce measures that are more sensitive to subtle speech and language changes and better suited to remote, lower burden, and higher frequency assessments. Such accessible endpoints could promote more inclusive and efficient clinical research that better captures the effects of disease progression or successful intervention on speech and language behaviors in AD.[18–20]

In this study, we analyzed longitudinal speech recordings from an early AD cohort and developed a novel composite measure to characterize progressive changes to speech and language patterns. The performance of this novel endpoint was compared to established clinical and cognitive assessments to determine the utility of this approach for developing new methods to monitor disease progression and detect response to treatment.

RESEARCH IN CONTEXT

1. **Systematic review**: We searched databases including PubMed and GoogleScholar for previous research relating to "speech", "language", "Alzheimer's disease," and "mild cognitive impairment". While many studies on cross-sectional differences in speech and language patterns in AD exist, few measure longitudinal changes with disease progression. Relevant studies are cited and summarized.
2. **Interpretation**: This study uses open-ended, naturalistic speech and automated speech analysis methods to determine the aspects of speech and language that change over time in early AD. This study adds to previous research on this topic by proposing a novel composite score reflecting both acoustic and linguistic changes to speech that occur in early AD and progress.
3. **Future directions**: The novel speech composite proposed in this study requires further validation and replication in larger and more diverse samples. Validation against other measures of disease progression and comparisons to longitudinal changes in healthy aging are also needed to further verify the disease relevance of this measure.

2 | METHODS

2.1 | Participants

We analyzed data from a subset of English-speaking US-based participants in the Tauriel phase 2 trial of semorinemab (NCT03289143).[21] All participants were 50-80 years old, met the National Institute on Aging/Alzheimer's Association core clinical criteria for probable AD dementia (mild AD) or mild cognitive impairment due to AD (prodromal AD), demonstrated evidence of cerebral AD pathology confirmed by amyloid-beta (Aβ) positron emission tomography (PET) scan ([^{18}F]florbetaben, [^{18}F]florbetapir, [^{18}F]flutemetamol, or [^{18}F]NAV4694 via visual read) or Aβ(1-42) levels [$\leq$1000 pg/mL, Elecsys β-amyloid (1-42) cerebrospinal fluid (CSF) immunoassay; Roche Diagnostics, Penzberg, Germany], and had scores on the Mini Mental State Examination (MMSE)[22] of $\geq$20 points and a Clinical Dementia Rating (CDR)[23] Global Score of 0.5 or 1. Although this was a double-blind, randomized, placebo-controlled trial, given the similar performance of the semorinemab and placebo arms on all clinical outcome measures,[21] participants from all study arms were combined into a single group.

This study was approved by each center's Institutional Review Board/Ethics Committee and conducted in accordance with the Declaration of Helsinki and the International Conference on Harmonization E6 Guidelines for Good Clinical Practice. All participants and/or their legally authorized representatives provided written informed consent.

2.2 | Clinical assessments

As part of the clinical trial design, participants were assessed by a trained clinical rater at screening, baseline, and at 6-, 12-, 17-, and 18-month follow up assessments, between October 2017 and July 2020. Primary and secondary endpoints included cognitive and functional measures including the CDR Sum of Boxes (CDR-SB), the 13-item version of Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog13),[24] the Repeatable Battery for Assessment of Neuropsychological Status (RBANS),[25] the MMSE, and the Alzheimer's Disease Cooperative Study Group-Activities of Daily Living Inventory (ADCS-ADL).[26] Baseline demographics and cognitive assessment scores are listed in Table 1.

TABLE 1 Baseline demographics and scores on cognitive and functional assessments.

	Training dataset ($n = 101$)	Testing dataset ($n = 29$)
Sex, female *n* (%)	58 (57)	17 (59)
Age, mean years (SD)	69.3 (7.0)	68.8 (8.5)
Education, high school graduate or more *n* (%)	100 (99)	29 (100)
Race/ethnicity, white *n* (%)	96 (95)	29 (100)
APOE status, $\varepsilon 4+$ *n* (%)	77 (76)	19 (66)
CDR-SB, total score (SD)	4 (1.63)	4 (1.48)
ADAS-Cog, total score (SD)	28 (7.23)	26 (7.81)
RBANS, total score (SD)	64 (12.02)	68 (12.81)
MMSE, total score (SD)	23 (3.42)	24 (3.62)
ADCS-ADL, total score (SD)	69 (5.44)	70 (5.87)

Note: A separate sample ($n = 29$) from the same trial, matched on demographic and clinical characteristics, was held out as an independent test dataset.

Abbreviations: ADAS-Cog, Alzheimer's Disease Assessment Scale–Cognitive Subscale; ADCS-ADL, Alzheimer's Disease Cooperative Study–Activities of Daily Living Scale.; CDR-SB, Clinical Dementia Rating–Sum of Boxes; MMSE, Mini-Mental State Examination; RBANS, Repeatable Battery for the Assessment of Neuropsychological Status; SD, standard deviation.

