Journal of Alzheimer's Disease
2025, Vol. 107(1) 154–169
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/13872877251359967
journals.sagepub.com/home/alz
Sage | IOS Press

Associations between digital speech features of automated cognitive tasks and trajectories of brain atrophy and cognitive decline in early Alzheimer's disease

Qingyue Li[1], Stefanie Koehler[2], Alexandra Koenig[3,4], Martin Dyrba[2], Elisa Mallik[4], Nicklas Linz[4], Josef Priller[5,6,7,8], Eike Spruth[5,6], Slawek Altenstein[5,6], Jens Wiltfang[9,10,11], Inga Zerr[9,12], Claudia Bartels[10], Franziska Maier[13], Ayda Rostamzadeh[13], Emrah Duezel[14,15], Wenzel Glanz[14,15], Enise I Incesoy[14], Michaela Butryn[14,15], Christoph Laske[16,17], Sebastian Sodenkamp[16,18], Matthias HJ Munk[16,18], Bjoern Falkenburger[19,20], Antje Osterrath[19,20], Ingo Kilimann[1,2], Melina Stark[21,22], Luca Kleineidam[21,22], Michael T Heneka[23], Annika Spottke[21,24], Michael Wagner[21,22], Frank Jessen[13,21,25], Gabor C Petzold[21,26], Fedor Levin[2] and Stefan Teipel[1,2]

[1]Department of Psychosomatic Medicine, Rostock University Medical Center, Rostock, Germany
[2]German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany
[3]Cognition-Behavior-Technology group (Cobtek), Université Côte d'Azur, Nice, France
[4]ki:elements GmbH, Saarbrücken, Germany
[5]German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany
[6]Department of Psychiatry and Psychotherapy, Charité, Berlin, Germany
[7]University of Edinburgh and UK DRI, Edinburgh, UK
[8]Department of Psychiatry and Psychotherapy, School of Medicine and Health, Technical University of Munich, and German Center for Mental Health (DZPG), Munich, Germany
[9]German Center for Neurodegenerative Diseases (DZNE), Goettingen, Germany
[10]Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, University of Goettingen, Goetthingen, Germany
[11]Neurosciences and Signaling Group, Institute of Biomedicine (iBiMED), Department of Medical Sciences, University of Aveiro, Aveiro, Portugal
[12]Department of Neurology, University Medical Center, Georg August University, Goettingen, Germany
[13]Department of Psychiatry, University of Cologne, Medical Faculty, Cologne, Germany
[14]German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany
[15]Institute of Cognitive Neurology and Dementia Research (IKND), Otto-von-Guericke University, Magdeburg, Germany
[16]German Center for Neurodegenerative Diseases (DZNE), Tuebingen, Germany
[17]Section for Dementia Research, Hertie Institute for Clinical Brain Research and Department of Psychiatry and Psychotherapy, University of Tuebingen, Tuebingen, Germany
[18]Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany
[19]German Center for Neurodegenerative Diseases (DZNE), Dresden, Germany
[20]Department of Neurology, University Hospital Carl Gustav Carus, Technische Universitaet Dresden, Dresden, Germany
[21]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[22]Department of Cognitive Disorders and Old Age Psychiatry, University Hospital Bonn, Bonn, Germany
[23]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg
[24]Department of Neurology, University of Bonn, Bonn, Germany
[25]Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany
[26]Division of Vascular Neurology, Department of Neurology, University Hospital Bonn, Bonn, Germany

Corresponding author:
Qingyue Li, Department of Psychosomatic Medicine, University Rostock Medical Center, Gehlsheimer Straße 20, 18147 Rostock, Germany.
Email: Qingyue.Li@med.uni-rostock.de

Handling Associate Editor: Rebecca Amariglio

Abstract

Background: Speech-based features extracted from telephone-based cognitive tasks show promise for detecting cognitive decline in prodromal and manifest dementia. Little is known about the cerebral underpinnings of these speech features.

Objective: To examine associations between speech features, brain atrophy, and longitudinal cognitive decline in individuals at risk for Alzheimer's disease (AD).

Methods: Healthy volunteers, individuals with subjective cognitive decline, and those with mild cognitive impairment completed phonebot-guided semantic verbal fluency (SVF) and 15-word verbal learning task (VLT). Speech features were automatically extracted, and a global cognitive score (SB-C score) was computed. We analyzed data from 161 participants for cognitive trajectories, 141 for cross-sectional brain atrophy, and 102 for longitudinal brain changes. Analyses were conducted using multiple linear regressions, mixed-effects models, and voxel-based morphometry.

Results: The SB-C score was associated with bilateral hippocampal volumes, SVF features were primarily associated with left hemisphere regions, including the inferior frontal, parahippocampal, and superior/middle temporal gyri ($p_{\text{uncorr}} < 0.001$). SB-C score, SVF correct counts, and VLT delayed recall were associated with atrophy rates in the hippocampal/parahippocampal gyrus and left middle/inferior temporal gyri ($p_{\text{FDR}} < 0.05$). These features were also associated with cognitive decline assessed via Preclinical Alzheimer's Cognitive Composite 5, SVF, and Wordlist learning delayed recall ($p_{\text{FDR}} < 0.01$). Word frequency and temporal cluster switches showed varying associations with cognitive trajectories. Other features did not show robust associations.

Conclusions: In this study, we highlight the potential of digital speech features for identifying brain atrophy and cognitive decline over time in at-risk AD populations.

Keywords

Alzheimer's disease, atrophy, cognition, cognitive decline, early diagnosis, natural language processing, speech

Received: 7 February 2025; accepted: 21 May 2025

Introduction

Changes in speech-related cognitive functions are prevalent in Alzheimer's disease (AD) and often emerge in its earliest stages.[1,2] People with mild cognitive impairment (MCI) commonly exhibit distinctive speech patterns, such as more frequent pauses and reduced use of diverse or uncommon words.[2–4] Additionally, reduced cerebrospinal fluid amyloid-β (Aβ) levels in individuals with subjective cognitive decline (SCD) were associated with reduced use of specific words, including concrete nouns and content words.[5] These early speech alterations make speech analysis a viable tool for identifying the risk of AD and early detection of cognitive decline.[1,2]

Cognitive tasks, such as the Semantic Verbal Fluency (SVF) task and the Verbal Learning Task (VLT), are prime candidates for speech-based analysis in the early stages of AD.[6,7] These verbally administered tests assess a range of cognitive functions that are typically impaired in the AD spectrum compared with healthy elderly,[8–10] including executive processing speed, working memory, lexical access and semantic memory (SVF),[9] and episodic memory (VLT delayed recall).[10] Studies have also shown that baseline performance on SVF and VLT was associated with cognitive decline over time, even in preclinical stages.[11,12] For instance, one study reported that both preclinical individuals who later progressed to dementia and those already diagnosed with dementia experienced significant verbal fluency decline over time, with marked baseline impairment in SVF compared to cognitively healthy individuals.[11] Similarly, a study of individuals with SCD found that those who developed AD dementia after two years had significantly lower baseline VLT scores and more frequently performed poor on the delayed recall task.[12]

Beyond traditional task scores, research has suggested that more fine-grained speech features from these tasks might reflect subtle cognitive changes.[13,14] For example, semantic cluster size (the number of related words produced consecutively within a subcategory) and semantic cluster switches (transitions between different semantic subcategories) from SVF may reflect semantic memory retrieval and executive control, respectively.[13] With advancements in automatic speech recognition and natural language processing technologies, it has become possible to not only analyze these semantic features but also extract acoustic features, such as the timing and pacing of verbal output.[15–19] Evidence indicated that combining acoustic and semantic features with task scores was effective in achieving higher classification accuracy for machine learning classifiers trained for separating participants with SCD and MCI/dementia, compared to using task scores alone.[20,21]

While previous research has explored associations between digital speech features and gold-standard AD biomarkers, such as Aβ status and p-Tau levels,[22,23] critical gaps remain. For instance, one study employed an artificial intelligence text-pair evaluation model to analyze story recall tasks, demonstrating that speech-based assessments offer an accessible and scalable method for screening MCI and Aβ positivity.[23] However, limited work has investigated how digital speech features correlate with AD staging markers like brain atrophy or with longitudinal cognitive changes.

Brain atrophy, assessed through structural MRI, is a key staging marker of AD[24] and a promising candidate for investigating associations with digital speech features in early AD. Brain atrophy can be detected up to three years before diagnosis, even in the absence of overt cognitive symptoms.[25] AD-related atrophy often affects brain regions associated with cognitive functions essential to speech-based cognitive assessments, including memory, executive function, and semantic retrieval.[26–29] For example, decreases in gray matter (GM) in the hippocampus, parahippocampal gyrus, and basal forebrain are closely linked to different types of memory impairment commonly seen in the AD spectrum.[26,27] The anterior cingulate cortex (ACC) is crucial for executive function, and its atrophy in AD is well-documented.[28] Moreover, the middle and inferior temporal gyri, which are involved in semantic retrieval, may show atrophy and have been associated with future progression to AD dementia in non-demented individuals.[29]

In this study, we aimed to bridge these research gaps by examining the association between speech-based features derived from remote phone-based cognitive assessments and MRI-based brain measures, as well as longitudinal cognitive trajectories. Using data from the ongoing Prospect-AD study,[30] we analyzed digital speech features from SVF and VLT delayed recall to test three hypotheses: In pre-dementia AD stages, digital speech features are associated with (1) task-related regional brain volumes, (2) atrophy rate in these regions, and (3) domain-specific longitudinal cognitive trajectories. Establishing these associations could potentially enhance the clinical value of remote AD-risk assessment tools, aiding in the more efficient allocation of clinical resources.