2.3 | Speech sample processing

Participants were audio recorded on tablet computers (Virgil platform, WCG MedAvante-ProPhase, Hamilton NJ) as they completed the CDR interview, which enabled passive collection of real-world speech samples with no additional participant or assessor burden. Trained transcriptionists listened to each interview and divided the audio into segments based on the section of the CDR interview. Recordings were diarized to isolate and identify the participant's speech relative to the interviewer's speech. The participant's speech was transcribed into text transcripts, including identifiable words, unidentifiable words, and filled (e.g., "um", "uh") and unfilled (i.e., silent) pauses. Transcriptionists identified if CDR questions were skipped, repeated, or out of order.

Transcriptionists flagged any recordings that: were void of participant speech, contained non-English participant speech, or had very poor audio quality (e.g., participant inaudible for more than 50% of the recording); these recordings were not analyzed further. Of a total of 1100 audio samples processed, 178 recordings (16%) could not be analyzed: 113 (10%) due to poor audio quality, 53 (5%) due to inaudible or missing participant speech, 6 (0.5%) due to spoken language other than English, and 6 (0.5%) due to other errors.

Speech analysis focused on the section of the CDR interview in which participants are asked to describe recent experiences within the past week and month. These segments represent the most naturalistic and open-ended speech captured in the CDR interview and were found to contain the longest speech samples (mean audio duration in seconds = 216.64, SD = 125.85 at study baseline). Segmented, diarized audio samples, containing only the participant's speech, and their accompanying text transcripts were analyzed using the Winterlight platform (www.winterlightlabs.com), which uses machine learning and natural language processing techniques to generate over 500 individual variables describing the speech and language patterns of any speech recording. This pipeline performs data processing and feature extraction using Python-based standard acoustic and language processing libraries (e.g., spaCy, Stanford parser, Praat/Parselmouth[27–30]) and custom code.

The extracted speech variables quantify the acoustic and linguistic properties of speech, making it possible to measure acoustic (e.g., properties of the sound wave), lexical (e.g., rates and types of words used), semantic (e.g., semantic relatedness of utterances), and syntactic (e.g., grammatical constructions and complexity) features. Previous research using this pipeline has used speech to distinguish participants with AD from healthy controls speech and explore which specific features are associated with AD symptomatology.[9,31–34]

2.4 | Statistical analysis

The dataset was split into training (78% of participants) and testing (22% of participants) datasets, for the purposes of developing and testing the composite score on independent data. Exploratory analyses on the training dataset using linear mixed effects models were performed to determine which aspects of speech and language exhibited consistent changes over the 18-month duration of the trial. The testing dataset was unseen during exploratory analyses and composite development. For each speech feature, a linear mixed model was tested with fixed effects of time (baseline, 6-months, 12-months, 18-months), age, sex, and level of education, and random intercepts by subject. Linear mixed effects models were also used to test the effect of time on clinical and speech composite scores. Test-retest reliability was evaluated between screening and baseline assessments and between 17- and 18-month assessments using intraclass correlations (ICC). Correlations between features, and between features and clinical scores, were evaluated using Pearson correlations.

To combine the selected features into a composite score, positive or negative equal weights were assigned to each feature, based on the direction of the time effect in the linear model. To combine speech features that have different units, each feature was z-scored by subtracting the mean value of the speech feature across all participants and timepoints, and then dividing by the standard deviation. Once all the features were standardized as z-scores, their signs were adjusted so that they all progressed in the same direction over time and they were summed to form a single composite score.

Analyses were replicated in the test dataset to test the generalizability of results. Statistical analyses were performed using R Statistical Software version 4.1.1,[35] with R packages tidyverse 1.3.1 [36] for data cleaning and processing, lmerTest 3.1-3[37] for linear mixed models, irr 0.84.1 [38] for intraclass correlation tests and ggplot2 3.3.5[39] for visualizations.