Methods

Study design

Prospect-AD is a longitudinal, prospective observational study designed to evaluate the efficacy of digital speech features in early AD detection.[30] It integrates remote speech data collection into ongoing studies across Europe, including DESCRIBE and DELCODE in Germany, which began speech data collection in 2022. DELCODE[31] and DESCRIBE, initiated in 2014 and 2015 respectively, are longitudinal, multicenter studies conducted by the German Centre for Neurodegenerative Diseases (DZNE) throughout Germany. The study protocols included clinical assessments, neuropsychological tests, and MRI scans, while a subset of participants underwent cerebrospinal fluid (CSF) sampling and blood testing. Up-to-date cohort information is available at https://www.dzne.de/en/research/studies/clinical-studies/delcode/ and https://www.dzne.de/en/research/studies/clinical-studies/describe/.

All data collection adhered to ethical standards for medical research involving human participants, following the Declaration of Helsinki and the European General Data Protection Regulation. The German Prospect-AD study received approval from local institutional review boards: University Hospital Bonn (291/22), Ethics Committee of the Technical University of Dresden (BO-EK-263062022), University Medical Center Goettingen (27/9/22), University of Cologne Faculty of Medicine (22-1284), University Hospital Magdeburg Medical Faculty (80/22), University Medical Center Rostock (A2021-0256), and University Hospital Tuebingen (551/2022BO2). For Charité Universitätsmedizin Berlin, no separate approval was required, as it relied on the one from Rostock as the lead site.

Participants

To date, 234 participants aged 50 years or older, with cognitive states ranging from normal to MCI and a CDR global score ≤0.5, have been recruited for the Prospect-AD study. Key exclusion criteria included unstable medical conditions, psychiatric disorders, and substance abuse, with full inclusion and exclusion criteria outlined in the study protocol.[30] All participants provided written informed consent for speech data collection.

After exclusions (due to the lack of MRI scans, unavailable follow-up data, and other causes), 141 participants with at least one qualified MRI scan were included in the cross-sectional MRI dataset. Additionally, 102 participants with at least two MRI scans were included in the longitudinal MRI dataset. A separate longitudinal cognitive dataset included 161 participants with at least two clinical visits (Figure 1). We included the following participant groups: Healthy controls (HC), first-degree relatives of AD, SCD, and MCI.

Diagnosis and clinical assessment

Diagnoses of participants were made by local clinicians. HC were defined as individuals without cognitive complaints, showing no signs of cognitive impairment on cognitive assessments, and without diagnosed neurological or

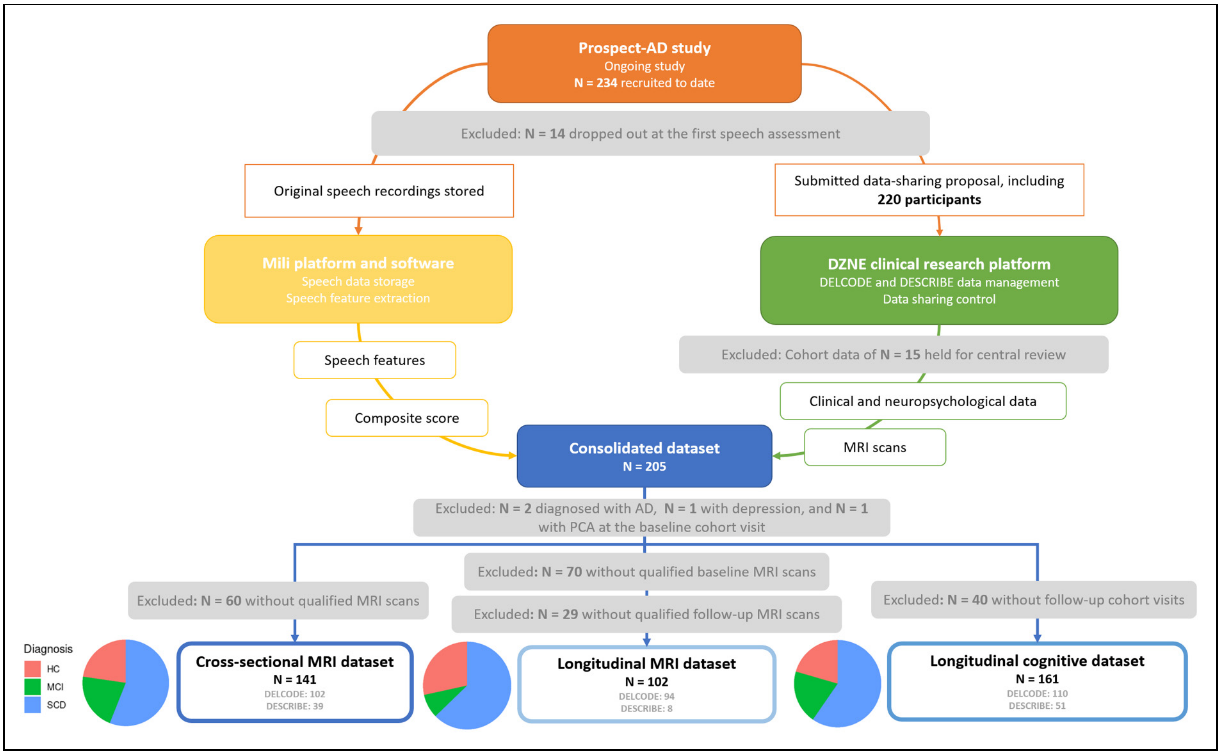
Prospect-AD study
Ongoing study
N = 234 recruited to date
Excluded: N = 14 dropped out at the first speech assessment
Original speech recordings stored
Submitted data-sharing proposal, including
220 participants
Mili platform and software
Speech data storage
Speech feature extraction
DZNE clinical research platform
DELCODE and DESCRIBE data management
Data sharing control
Speech features
Excluded: Cohort data of N = 15 held for central review
Clinical and neuropsychological data
Composite score
Consolidated dataset
N = 205
MRI scans
Excluded: N = 2 diagnosed with AD, N = 1 with depression, and N = 1 with PCA at the baseline cohort visit
Excluded: N = 70 without qualified baseline MRI scans
Excluded: N = 60 without qualified MRI scans
Excluded: N = 29 without qualified follow-up MRI scans
Excluded: N = 40 without follow-up cohort visits
Diagnosis
HC
MCI
SCD
Cross-sectional MRI dataset
N = 141
DELCODE: 102
DESCRIBE: 39
Longitudinal MRI dataset
N = 102
DELCODE: 94
DESCRIBE: 8
Longitudinal cognitive dataset
N = 161
DELCODE: 110
DESCRIBE: 51

Figure 1. Data overview of the ongoing prospect-AD study.

psychiatric conditions. First-degree relatives of AD were required to be cognitively normal but had at least one family member with a confirmed AD diagnosis. The criteria for SCD followed the consensus of the SCD Initiative Working Group, defining SCD as a self-reported, ongoing cognitive decline lasting at least six months without objective cognitive impairment and unrelated to sudden life events.[32] MCI (amnestic and non-amnestic) and AD dementia were diagnosed according to the National Institute on Aging-Alzheimer's Association (NIA-AA) guidelines.[33,34] After reviewing neuropsychological test results and available amyloid status, first-degree relatives of AD patients were grouped with cognitively healthy individuals to form the HC group in our dataset.

Each participant underwent a standardized assessment based on the criteria established for the DELCODE and DESCRIBE cohorts,[30,31] including psychiatric and neurological evaluations, clinical history, and questionnaires assessing disease severity and daily functioning. Both cohorts followed an annual follow-up schedule to collect clinical, neuropsychological, and MRI data. A comprehensive battery of neuropsychological tests was administered at each visit.[30] For our analysis, we used the Mini-Mental State Examination (MMSE)[35] and the Preclinical Alzheimer Cognitive Composite 5 (PACC5)[36] to assess general cognitive performance, the latter being available only in the DELCODE cohort. We also included paper-pencil-based SVF (animal category) and 10-word list learning (WLL) delayed recall to evaluate semantic and episodic memory, respectively.

Imaging acquisition and preprocessing

All structural MRI scans were acquired using Siemens 3T MRI scanners with harmonized study procedures and a standardized acquisition protocol across all sites.[31] The central DZNE imaging team performed quality checks on all scans to ensure compliance with study standards. Structural MRI data were preprocessed using Statistical Parametric Mapping (SPM12, revision 7487) and the CAT12 toolbox with DARTEL registration implemented in MATLAB (R2020a).[37] For cross-sectional analyses, we selected the most recent MRI scans for each participant that were closest in time to the Mili assessment. A total of 141 MRI scans were preprocessed using the CAT12 (revision 1202) general segmentation pipeline to obtain native-space GM files.[38–40] For longitudinal analyses, 553 longitudinal MRI scans were preprocessed using the CAT12.8 (revision 1872) longitudinal segmentation pipeline, optimized for atrophy detection to enhance measurement accuracy and reduce bias.[37,40] All segmented GM files underwent CAT12 sample homogeneity checks and outlier detection, with flagged images visually inspected for quality assurance before inclusion in the analysis.

We estimated regional GM volumes for 48 cortical and 4 subcortical regions per hemisphere using the

Harvard-Oxford atlas with a probability threshold of 0.5. An additional basal forebrain (BF) mask was applied to extract BF volumes.[41] Regional volumes were adjusted to individual baseline total intracranial volume (TIV) to correct for head size differences.[40,42]

Individual annual GM atrophy rate was estimated using voxel-wise difference maps with the *ImCalc* toolbox in SPM12, using the formula: $\Delta GM_{volume} = (GM_{baseline} – GM_{latest}/(t_{latest} – t_{baseline})$.[43] $\mathrm{GM}_{\mathrm{baseline}}$ and $\mathrm{GM}_{\mathrm{latest}}$ represent the first and last available follow-up scan per participant, and (t_{latest}–t_{baseline}) is the time interval between scans. For voxel-based morphometry (VBM) analyses, baseline GM images and difference maps were smoothed with an 8 mm Gaussian kernel.

Mili assessment and speech data processing

Speech data were collected through six automated phone-based assessments over 15 months, with calls scheduled every 3 months. The phonebot 'Mili' automatically called and guided participants through verbal-based tasks, including immediate verbal learning encoding (VLT, 15 words), semantic verbal fluency (SVF, animal category), narrative storytelling, and verbal learning delayed recall, respectively.[30] Our analyses focused on speech features from the first Mili assessment (hereafter referred to as 'Mili').

Speech recordings from both SVF and VLT underwent automated processing using the proprietary SIGMA speech analysis pipeline developed by ki:elements.[44,45] The SIGMA pipeline includes automatic speech recognition, which is optimized for German and integrated with a local language corpus to transcribe verbal responses and extract semantic and task-related features, such as word frequency and correct counts. High agreement was observed between automated and manual SVF transcriptions in German, particularly for word count, which also showed strong discrimination between cognitively impaired and unimpaired individuals.[44] In parallel, voice analysis of the original audio is used to derive acoustic features, including measures like temporal cluster switches, based on intra-word pause durations. A composite cognitive score (SB-C score) and sub-domain scores were automatically computed from these speech features.[45] The workflow of the speech data processing was described in the previous report,[45] and a full list of features derived from SVF and VLT delayed recall is detailed in Supplemental Table 1.

Selection of variables of interest

Regions of interest for volumetry. Relevant regional volumes were combined into larger regions of interest (ROIs) based on their functional roles in cognitive tasks. This included regions involved in language processing,[46,47] executive speed and working memory, BF (memory and attention),[27,48] and hippocampus and parahippocampus (episodic memory). Regions associated with semantic memory retrieval[49] were selected for the SVF features analyses, while regions involved in episodic memory were chosen for VLT delayed recall features analyses (Table 1).

Speech features. We selected Global cognition and task correct response metrics, including SB-C cognition score, SVF correct count, and VLT delayed recall count.[20,21,45] Fine-grained features were chosen based on prior evidence and their ability to detect subtle underlying cognitive processing. From SVF, we included semantic cluster size and switch counts due to their established associations with semantic memory and executive function.[13] Word frequency was added given its relevance to early word-finding difficulties.[4] For the VLT task, serial cluster size was selected to capture sequential recall patterns linked to episodic memory organization.[50] To complement semantic features, we also incorporated temporal cluster size and switch counts, as prior work suggests that combining temporal and semantic information helps investigate exploitation and exploration patterns, which serve as indicators for semantic memory retrieval and executive control processes.[18] Detailed feature descriptions and their cognitive associations are provided in Table 2.

Statistical analysis

Linear regression and linear mixed-effect models were conducted in R (v4.1.3) accessed via RStudio, using the *lm* base function and the *lmerTest* package, respectively. The false discovery rate (FDR) was controlled using the Benjamini-Hochberg procedure to correct for multiple comparisons ($p_{\mathrm{FDR}} < 0.05$). VBM analyses were performed using SPM12 (revision 7487) in MATLAB (R2020a), with FDR correction applied. For all analyses, sex, diagnosis, and scan sites were treated as dummy variables, with female, HC, and one site as the reference categories.

Hypothesis-driven analyses. We used linear regression models to test associations between ROI GM volumes and digital speech features in the cross-sectional MRI dataset. The models controlled for sex, age, education years, diagnosis at the scan, scan site, and the time gap between the MRI scan and Mili assessment. Regression model assumptions were assessed using standard diagnostic procedures. Box-Cox transformations were applied to the dependent variables to improve model fit when violations were detected. Linear mixed effects models were used to examine the associations between digital speech features and brain atrophy/cognitive decline over time in the longitudinal datasets. Fixed effects included speech features, follow-up years, sex, age, education years, diagnosis, the time gap in years between the baseline scan/cognitive

Table 1. Regions of interest (ROIs) analyzed for different digital speech features.

			Relevant cognitive measure			
Function	Region	Hemisphere	SVF features	VLT delayed recall features	Task correct replies	Global cognitive score
Language processing	Inferior frontal gyrus (Broca's area)	Left	√	√	√	√
	Posterior superior temporal gyrus (Wernicke's area)	Left	√	√	√	√
	Inferior parietal lobule	Left	√	√	√	√
Executive function	Dorsolateral prefrontal cortex (DLPFC)	Bilateral	√	√	√	√
	Anterior cingulate cortex (ACC)	Bilateral	√	√	√	√
Memory	Basal forebrain (BF)	Bilateral	√	√	√	√
Semantic memory	Middle temporal gyrus (MTG)	Left	√		√	√
	Inferior temporal gyrus (ITG)	Left	√		√	√
Episodic memory	Hippocampus	Bilateral		√	√	√
	Parahippocampal gyrus	Bilateral		√	√	√

tests and Mili (time gap), the interaction between diagnosis and follow-up years, and the interaction between speech features and follow-up years. For longitudinal MRI analysis, we also controlled for scan sites. Models included random intercepts and slopes[51] over time nested within participants, with the default unstructured covariance matrix. Model assumptions were visually inspected and confirmed to be adequately met. As a sensitivity analysis, we repeated all longitudinal models, excluding MCI participants, to assess potential bias related to follow-up time differences across diagnostic groups.

Data visualizations and model summaries were generated using the '*ggplot2*' and '*sjPlot*' packages.

Data-driven analyses. To further investigate the association between digital speech features and local atrophy, we performed cross-sectional and longitudinal VBM analyses on modulated, normalized, and smoothed baseline GM images and difference maps. We conducted multiple linear regression models controlled for sex, age, education years, diagnosis, TIV, and time gap. One-sample t-tests were applied to assess positive associations between GM volumes and digital features, with a negative association expected for SVF 'Word frequency' and 'Semantic cluster switches'. Results were thresholded at $p_{\mathrm{FDR}} < 0.05$ and visualized using the *bspmView* toolbox, with an extent threshold k set at > 50 voxels. A liberal threshold (uncorrected $p < 0.001$) was also considered when no significant clusters emerged after multiple comparison correction, as recommended to balance Type I and Type II error control.[52]

Results

Descriptive statistics

In the cross-sectional dataset, participants' ages at the scans ranged from 55 to 88 years (mean age = 72.6 ± 6.4 years). The median time gap between the latest scan and the phone-based assessment with Mili was 1.5 years (range: −1.0–9.0 years), with no significant difference across diagnostic groups (Table 3). Baseline TIV-adjusted brain volumes and digital speech feature performance are shown in Supplemental Tables 2 and 3. Participants in the longitudinal MRI dataset had a median of 5 follow-up scans (range: 1–8), with significant differences observed between diagnostic groups (MCI < HC, $p = 0.015$; SCD < HC, $p = 0.008$). Participants in the longitudinal cognitive dataset had a median follow-up duration of 5 years (range: 1–8 years), with significant differences observed between diagnostic groups (MCI < SCD < HC, $p < 0.01$). The MCI group also had fewer females compared to the other groups. Demographic information and values of digital speech features for both the longitudinal MRI and cognitive datasets are provided in Supplemental Tables 4–7.

Cross-sectional analyses

ROI-based analyses. Adjusted GM volumes of the ACC were log-transformed ($\lambda \approx 0$), while hippocampal and parahippocampal volumes were squared ($\lambda \approx 2$). After transformation, all models met the assumptions of linear regression. We did not find any significant associations between digital speech features and ROI volumes at the $p_{\mathrm{FDR}} < 0.05$ level. The complete results with uncorrected p-values can be found in Supplemental Table 8.