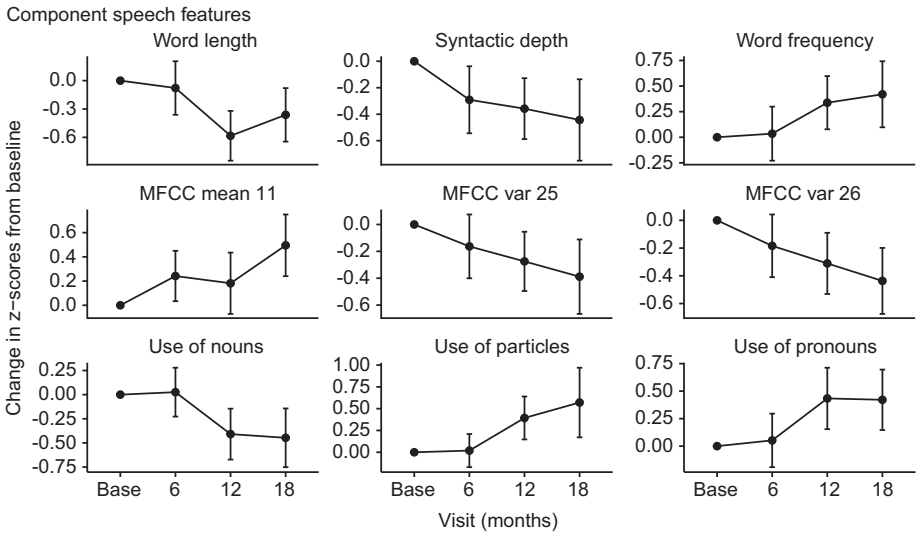

FIGURE 1 Longitudinal change in speech features. Mean change from baseline at each assessment for selected speech features with significant linear change over the course of the study. Error bars represent 95% confidence intervals.

2.5 | Data availability

Qualified researchers may request access to individual patient level data through the clinical study data request platform (https://vivli.org/). Further details on Roche's criteria for eligible studies are available here: https://vivli.org/members/ourmembers/. For further details on Roche's Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see https://www.roche.com/innovation/process/clinical-trials/data-sharing/.

3 | RESULTS

3.1 | Speech feature selection

The trajectory of individual speech features over time was examined using linear models, selecting those that showed evidence of consistent longitudinal change over the duration of the trial for further analysis. Nine speech features had effects of time significant at $p < 0.001$, suggesting consistent and progressive change over the study period.

The trajectories of the selected nine speech features are shown in Figure 1 and Table 2. Six of the features were linguistic features, representing word length, word frequency, syntactic depth, use of nouns, use of particles, and pronoun-to-noun ratio. Word frequency was calculated by averaging the estimated frequency of each word based on published norms,[40] and syntactic depth was calculated by averaging the number of levels in the syntactic tree representation of each utterance. All of the trajectories are in the expected directions, with participants using shorter and more frequent words, simpler sentence syntax, fewer nouns, and more particles and pronouns over time. The three other features represent acoustic aspects of the speech sample derived from transformations of the power spectrum of the recording. These Mel-frequency Cepstral Coefficient (MFCC) features correspond to the mean of the 11th MFCC coefficient (MFCC mean 11), the variance of the first derivative of the 11th MFCC coefficient (MFCC var 25) and the variance of the first derivative of the 12th MFCC coefficient (MFCC var 26).

As a validity check, intra-class correlations (ICC) were computed for selected speech features based on the screening and baseline assessments (up to 8-week interval per protocol), and the 17- and 18-month sessions. The ICC estimates for the nine features showing linear change over time ranged from 0.20 to 0.70 (Table 2). The MFCC variance features had the highest and most consistent ICC estimates (> 0.65 for all comparisons) while the linguistic features had more variable ICC estimates. The overall lower test-retest reliability of the linguistic features is likely attributable to the open-ended, unstructured nature of the speech task, with content varying across individuals and assessments.

To determine if selected features were measuring independently changing aspects of speech and language, correlations were computed between the selected speech features. Features had low to moderate correlations with one another (r = 0.01–0.65), with the exception of the two MFCC variance features and the noun and pronoun-to-noun ratio features, with both pairs having high correlations (r > 0.9). Since only two pairs of features had high correlations, we determined that the speech features were mostly independent from one another and were largely measuring unique aspects of underlying speech and language patterns.

3.2 | Speech composite generation and validation

3.2.1 | Composite score generation

The objective of this analysis was to create a novel composite score combining the aspects of speech that changed over time in this study cohort, aiming to maximize sensitivity to speech and language changes

TABLE 2 Effect of change over time, and test-retest reliability for selected speech features with significant linear effects of change over time.

Speech feature	Effect of time	ICC (Screening/Baseline)	ICC (17-/18-month assessment)
MFCC variance 26	$\beta = -0.013, p < 0.0001$	ICC = 0.70, $p < 0.0001$	ICC = 0.67, $p < 0.0001$
Average word length	$\beta = -0.032, p < 0.0001$	ICC = 0.35, $p = 0.0002$	ICC = 0.37, $p = 0.001$
MFCC variance 25	$\beta = -0.013, p < 0.0001$	ICC = 0.69, $p < 0.0001$	ICC = 0.67, $p < 0.0001$
Pronoun-to-noun ratio	$\beta = 0.015, p = 0.0001$	ICC = 0.47, $p < 0.0001$	ICC = 0.33, $p = 0.004$
Use of particles	$\beta = 0.003, p = 0.0003$	ICC = 0.30, $p = 0.002$	ICC = 0.38, $p = 0.0009$
Use of nouns	$\beta = -0.005, p = 0.0003$	ICC = 0.49, $p < 0.0001$	ICC = 0.20, $p = 0.05$
Word frequency	$\beta = 0.017, p = 0.0004$	ICC = 0.51, $p < 0.0001$	ICC = 0.38, $p = 0.0009$
Average syntactic depth	$\beta = -0.039, p = 0.0007$	ICC = 0.42, $p < 0.0001$	ICC = 0.67, $p < 0.0001$
MFCC mean 11	$\beta = 0.005, p = 0.0008$	ICC = 0.29, $p = 0.003$	ICC = 0.29, $p = 0.008$