Voxel-based morphometry analyses. No clusters survived multiple comparison correction (all $p_{\mathrm{FDR}} > 0.05$). However, associations were identified at an uncorrected voxel-wise threshold of $p < 0.001$. GM volume was associated with several digital speech features, including SB-C cognition score, SVF correct count, and SVF word frequency. In regions such as the bilateral hippocampus,

Table 2. Speech features for analyses and their descriptions.

Category	Features	Description	Functional interpretation
global cognition	SB-C Cognition score[45]	A composite score derived from over 50 automatically extracted speech features from SVF and VLT.	Serves as a global cognitive performance marker, reflecting learning and memory, executive function, and processing speed.
Task score	SVF correct count	The total number of animal names correctly produced in one minute.	Assesses lexical retrieval, semantic memory access, and executive function.
	VLT delayed recall	Number of correctly remembered words in the delayed recall trial.	Assesses encoding efficiency and episodic memory.
SVF features	Word frequency	Average frequency of produced animal names based on a German corpus.	Reflects lexical accessibility; lower frequency may suggest richer lexical retrieval.
	Semantic cluster size	Average number of semantically related words produced consecutively in a subcategory (e.g., farm animals).	Reflects semantic memory and lexical retrieval processes.
	Semantic cluster switches	The total number of switches between semantic subcategories (n_clusters - 1).	Reflects executive functions (flexibility, strategy and attention control).
	Temporal cluster size	Average size of word groups produced in rapid succession (based on pauses).	Reflects automatic lexical access and semantic memory strength, exploitation.
	Temporal cluster switches	Number of transitions between temporal clusters (n_clusters - 1).	Reflects executive functions (cognitive flexibility, strategic search, inhibition), exploration.
VLT delayed recall features	Serial clusters	Average size of word sequences recalled in original learning order.	Reflects passive manner of encoding strategy.
	Temporal cluster size	Average size of word groups produced in rapid succession (based on pauses).	Reflects episode memory retrieval process and exploitation strategy.
	Temporal cluster switches	Number of transitions between temporal clusters (n_clusters - 1).	Reflects retrieval search strategies (more switches = less organized retrieval).

inferior frontal gyrus, superior and middle temporal gyri, fusiform gyrus, and insula (Figure 2). In contrast, SVF semantic cluster size, SVF temporal cluster size, and VLT delayed recall temporal cluster switches did not exhibit significant associations with any brain regions (Supplemental Table 9).

Longitudinal analyses

ROI-based analyses. Focusing on the interaction between digital speech features and follow-up years, we found that higher SB-C scores were significantly associated with slower atrophy of the bilateral hippocampus, parahippocampal gyrus, ACC, and left MTG and IFG (Figure 3; Supplemental Table 10). Higher SVF correct counts were significantly associated with slower atrophy rate in the bilateral hippocampus (B = 0.001, [95% CI 0.0005, 0.002], $p_{\text{FDR}} = 0.01$), parahippocampal gyrus (0.002, [0.0006, 0.0025], $p_{\text{FDR}} = 0.01$), ACC (0.002, [0.001, 0.003], $p_{\text{FDR}} = 0.003$), left MTG (0.001, [0.0003, 0.002], $p_{\text{FDR}} = 0.02$), left ITG (0.001, [0.0004, 0.002], $p_{\text{FDR}} = 0.01$), and left posterior superior temporal gyrus (0.0002, [0.00002, 0.0003], $p_{\text{FDR}} = 0.05$). Higher VLT delayed recall correct counts were associated with a slower atrophy in the bilateral hippocampus (0.003, [0.001, 0.005], $p_{\text{FDR}} = 0.01$), parahippocampal gyrus (0.002, [0.001, 0.004], $p_{\text{FDR}} = 0.03$), left MTG (0.002, [0.001, 0.004], $p_{\text{FDR}} = 0.02$), and left ITG (0.002, [0.001, 0.003], $p_{\text{FDR}} = 0.02$). In addition, SVF word frequency was associated with atrophy rate in the bilateral ACC (−0.043, [−0.071, −0.016], $p_{\text{FDR}} = 0.03$). No speech markers were associated with longitudinal atrophy in the left inferior frontal gyrus and inferior parietal lobule, DLPFC, or BF. The interaction effects of speech features and follow-up time on regional brain volumes are visualized in Supplemental Figure 1 and detailed in Supplemental Table 10. The sensitivity analysis excluding MCI participants remained largely consistent with the primary analyses, with no substantial changes in the significance or direction of the associations.

Voxel-based morphometry analyses. In the complementary VBM regression analyses, no associations survived at the $p_{\text{FDR}} < 0.05$ level. When applying an uncorrected threshold of $p < 0.001$, we observed significant associations between SVF features and GM volume changes over time in voxels located in areas such as the fusiform gyrus, inferior frontal gyrus, precuneus, and ACC. VLT delayed recall features were significantly associated with changes in volumes in the right precuneus, inferior frontal gyrus, and middle temporal gyrus. Further associations were observed with SVF semantic cluster switches, SVF temporal cluster

Table 3. Demographic information of participants in the cross-sectional MRI dataset.

Mean (SD)	Overall (N = 141)	HC (N = 32)	SCD (N = 79)	MCI (N = 30)	ANOVA p
Age at scan	72.6 (6.4)	73.7 (6.0)	71.6 (6.5)	74.1 (6.3)	0.113
Median [Min, Max]	71.9 [55.4, 87.8]	71.7 [60.8, 85.2]	71.9 [55.4, 87.8]	76.2 [58.6, 84.6]	
Sex (f, %)	78 (55.3%)	24 (75.0%)	40 (50.6%)	14 (46.7%)	$\chi^2(2) = 6.6$, $p = .036$*; post hoc n.s.
Education (y)	14.9 (2.8)	14.8 (2.6)	15.2 (2.8)	14.6 (2.9)	0.547
Time gap (y)	1.5 (1.8)	2.0 (2.4)	1.5 (1.7)	1.2 (1.3)	0.238
Median [Min, Max]	1.0 [−1.0, 9.0]	1.0 [0.0, 9.0]	1.1 [−0.1, 8.0]	1.0 [−1.0, 4.9]	

*$p < 0.05$ **$p < 0.01$ ***$p < 0.001$.

switches, SVF temporal cluster size, and VLT delayed recall temporal cluster switches in regions including the left MTG, ACC, and parahippocampal gyrus (Supplemental Table 11).

Analysis of cognitive decline

Higher SB-C cognition scores ($p_{\text{FDR}} < 0.01$), SVF correct counts ($p_{\text{FDR}} < 0.01$), and VLT delayed recall performance ($p_{\text{FDR}} < 0.001$) were significantly associated with slower cognitive decline on paper-pencil-based PACC5, SVF, and WLL delayed recall. (Figure 4; Supplemental Table 12).

For the automated SVF task, participants who used less frequent words experienced slower cognitive decline, as measured by paper-pencil-based PACC5 score (−0.09, [−0.16, −0.02], $p_{\text{FDR}} = 0.03$) and WLL delayed recall (−0.29, [−0.46, −0.13], $p_{\text{FDR}} = 0.003$). More frequent temporal cluster switches were associated with slower cognitive decline, significantly shown on traditional WLL delayed recall (0.03, [0.01, 0.05], $p_{\text{FDR}} = 0.05$). In contrast, features 'Semantic cluster size', 'Temporal cluster size', and 'Semantic cluster switches' did not show significant interaction effects on any of the cognitive assessment scores. For the automated VLT delayed recall, participants who exhibited more frequent temporal switches experienced slower cognitive decline as measured by MMSE (0.03, [0.002, 0.05], $p_{\text{FDR}} = 0.04$), PACC5 (0.02, [0.01, 0.03], $p_{\text{FDR}} = 0.002$), and WLL delayed recall (0.04, [0.01, 0.07], $p_{\text{FDR}} = 0.01$). We did not find any significant associations between paper-pencil-based cognitive trajectories and the features 'Serial cluster size' and 'Temporal cluster size'. All interaction effects of speech features and follow-up time on cognitive scores are visualized in Supplemental Figure 2 and detailed in Supplemental Table 12. The sensitivity analysis excluding MCI participants remained largely consistent with the primary analyses, with no substantial changes in the significance or direction of the associations.

Discussion

In this study, we explored whether digital speech features from remote SVF and VLT delayed recall could reflect brain atrophy and cognitive decline in the very early stages of AD. Our findings show that the automatically calculated SB-C cognition score, along with correct count measures from both tasks (SVF and VLT delayed recall), showed trends of association with cross-sectional brain volumes in regions related to early AD brain pathology. These features were also associated with regional brain atrophy and cognitive trajectories over time. Notably, temporal cluster switches and word frequency emerged as secondary markers of brain regional atrophy and cognitive trajectories. Semantic features showed limited associations.

Together, these results highlight the potential of automated speech-based features, especially the composite score and word counts, as early indicators of brain structural and cognitive changes.

Association between digital speech features and volumes of relevant brain regions

We first hypothesized that digital speech features would be associated with MRI-based volumes of relevant brain regions. VBM analyses partially supported this. The SB-C score showed trend-level associations with bilateral hippocampal volumes in both uncorrected ROI and VBM analyses. This finding aligns with prior research reporting associations between composite cognitive scores and hippocampal volume, although such studies have typically used global scores derived from traditional paper-and-pencil assessments.[53,54] As the SB-C score is an automatically calculated digital measure, it highlights its potential utility for scalable and remote cognitive assessment. Additionally, speech features showed associations with the fusiform gyrus and temporal gyri in exploratory analyses. These regions are known to undergo atrophy early in the AD continuum, as supported by a previous study.[25]

While these preliminary findings point to the potential of automated digital assessments in reflecting early structural brain changes, they should be interpreted with caution. Notably, effects were observed in uncorrected VBM but not in ROI analyses and did not survive multiple comparison correction. The distinction likely reflects both methodological and statistical considerations. First, the voxel-based approach may offer greater sensitivity to subtle or localized

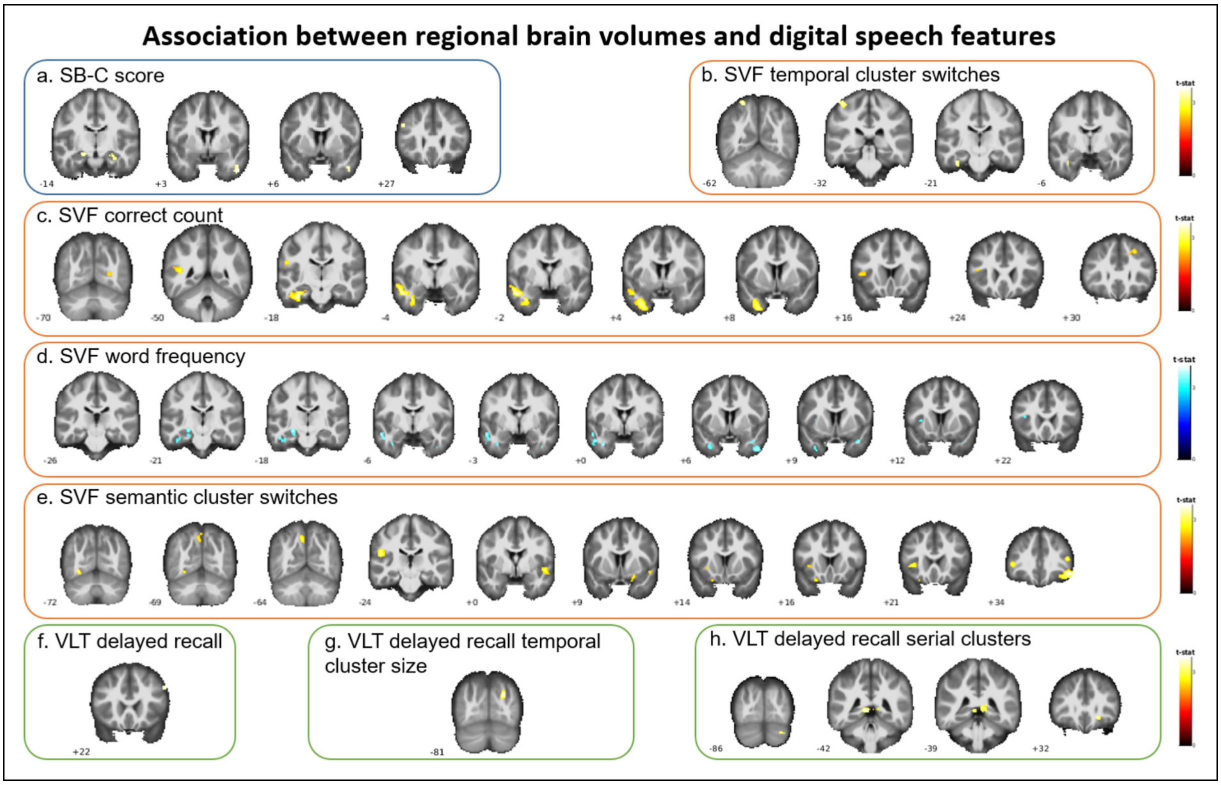
Association between regional brain volumes and digital speech features
a. SB-C score
-14
+3
+6
+27
b. SVF temporal cluster switches
-62
-32
-21
-6
c. SVF correct count
-70
-50
-18
-4
-2
+4
+8
+16
+24
+30
d. SVF word frequency
-26
-21
-18
-6
-3
+0
+6
+9
+12
+22
e. SVF semantic cluster switches
-72
-69
-64
-24
+0
+9
+14
+16
+21
+34
f. VLT delayed recall
+22
g. VLT delayed recall temporal cluster size
-81
h. VLT delayed recall serial clusters
-86
-42
-39
+32
t-stat

Figure 2. VBM-based associations between digital speech features and regional GM volumes.
Images are displayed in coronal view (neurological convention: left = left hemisphere). Statistically significant voxels (uncorrected $p < 0.001$) are overlaid on the MNI152 brain template. All models controlled for age, sex, education, TIV, scan site, time gap between MRI and speech assessment, and diagnosis.

changes, which could be missed in the more predefined ROI approach.[55] Especially when identifying very early-stage changes that might be spatially restricted and localized to some subregions within larger brain areas. Additionally, conservative thresholds in VBM analysis, while minimizing Type I errors, may limit the detection of subtle effects, increasing the likelihood of Type II errors.[52] In cross-sectional studies, a liberal threshold might still detect meaningful effects when supported by replication and meta-analysis evidence.[52] Therefore, with further validation across cohorts, digital speech assessment may become a useful component of early-AD identification strategies.

Association between digital speech features and atrophy rate of relevant brain regions

Our results provide partial evidence for associations between digital speech features and brain atrophy rates. Higher composite scores and word counts were significantly associated with slower atrophy rates across multiple regions involved in different cognitive processes,[26,28,49] including the hippocampus and parahippocampal gyrus. These findings complement our cross-sectional results and align with prior work highlighting hippocampal atrophy as an AD staging marker for both episodic memory and lexical-semantic processing.[56]

Among fine-grained features, only SVF word frequency was associated with ACC atrophy rate, though the association was modest. Smaller associations were also observed for the SB-C score and SVF correct counts with ACC atrophy, which may reflect the difficulty of using digital speech features to capture subtle changes in the ACC. This may be because ACC atrophy typically appears later in AD progression and shows greater loss in individuals with MCI who eventually progress to AD.[25,57] Other prefrontal regions may follow a similar trajectory, which could explain the lack of significant associations with additional speech features at these very early disease stages.[25] Given that our participants were in the very early stages of AD, with only 21% diagnosed with MCI, these later-stage atrophy patterns may not have been captured. Furthermore, data-driven analyses did not show any significant results after FDR correction, likely due to limitations in estimating atrophy rates using only two scans per participant, reducing the statistical robustness provided by mixed-effects models, where up to nine scans per participant were included.

These results suggest that overall performance on digital assessments may be more reflective of atrophy rates in