Abbreviations: ICC, intraclass correlations; MFCC, Mel-frequency cepstral coefficient.

that can be detected in early AD. As described in the methods, selected speech features were standardized, assigned a positive or negative weight based on the direction of the time effect, and summed to create a composite score. The estimated test-retest reliability of the resulting composite score was ICC = 0.55 ($p < 0.0001$) between 17 and 18 months and ICC = 0.50 ($p < 0.0001$) between screening and baseline.

3.2.2 | Comparison with clinical scores

To compare the effect of change over time for the speech composite with other trial endpoints, all clinical scores and the speech composite score were standardized (z-scored) by subtracting the mean and dividing by the standard deviation of each score across all participants and timepoints. The effect of change over time was tested for all standardized scores using linear mixed models with fixed effects of time, age, sex, and level of education, and random intercepts by subject. Compared to clinical endpoints, the speech composite score ($\beta = 0.29$, $p < 0.0001$) had similar effects of change over time to the CDR-SB ($\beta = 0.30, p < 0.0001$) and ADCS-ADL ($\beta = -0.30, p < 0.0001$), with a numerically larger effect of time than the ADAS-Cog ($\beta = 0.22$, $p < 0.0001$), RBANS ($\beta = -0.15$, $p < 0.0001$), and MMSE ($\beta = -0.23$, $p < 0.0001$) (Figure 2A).

Since the speech composite is ostensibly measuring language abilities, comparisons were also made with the language subscores of the clinical endpoints, including the Language Index ($\beta = -0.19, p < 0.0001$) from the RBANS, the Spoken Language Ability (SLA) ($\beta = 0.09, p = 0.02$) and Word Finding Difficulty (WFD) scores ($\beta = 0.12$, $p = 0.007$) from the ADAS-Cog, and a previously published language composite from the ADAS-Cog[41] ($\beta = 0.16$, $p < 0.0001$), which includes the SLA, WFD, Naming Objects and Fingers, Language Comprehension, and Remembering Test Instructions items. The speech composite score had numerically greater effects of change over time relative to the other available language scores (Figure 2B).

Significance testing of the differences between the time effects across the different endpoints and subscores was conducted by computing the 95% confidence interval of the time effect and comparing across measures. The slope of change over time for the speech composite was significantly greater than the slope for the RBANS and the ADAS-Cog language subscales, but the other differences did not reach significance (Figure 2C).

Correlations were computed between change in the speech composite score and change in the clinical scores from baseline to endpoint (18 months). Change in the speech composite had the highest correlations with change on the ADAS-Cog WFD (r = 0.49, df = 52, $p = 0.0002$), CDR-SB (r = 0.45, df = 52, $p = 0.0006$), and ADAS-Cog language composite (r = 0.44, df = 52, $p = 0.0008$), with increases in all four scores indicating greater impairment (Figure 3A). Correlations between the speech composite score and clinical scores at baseline were also evaluated (Figure 3B). At baseline, the speech composite score exhibited the strongest correlations with the CDR-SB (r = 0.37, df = 98, $p = 0.0002$), MMSE (r = −0.32, df = 98, $p = 0.001$), and ADCS-ADL (r = −0.29, df = 98, $p = 0.003$), again with higher speech composite scores consistent with greater impairment. All reported correlations remain significant using an adjusted significance threshold of $\alpha = 0.0056$, Bonferroni-corrected for multiple comparisons.

3.2.3 | Generalizability of speech composite

To test generalizability, the composite score was computed in the held-out testing dataset. The testing dataset contained 29 subjects at baseline and 21 subjects at endpoint. To calculate the speech composite, each component feature was standardized using the means and standard deviations from the training set, signs were adjusted, and features were summed to compute the composite score, which was then z-scored using the mean and standard deviation from the training set. All individual features had the same direction of change from baseline to endpoint in the test and training datasets.