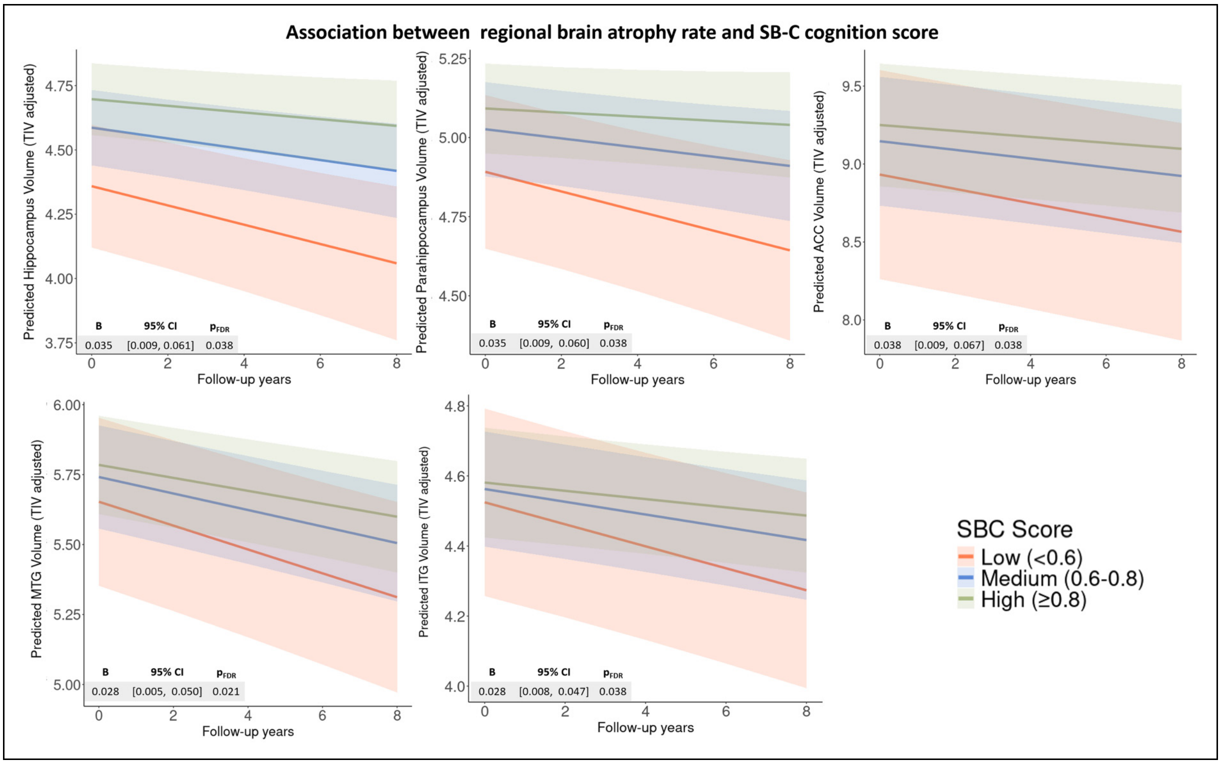

Figure 3. ROI-based associations between regional brain atrophy rate and SB-C cognition score.
Predicted regional brain volumes (y-axis) were derived from linear mixed-effects models adjusted for age, sex, education, diagnosis, scan site, and the time gap between baseline MRI and speech assessment. Models included random intercepts and slopes to account for repeated measurements. Regression lines represent model-predicted atrophy trajectories over time, with shaded ribbons indicating 95% confidence intervals. For visualization purposes, SB-C scores were divided into tertiles (low, middle, high) based on their distribution, with the 33rd and 67th percentiles as the cutoff points. The regression lines illustrate atrophy trajectories for each tertile group.

early-affected medial temporal regions than in later-vulnerable frontal areas, despite the latter's key role in language processing. As a longitudinal speech dataset is developed in our Prospect-AD study, future studies can refine these associations and assess their utility for monitoring atrophy progression through parallel longitudinal analyses.

Association between digital speech features and cognitive trajectories

We thirdly hypothesized that digital speech features would be associated with cognitive trajectories, which was partially supported. The SB-C cognition score, SVF correct count, and VLT delayed recall demonstrated significant associations with cognitive trajectories in paper-and-pencil-based PACC5, SVF, and WLL delayed recall outcomes. PACC5, as a widely used cognitive composite score for detecting early cognitive decline,[36] showed a significant relationship with the SB-C score, suggesting that the digital composite score may offer comparable sensitivity to traditional neuropsychological measures. These findings also extend prior work[11,12,20,21] by demonstrating that automated digital scores not only align with paper-and-pencil-based measures at a single time point but also track cognitive trajectories over time, highlighting the potential of digital assessments as a remote approach for detecting cognitive decline.

In contrast, only one significant association was found between the MMSE trajectory and VLT delayed recall temporal cluster switches. This contrasts with previous cross-sectional findings from a Dutch cohort that reported a significant correlation between the SB-C cognition score and MMSE score ($r = 0.54$, $p < 0.001$, $d = 1.28$).[45] This discrepancy may be due to differences in sample composition. Notably, the Dutch cohort included a higher proportion of participants with MCI and dementia (49% of the sample), whereas we had only 20% with MCI. This contrast may also underscore the limitations of the MMSE as a stand-alone tool for detecting early-stage cognitive decline.[35] Ceiling effects[58] and limited score variability in our SCD-predominant sample may have contributed, as reflected in the narrow distribution and relatively stable individual trajectories observed over time.

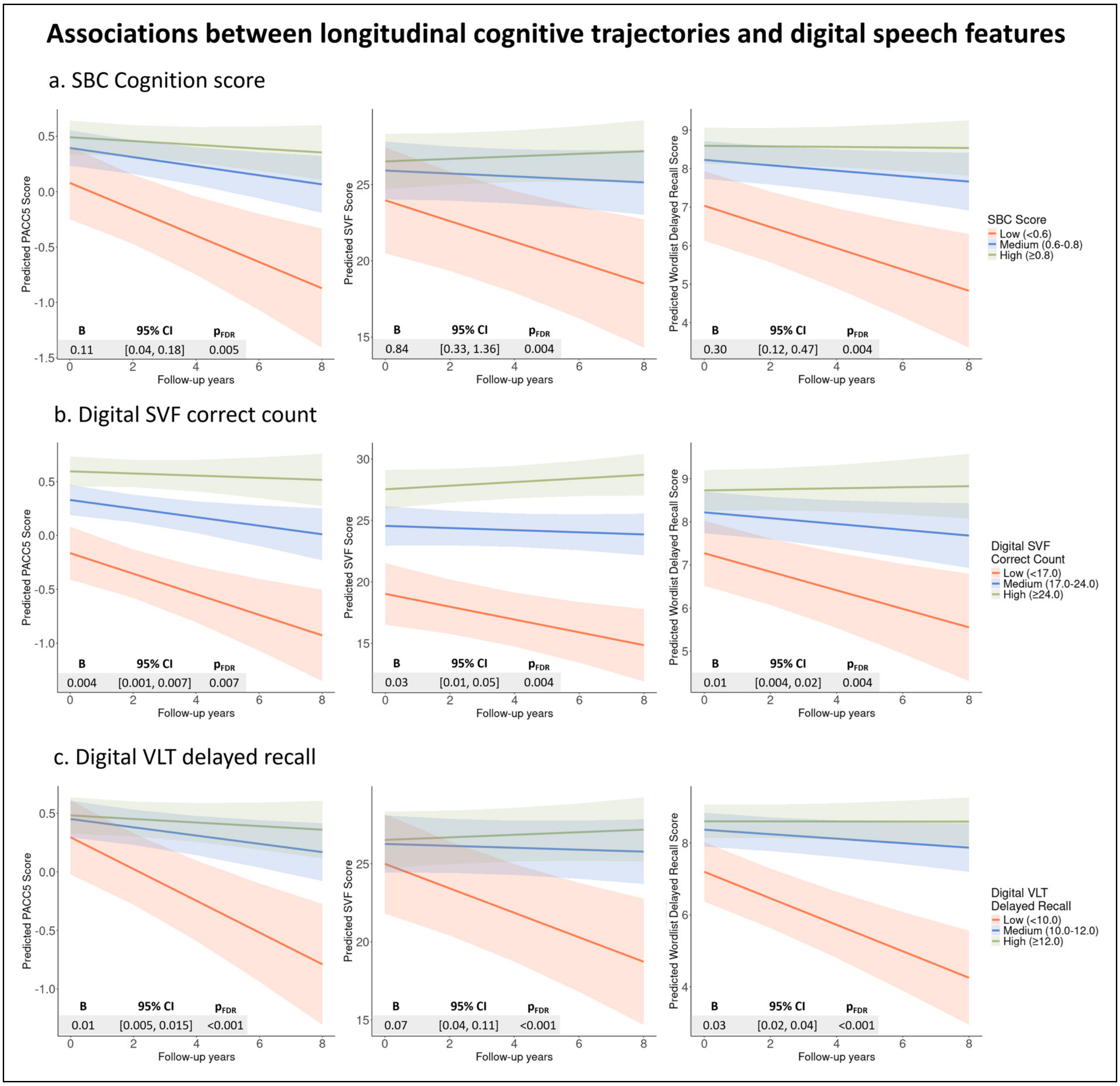

Figure 4. Associations between longitudinal cognitive trajectories via traditional neurocognitive assessments and digital speech features.

Predicted cognitive scores (y-axis) are derived from linear mixed-effects models adjusted for age, sex, education years, diagnosis, and the time gap between baseline neurocognitive testing and speech assessment. Models included random intercepts and slopes to account for repeated measurements. Regression lines represent model-predicted cognitive trajectories, with shaded ribbons indicating 95% confidence intervals. For visualization purposes, the values of digital speech features were divided into tertiles (low, middle, high) based on their distribution, with the 33rd and 67th percentiles as the cutoff points. The regression lines illustrate atrophy trajectories for each tertile group.

Among the fine-grained features, only word frequency and temporal cluster switches were associated with cognitive changes. Word frequency may reflect increased reliance on high-frequency, familiar words as semantic memory begins to decline, a pattern seen in early AD.[4] This feature may therefore act as a subtle marker of compensatory strategies in cognitive processing, with previous studies suggesting that word frequency reflects episodic recognition rather than just semantic-lexical processing.[59] Temporal features, which may reflect executive control and processing speed, have shown diagnostic value and correlations with cognitive profiles in SVF tasks.[18,19] However, research on these pause-oriented features in the VLT remains limited, replication in future studies is essential before concluding their reliability and utility in this context.

From a practical standpoint, automated speech-based measures offer key advantages over traditional testing. While individual digital scores like SVF and VLT delayed recall correct counts may still be affected by practice effects, the SB-C score appears more robust to such biases. As an integrative measure derived from multiple automatically extracted features, SB-C captures cognitive performance with greater nuance. Its automated, multidimensional design allows for a more refined and potentially more stable representation of cognitive change over time compared to traditional assessments.

Strengths and limitations

A key strength of our study is its use of fully automated and remote assessments to explore the associations between speech patterns and brain atrophy. We employed a comprehensive approach that included both ROI-based and voxel-wise analyses of GM volume conducted through cross-sectional and longitudinal methods. Additionally, our focus on longitudinal trajectories of cognitive scores from established neuropsychological tests provides insights into the discovered relationship between cognitive decline and speech features in individuals at pre-dementia stages, addressing a gap in prior studies, which have largely relied on cross-sectional data.