The overall effect of change over time on the composite score was comparable in the testing dataset ($\beta = 0.26$, $p < 0.0001$), as were the ICC values (screening-baseline: ICC = 0.69, $p < 0.0001$; 17–18 months: ICC = 0.64, $p = 0.001$), supporting the generalizability of the speech composite to the held-out dataset, which was not used in selecting the

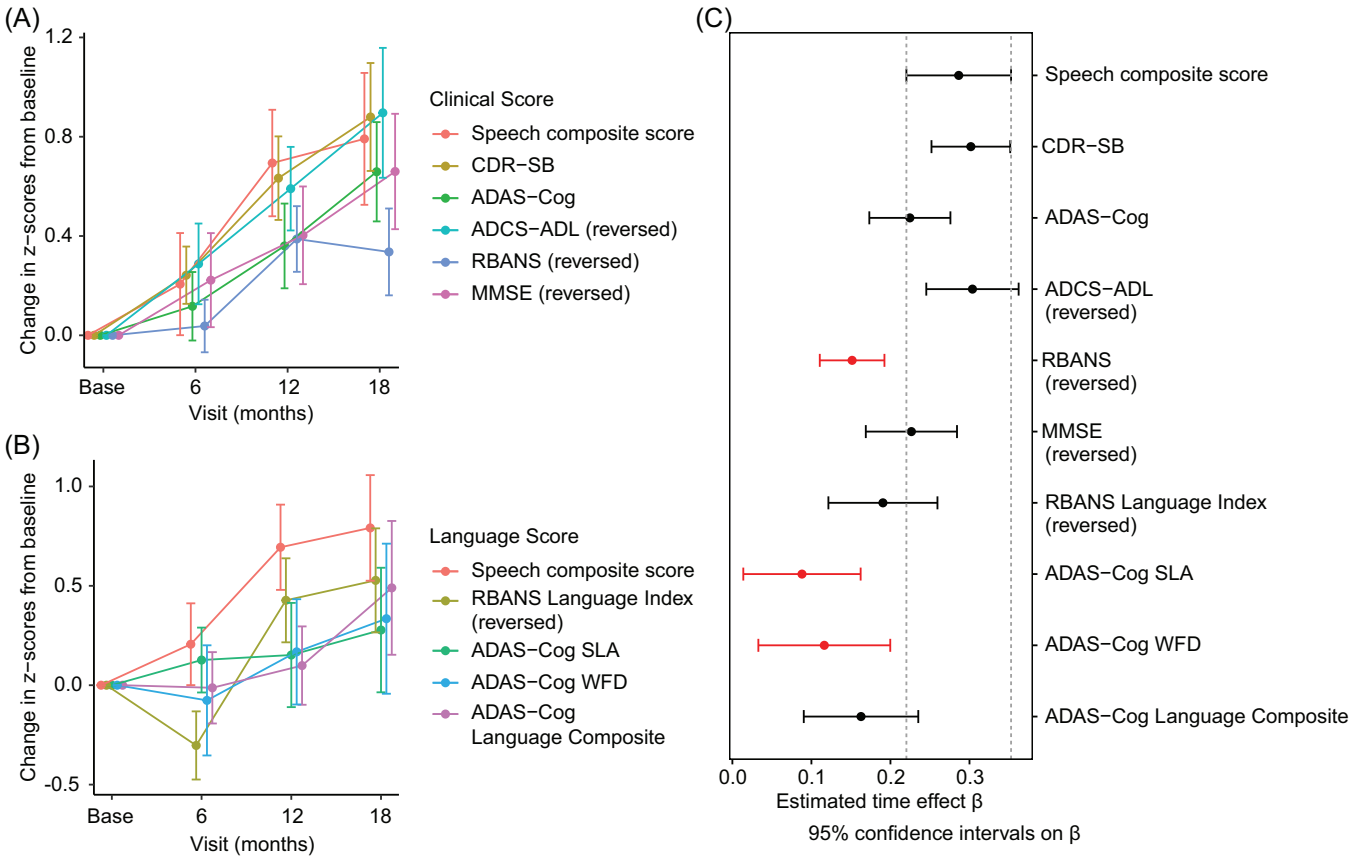

FIGURE 2 Speech composite score. Mean change from baseline at each assessment for the speech composite score compared with (A) clinical assessment total scores ad (B) clinical language subscores. Error bars represent 95% confidence intervals for the mean of change. (C) The plot represents estimated time effects (slopes) from linear mixed models using standardized values, along with the 95% confidence intervals (error bars) for the estimated time effects. The 95% confidence interval for the composite score is indicated by dotted lines. The scores that have significantly different effects of time compared to the speech composite score are colored in red.

component speech features. Importantly, the trajectory of change over time also appeared similar in the testing dataset, albeit with greater variance due to the smaller sample size (Figure 4).

4 | DISCUSSION

In this study of individuals with prodromal-to-mild AD, an automated speech processing pipeline was used to identify progressive changes in speech and language patterns over an 18-month trial period and to derive a new speech-based composite score from short, open-ended speech samples. This composite score was designed to maximize sensitivity to speech and language changes that occur in early AD, and it could be used as an endpoint to track changes in language patterns in future AD studies and trials. This measure is low-burden, requiring only a few minutes of open-ended speech from the patient, and its computation is objective, making it well-suited for remote or higher frequency assessment. The analysis pipeline included human transcribers to segment, diarize, and transcribe the audio with the highest accuracy, but using automatic speech recognition (ASR) algorithms to perform these steps could enable the computation of the composite score to be fully automated, allowing faster processing and greater scalability.

Exploratory, data-driven methods were utilized to identify speech features showing longitudinal change. Many of the selected features are consistent with previous research on speech and language changes in AD. In previous works using the same processing pipeline,[9,31] in an independent sample of individuals with AD and healthy controls, AD was associated with shorter word length, higher pronoun-to-noun ratio, and reduced use of nouns. These findings are broadly consistent with clinical observations of so-called "empty speech" in AD, which contains fewer content words and more circumlocutions.[1] In the present study, these same measures were shown to progressively worsen over time in early AD participants. Consistent with these results, other prior work has shown evidence for reduced semantic content and simplified syntax in AD, consistent with the noun and pronoun features and the reductions in syntactic depth.[2,14,42–44] In a systematic review of classification studies differentiating AD from MCI and/or healthy controls based on speech, changes to semantic and lexical features including word length, word frequency, and use of word types (including nouns and pronouns) were observed,[11] consistent with the features showing progression over time in this study. While the acoustic features

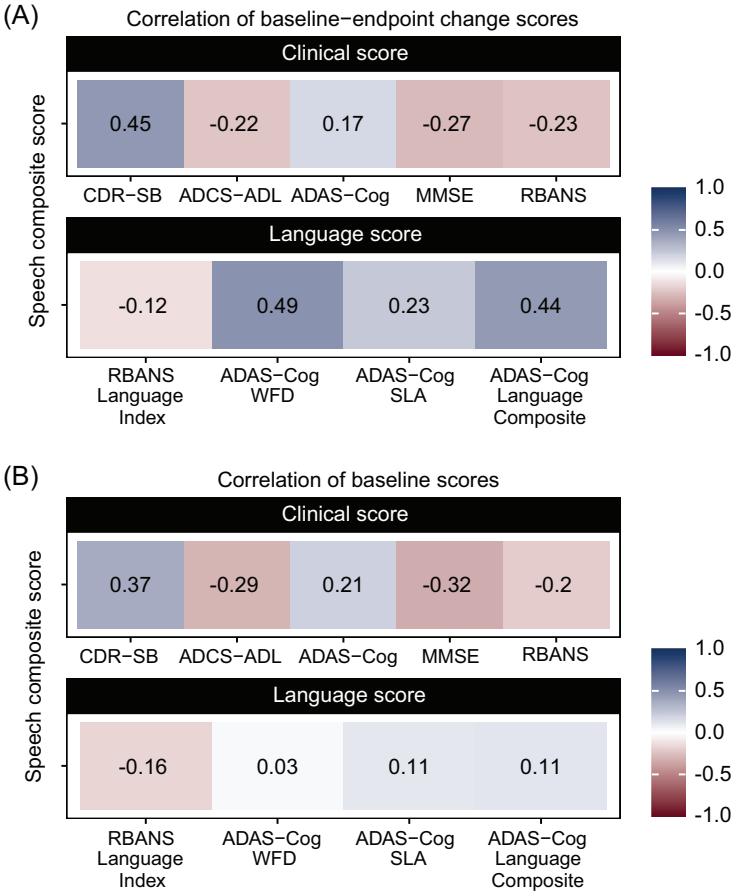

FIGURE 3 Associations with clinical scores. Pearson correlations calculated between speech composite score and clinical and language scores, using (A) longitudinal change scores and (B) baseline scores.

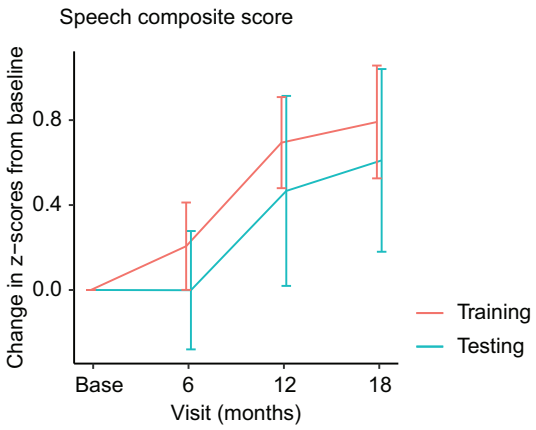

FIGURE 4 Generalizability of the speech composite. Comparison of the speech composite score in the training dataset and a held-out, independent testing dataset ($n = 29$) from the same cohort. The composite score had a similar trajectory of change from baseline over 18 months and similar test-retest reliability in both datasets. Error bars represent 95% confidence intervals.

identified in this study are less frequently reported in the literature and are harder to interpret, changes to acoustic features and vocal prosody have been consistently observed in AD, as evidenced by successful classification of AD based only on acoustic features in previous works.[45–47]

The selected variables cover several domains of speech and language, including vocabulary, word types, syntax, and acoustics. These differing variables may reflect different underlying deficits; for example, the decrease in nouns could reflect semantic memory impairments while simpler syntax could reflect overall cognitive decline. Further work is needed to replicate and further validate the speech-based composite score, the selected component variables, and link those to underlying cognitive or neurological changes. Similarly, while we chose to weight the variables equally in this study, assigning weights based on the magnitude of each variable's effect could help to tune the composite and increase sensitivity. Our initial validation results showed that novel speech composite score correlated with other measures of cognition at baseline, and that the degree of longitudinal change correlated with change on other language subscales, but more specific relationships remain a question for further study.