A key limitation of this study relates to its procedural design. As an add-on to existing cohorts, MRI scans and speech assessments were not conducted within a short time frame, which may have biased observed associations, potentially obscuring links between semantic features and regions involved in semantic processing.[60] In an upcoming cohort, we will have the opportunity to align speech collection more closely with clinical visits. Furthermore, the cohorts had different recruitment timelines and focuses, contributing to an imbalance in diagnostic group sizes and follow-up durations. Although we conducted longitudinal sensitivity analyses excluding MCI participants, potential bias from varying follow-up periods or missing data remains. Additionally, while Aβ status is known to influence cognitive performance,[61,62] limited availability in our datasets (e.g., n = 2 in the cross-sectional dataset) prevented us from examining it as a mediator. Lastly, since current digital speech features were captured from standard cognitive tasks based on word-response, they may not fully capture more natural speech patterns or identify richer indicators.[5,63] Future work will explore spontaneous speech tasks to identify more sensitive indicators of early AD-related changes.

Conclusion

Speech features from remote SVF and VLT delayed recall, along with the digital global cognition score, showed potential for detecting early signs of brain atrophy and longitudinal cognitive changes across multiple domains. Our findings motivate further research on digital speech features related to cognitive ability and brain atrophy in the at-risk stage of AD. Continued development and validation of these features could enhance their utility in early AD detection and contribute to scalable and cost-effective identification tools for at-risk populations.

Acknowledgments

We are grateful to the European Union's HORIZON-MSCA Doctoral Networks for supporting young researchers and to the Alzheimer's Drug Discovery Foundation for funding this study. We sincerely thank ki:elements for their contribution to AI-based software development and speech data collection. Special thanks to DZNE, all Prospect-AD sites, and the DELCODE and DESCRIBE research teams for their efforts in clinical participant recruitment, clinical assessments, neurocognitive and MRI data collection, and quality control. We appreciate the dedication and support of our study teams at Rostock University Medical Center and the DZNE Rostock site. We also acknowledge the valuable contributions of our co-authors to study design, project coordination, data curation, and manuscript preparation. Lastly, we extend our deepest gratitude to all Prospect-AD study participants for their invaluable participation.

ORCID iDs

Qingyue Li https://orcid.org/0009-0001-2559-4795
Stefanie Koehler https://orcid.org/0000-0002-7417-333X
Alexandra Koenig https://orcid.org/0000-0001-9960-9657
Martin Dyrba https://orcid.org/0000-0002-3353-3167
Nicklas Linz https://orcid.org/0000-0001-5178-3234
Josef Priller https://orcid.org/0000-0001-7596-0979
Eike Spruth https://orcid.org/0000-0002-8976-7309
Slawek Altenstein https://orcid.org/0000-0003-2753-5999
Jens Wiltfang https://orcid.org/0000-0003-1492-5330
Inga Zerr https://orcid.org/0000-0002-6722-2463
Claudia Bartels https://orcid.org/0000-0003-3023-9971
Franziska Maier https://orcid.org/0000-0002-9335-9594
Ayda Rostamzadeh https://orcid.org/0000-0001-5189-134X
Emrah Duezel https://orcid.org/0000-0002-0139-5388
Wenzel Glanz https://orcid.org/0000-0002-5865-4176
Enise I Incesoy https://orcid.org/0000-0003-2014-4098
Sebastian Sodenkamp https://orcid.org/0009-0004-9118-0621
Matthias HJ Munk https://orcid.org/0000-0002-5339-4045
Bjoern Falkenburger https://orcid.org/0000-0002-2387-526X
Antje Osterrath https://orcid.org/0009-0003-8460-760X
Ingo Kilimann https://orcid.org/0000-0002-3269-4452
Melina Stark https://orcid.org/0009-0005-3812-023X
Luca Kleineidam https://orcid.org/0009-0006-3309-6856
Michael T Heneka https://orcid.org/0000-0003-4996-1630
Annika Spottke https://orcid.org/0000-0001-9854-2972
Michael Wagner https://orcid.org/0000-0003-2589-6440
Frank Jessen https://orcid.org/0000-0003-1067-2102
Gabor C Petzold https://orcid.org/0000-0002-0145-8641

Fedor Levin https://orcid.org/0000-0002-0518-1715
Stefan Teipel https://orcid.org/0000-0002-3586-3194

Ethical considerations

All data collection adhered to ethical standards for medical research involving human participants, following the Declaration of Helsinki and the European General Data Protection Regulation. The German Prospect-AD study received approval from local institutional review boards: University Hospital Bonn (291/22), Ethics Committee of the Technical University of Dresden (BO-EK-263062022), University Medical Center Goettingen (27/9/22), University of Cologne Faculty of Medicine (22-1284), University Hospital Magdeburg Medical Faculty (80/22), University Medical Center Rostock (A2021-0256), and University Hospital Tuebingen (551/2022BO2). For Charité Universitätsmedizin Berlin, no separate approval was required, as it relied on the one from Rostock as the lead site.

Consent to participate

All participants provided written informed consent for speech data collection.

Author contributions

Qingyue Li: Conceptualization; Formal analysis; Methodology; Visualization; Writing – original draft.
Stefanie Koehler: Data curation; Investigation; Project administration; Writing – review & editing.
Alexandra Koenig: Conceptualization; Writing – review & editing.
Martin Dyrba: Methodology; Writing – review & editing.
Elisa Mallick: Methodology; Software; Writing – review & editing.
Nicklas Linz: Methodology; Software; Writing – review & editing.
Josef Priller: Data curation; Investigation; Writing – review & editing.
Eike Spruth: Data curation; Investigation; Writing – review & editing.
Slawek Altenstein: Data curation; Investigation; Writing – review & editing.
Jens Wiltfang: Data curation; Investigation; Writing – review & editing.
Inga Zerr: Data curation; Investigation; Writing – review & editing.
Claudia Bartels: Data curation; Investigation; Writing – review & editing.
Franziska Maier: Data curation; Investigation; Writing – review & editing.
Ayda Rostamzadeh: Data curation; Investigation; Writing – review & editing.
Emrah Düzel: Data curation; Investigation; Writing – review & editing.
Wenzel Glanz: Data curation; Investigation; Writing – review & editing.
Enise Inceroy: Data curation; Investigation; Writing – review & editing.
Michaela Butryn: Data curation; Investigation; Writing – review & editing.
Christoph Laske: Data curation; Investigation; Writing – review & editing.
Sebastian Sodenkamp: Data curation; Investigation; Writing – review & editing.
Matthias Munk: Data curation; Investigation; Writing – review & editing.
Bjoern Falkenburger: Data curation; Investigation; Writing – review & editing.
Antje Osterrath: Data curation; Investigation; Writing – review & editing.
Ingo Kilimann: Data curation; Investigation; Writing – review & editing.
Melina Stark: Data curation; Investigation; Writing – review & editing.
Luca Kleineidam: Data curation; Investigation; Writing – review & editing.
Michael Heneka: Data curation; Investigation; Writing – review & editing.
Annika Spottke: Data curation; Investigation; Writing – review & editing.
Michael Wagner: Data curation; Investigation; Writing – review & editing.
Frank Jessen: Data curation; Investigation; Writing – review & editing.
Gabor Petzold: Data curation; Investigation; Writing – review & editing.
Fedor Levin: Methodology; Writing – review & editing.
Stefan Teipel: Conceptualization; Investigation; Methodology; Supervision; Writing – review & editing.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Qingyue Li receives funding from the European Union's HORIZON-MSCA Doctoral Networks 2021 program under grant agreement No 101071485. The PROSPECT-AD study is partially funded by the Alzheimer's Drug Discovery Foundation Diagnostic Accelerator program (ADDF). The funders of this study had no role in study design, data collection, data analysis, data interpretation, or writing the report.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Stefanie Koehler has received funding from the Alzheimer Drug Discovery Foundation and Lilly Deutschland GmbH, as well as lecture fees from Eisai. Martin Dyrba has received research funding from the Deutsche Forschungsgemeinschaft and lecture fees from Europäische Fernhochschule Hamburg GmbH. Alexandra Koenig, Nicklas Linz, and Elisa Mallick are employed by the company ki: elements, which developed the application for the automatic phone assessment and calculated the speech biomarkers. Nicklas Linz owns shares in the ki: elements company. Claudia Bartels received funding from the German Alzheimer's Association and lecture fees from Lilly, Boehringer Ingelheim, and Eisai. Franziska Maier participated in scientific advisory boards of Lilly and Johnson & Johnson, and received lecture fees from Novo Nordisk. Jens Wiltfang has received funding from the BMBF, consulting fees from Immunogenetics, Noselab, Roboscreen, and lecture fees from Beijing Yibai Science and Technology Ltd, Gloryren, Janssen Cilag, Pfizer, Med Update GmbH, Roche Pharma, and Lilly. Jens Wiltfang participated in a Data Safety Monitoring Board or Advisory Board of Biogen, Abbott, Boehringer Ingelheim, Lilly, MSD Sharp & Dohme, and Roche. Bjoern Falkenburger has received funding from the Deutsche Forschungsgemeinschaft. Ingo Kilimann has received an unrestricted research grant from Lilly Germany. Stefan Teipel participated in scientific advisory boards of Roche Pharma AG, Biogen, Lilly, and Eisai, and received lecture fees from Lilly and Eisai. Stefan Teipel is a senior editor in 2025, Alexandra Koenig, Inga Zerr, and Ayda Rostamzadeh are editorial Board members, but were not involved in the peer-review process of this article, nor had access to any information regarding its peer review. The other authors declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Data availability statement

The data are not publicly accessible but may be provided upon reasonable request.