While these results are promising, there are a number of factors that may limit their interpretation. First, a healthy control comparison group was not included in the analysis, so the possibility cannot be eliminated that the observed longitudinal changes to speech and language patterns may reflect aging alone rather than disease progression. Future research should assess the composite score's performance in both healthy controls and across a range of AD severity. Second, while the sample size was comparable or larger than most of the study cohorts used in prior speech and language research in AD, it is still relatively small for model building purposes, and this work would benefit

from replication in a larger population. There may have been heterogeneity in the sample relating to AD phenotypes such as logopenic progressive aphasia or posterior cortical atrophy. The inclusion criteria for the study biased the sample toward a typical amnestic AD phenotype, but we cannot rule out the possibility that a small number of logopenic progressive aphasia patients may have also been included. Additionally, the sample consisted primarily of white, English-speaking participants in the United States with at least a high school diploma. This sample does not adequately reflect the global diversity of individuals with AD; different accents, as well as ethnic, educational, and linguistic backgrounds could have important implications for the language patterns being studied. Further work should explore whether the speech patterns observed in this study generalize to other groups with broader demographic characteristics and/or languages spoken. Finally, while the recent experience question in the CDR interview is open-ended and naturalistic, it still uses a specific prompt and was collected in a structured and controlled clinical interview setting. If the observed changes to speech and language are detectable in other forms of open-ended speech and other languages, this would increase the flexibility of this measure for use in non-clinical settings and broader cultural contexts.

To conclude, the results from this study represent the first steps in developing a novel speech-based measure to characterize progressive acoustic and linguistic changes that occur in AD. Such a measure could be used for disease detection and monitoring, and as an endpoint in future clinical research. For example, short, remotely administered, speech-based assessments could be collected over the course of a clinical trial to monitor disease progression and possible response to therapy. Further validation is required to replicate, test the generalizability and clinical meaningfulness of this measure.[48–50]

ACKNOWLEDGMENTS

The authors thank all of the study participants and their families, and all of the site investigators, study coordinators, and staff. This work was supported by Genentech, Inc. This study was funded by Genentech Inc., South San Francisco, CA.

CONFLICT OF INTEREST STATEMENT

J.R., M.X., J.N., and W.S. are full-time employees of Winterlight Labs, Inc. and have financial interests in Winterlight Labs, Inc. A.B. was an employee of Winterlight Labs, Inc. at the time the work was completed. A.O., M.H., S.H., M.N., and E.T. are full-time employees of Genentech, Inc. and shareholders in F. Hoffmann La Roche, Ltd. L.K. was an employee of Genentech, Inc. at the time the work was completed.

CONSENT STATEMENT

All human subjects provided informed consent in this study.

REFERENCES

1. Appell J, Kertesz A, Fisman M. A study of language functioning in Alzheimer patients. *Brain Lang*. 1982;17(1):73-91.
2. Croisile B, Ska B, Brabant MJ, et al. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain Lang*. 1996;53(1):1-19.
3. Illes J. Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's. *Brain Lang*. 1989;37(4):628-642.
4. Gold M, Amatniek J, Carrillo MC, et al. Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer's disease clinical trials. *Alzheimers Dement Transl Res Clin Interv*. 2018;4:234-242.
5. Kaye J, Amariglio R, Au R, et al. Using digital tools to advance Alzheimer's drug trials during a pandemic: the EU/US CTAD Task Force. *J Prev Alzheimers Dis*. 2021:1-7. Published online.
6. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *Npj Digit Med*. 2019;2(1):9.
7. Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: a review. *Front Psychol*. 2017;8.
8. de la Fuente Garcia S, Ritchie CW, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J Alzheimers Dis*. 2020;78(4):1547-1574.
9. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. Garrard P, ed. *J Alzheimers Dis*. 2015;49(2):407-422.
10. Liu Z, Paek EJ, Yoon SO, Casenhiser D, Zhou W, Zhao X. Detecting Alzheimer's disease using natural language processing of referential communication task transcripts. *J Alzheimers Dis*. 2022;86(3):1385-1398.
11. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J Am Med Inform Assoc*. Published online September 14, 2020:ocaa174.
12. Pulido MLB, Hernández JBA, Ballester MÁF, González CMT, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: a review. *Expert Syst Appl*. Published online January 2020:113213.
13. Stegmann G, Hahn S, Bhandari S, et al. Automated semantic relevance as an indicator of cognitive decline: out-of-sample validation on a large-scale longitudinal dataset. *Alzheimers Dement Diagn Assess Dis Monit*. 2022;14(1).
14. Mueller KD, Koscik 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.
15. Bertola L, Mota NB, Copelli M, et al. Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Front Aging Neurosci*. 2014;6.
16. Chen L, Asgari M, Gale R, Wild K, Dodge H, Kaye J. Improving the assessment of mild cognitive impairment in advanced age with a novel multi-feature automated speech and language analysis of verbal fluency. *Front Psychol*. 2020;11:535.
17. König A, Linz N, Tröger J, Wolters M, Alexandersson J, Robert P. Fully automatic speech-based analysis of the semantic verbal fluency task. *Dement Geriatr Cogn Disord*. 2018;45(3-4):198-209.
18. Dodge HH, Zhu J, Mattek NC, Austin D, Kornfeld J, Kaye JA. Use of high-frequency in-home monitoring data may reduce sample sizes needed in clinical trials. Quinn TJ, ed. *PLOS ONE*. 2015;10(9):e0138095.
19. Khozin S, Coravos A. Decentralized trials in the age of real-world evidence and inclusivity in clinical investigations. *Clin Pharmacol Ther*. 2019;106(1):25-27.
20. 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*. Published online October 9, 2022:alz.12797.
21. Teng E, Manser PT, Pickthorn K, et al. Safety and efficacy of Semorinemab in individuals with prodromal to mild alzheimer disease: a randomized clinical trial. *JAMA Neurol*. Published online June 13, 2022.
22. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state." *J Psychiatr Res*. 1975;12(3):189-198.
23. Morris JC. The clinical dementia rating (CDR). *Neurology*. 1993;43(11):2412-2412-a.
24. Mohs RC, Knopman D, Petersen RC, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's disease assessment scale that broaden its scope. *Alzheimer Dis Assoc Disord*. 1997;11.
25. Randolph C, Tierney MC, Mohr E, Chase TN. The repeatable battery for the assessment of neuropsychological status (RBANS): preliminary clinical validity. *J Clin Exp Neuropsychol*. 1998;20(3):310-319.
26. Galasko D, Bennett D, Sano M, et al. An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. *Alzheimer Dis Assoc Disord*. 1997;11.
27. Boersma P, Weenink D. Praat: Doing Phonetics by Computer. Published online 2010.
28. Chen D, Manning C. A Fast and Accurate Dependency Parser using Neural Networks. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014:740-750.
29. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Published online 2017.
30. Jadoul Y, Thompson B, de Boer B. Introducing Parselmouth: a python interface to Praat. *J Phon*. 2018;71:1-15.
31. 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;13:189.
32. 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.
33. Yancheva M, Fraser K, Rudzicz F. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In: *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics; 2015:134-139.
34. 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.
35. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020.
36. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw*. 2019;4(43):1686.
37. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. *J Stat Softw*. 2017;82(13):1-26.
38. Gamer M, Lemon J, Singh IFP. *Irr: Various Coefficients of Interrater Reliability and Agreement*.; 2019.
39. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag; 2016.
40. Brysbaert M, New B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav Res Methods*. 2009;41(4):977-990.
41. Verma N, Beretvas SN, Pascual B, Masdeu JC, Markey MK, The Alzheimer's Disease Neuroimaging Initiative. New scoring methodology improves the sensitivity of the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) in clinical trials. *Alzheimers Res Ther*. 2015;7(1):64.
42. Ahmed S, de Jager CA, Haigh AM, Garrard P. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology*. 2013;27(1):79-85.
43. Ahmed S, Haigh AMF, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*. 2013;136(12):3727-3737.
44. 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.
45. Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. 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.
46. Meilán JJG, Martínez-Sánchez F, Martínez-Nicolás I, Llorente TE, Carro J. Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. *Behav Neurol*. 2020;2020:1-10.
47. Meilán JJG, Martínez-Sánchez F, Carro J, López DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement Geriatr Cogn Disord*. 2014;37(5-6):327-334.
48. Goldsack JC, Coravos A, Bakker JP, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *Npj Digit Med*. 2020;3(1):55.
49. Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M. Evaluation of speech-based digital biomarkers: review and recommendations. *Digit Biomark*. Published online October 19, 2020:99-108.
50. Manta C, Patrick-Lake B, Goldsack JC. Digital measures that matter to patients: a framework to guide the selection and development of digital measures of health. *Digit Biomark*. 2020;4(3):69-77.

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: Robin J, Xu M, Balagopalan A, et al. Automated detection of progressive speech changes in early Alzheimer's disease. *Alzheimer's Dement*. 2023;15:e12445. https://doi.org/10.1002/dad2.12445