Supplemental material

Supplemental material for this article is available online.

References

1. Vigo I, Coelho L and Reis S. Speech- and language-based classification of Alzheimer's disease: a systematic review. *Bioengineering (Basel)* 2022; 9: 27.
2. Taler V and Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *J Clin Exp Neuropsychol* 2008; 30: 501–556.
3. Pastoriza-Domínguez P, Torre IG, Diéguez-Vide F, et al. Speech pause distribution as an early marker for Alzheimer's disease. *Speech Commun* 2022; 136: 107–117.
4. Vita MG, Marra C, Spinelli P, et al. Typicality of words produced on a semantic fluency task in amnesic mild cognitive impairment: linguistic analysis and risk of conversion to dementia. *J Alzheimers Dis* 2014; 42: 1171–1178.
5. Verfaillie SCJ, Witteman J, Slot RER, et al. High amyloid burden is associated with fewer specific words during spontaneous speech in individuals with subjective cognitive decline. *Neuropsychologia* 2019; 131: 184–192.
6. Maseda A, Lodeiro-Fernández L, Lorenzo-López L, et al. Verbal fluency, naming and verbal comprehension: three aspects of language as predictors of cognitive impairment. *Aging Ment Health* 2014; 18: 1037–1045.
7. McDonnell M, Dill L, Panos S, et al. Verbal fluency as a screening tool for mild cognitive impairment. *Int Psychogeriatr* 2020; 32: 1055–1062.
8. Murphy KJ, Rich JB and Troyer AK. Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia. *J Int Neuropsychol Soc* 2006; 12: 570–574.
9. Shao Z, Janse E, Visser K, et al. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front Psychol* 2014; 5: 772.
10. Hamel R, Köhler S, Sistermans N, et al. The trajectory of cognitive decline in the pre-dementia phase in memory clinic visitors: findings from the 4C-MCI study. *Psychol Med* 2015; 45: 1509–1519.
11. Clark LJ, Gatz M, Zheng L, et al. Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer's disease. *Am J Alzheimers Dis Dement* 2009; 24: 461–468.
12. Estévez-González A, Kulisevsky J, Boltes A, et al. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging. *Int J Geriatr Psychiatry* 2003; 18: 1021–1028.
13. Troyer AK, Moscovitch M and Winocur G. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology* 1997; 11: 138–146.
14. Bruno D, Grothe MJ, Nierenberg J, et al. Output order and variability in free recall are linked to cognitive ability and hippocampal volume in elderly individuals. *Neuropsychologia* 2016; 80: 126–132.
15. Khurana D, Koli A, Khatter K, et al. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl* 2023; 82: 3713–3744.
16. Rudzicz F. Automatic speech recognition (ASR). In: *Clear speech. Synthesis lectures on assistive, rehabilitative, and health-preserving technologies*. Cham: Springer International Publishing, 2016, pp.17–22.
17. Beltrami D, Gagliardi G, Rossini Favretti R, et al. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Front Aging Neurosci* 2018; 10: 369.
18. Tröger J, Linz N, König A, et al. Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease. *Neuropsychologia* 2019; 131: 53–61.
19. Linz N, Lundholm Fors K, Lindsay H, et al. Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment. In: *Proceedings of the sixth workshop on computational linguistics and clinical psychology*. Minneapolis, MN: Association for Computational Linguistics, 2019, pp.103–113.
20. 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: 667–676.

21. Possemis N, Ter Huurne D, Banning L, et al. The reliability and clinical validation of automatically-derived verbal memory features of the verbal learning test in early diagnostics of cognitive impairment. *J Alzheimers Dis* 2024; 97: 179–191.
22. Hajjar I, Okafor M, Choi JD, et al. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. *Alzheimers Dement Diagn Assess Dis Monit* 2023; 15: e12393.
23. 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: fcac231.
24. Jack CR, Andrews JS, Beach TG, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's association workgroup. *Alzheimers Dement* 2024; 20: 5143–5169.
25. Whitwell JL. Progression of atrophy in Alzheimer's disease and related disorders. *Neurotox Res* 2010; 18: 339–346.
26. Paola M, Macaluso E, Carlesimo GA, et al. Episodic memory impairment in patients with Alzheimer's disease is correlated with entorhinal cortex atrophy: a voxel-based morphometry study. *J Neurol* 2007; 254: 774–781.
27. Grothe MJ, Heinsen H, Amaro E, et al. Cognitive correlates of basal forebrain atrophy and associated cortical hypometabolism in mild cognitive impairment. *Cereb Cortex* 2016; 26: 2411–2426.
28. Jones BF, Barnes J, Uylings HBM, et al. Differential regional atrophy of the cingulate gyrus in Alzheimer disease: a volumetric MRI study. *Cereb Cortex* 2005; 16: 1701–1708.
29. Convit A, De Asis J, De Leon MJ, et al. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiol Aging* 2000; 21: 19–26.
30. König A, Linz N, Baykara E, et al. Screening over speech in unselected populations for clinical trials in AD (PROSPECT-AD): study design and protocol. *J Prev Alzheimers Dis* 2023; 10: 314–321.
31. Jessen F, Spottke A, Boecker H, et al. Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). *Alzheimers Res Ther* 2018; 10: 15.
32. Jessen F, Amariglio RE, Van Boxtel M, et al. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimers Dement* 2014; 10: 844–852.
33. 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: 270–279.
34. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia 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: 263–269.
35. Arevalo-Rodriguez I, Smailagic N, Roqué-Figuls M, et al. Mini-mental state examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev* 2021; 27: CD010783.
36. Papp KV, Rentz DM, Orlovsky I, et al. Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: the PACC5. *Alzheimers Dement Transl Res Clin Interv* 2017; 3: 668–677.
37. Gaser C, Dahnke R, Thompson PM, et al. CAT: a computational anatomy toolbox for the analysis of structural MRI data. *Gigascience* 2024; 13: giae049.
38. Ashburner J. A fast diffeomorphic image registration algorithm. *NeuroImage* 2007; 38: 95–113.
39. Teipel SJ, Fritz HC, Grothe MJ, et al. Neuropathologic features associated with basal forebrain atrophy in Alzheimer disease. *Neurology* 2020; 95: e1301–e1311.
40. Levin F, Grothe MJ, Dyrba M, et al. Longitudinal trajectories of cognitive reserve in hypometabolic subtypes of Alzheimer's disease. *Neurobiol Aging* 2024; 135: 26–38.
41. Kilimann I, Grothe M, Heinsen H, et al. Subregional basal forebrain atrophy in Alzheimer's disease: a multicenter study. *J Alzheimers Dis* 2014; 40: 687–700.
42. Dyrba M, Hanzig M, Altenstein S, et al. Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. *Alzheimers Res Ther* 2021; 13: 191.
43. Crivello F, Tzourio-Mazoyer N, Tzourio C, et al. Longitudinal assessment of global and regional rate of grey matter atrophy in 1,172 healthy older adults: modulation by sex and age. *PLoS One* 2014; 9: e114478.
44. König A, Köhler S, Tröger J, et al. Automated remote speech-based testing of individuals with cognitive decline: Bayesian agreement of transcription accuracy. *Alzheimers Dement Diagn Assess Dis Monit* 2024; 16: e70011.
45. Tröger J, Baykara E, Zhao J, et al. Validation of the remote automated ki:e speech biomarker for cognition in mild cognitive impairment: verification and validation following DiME V3 framework. *Digit Biomark* 2022; 6: 107–116.
46. Geschwind N. The organization of language and the brain: language disorders after brain damage help in elucidating the neural basis of verbal behavior. *Science* 1970; 170: 940–944.
47. Poeppel D and Hickok G. Towards a new functional anatomy of language. *Cognition* 2004; 92: 1–12.
48. Nemy M, Dyrba M, Brosseron F, et al. Cholinergic white matter pathways along the Alzheimer's disease continuum. *Brain* 2023; 146: 2075–2088.
49. Jackson RL. The neural correlates of semantic control revisited. *NeuroImage* 2021; 224: 117444.
50. Meijs C, Hurks P, Rozendaal N, et al. Serial and subjective clustering on a verbal learning test (VLT) in children aged 5–15: the nature of subjective clustering. *Child Neuropsychol* 2013; 19: 385–399.
51. Heisig JP and Schaeffer M. Why you should *always* include a random slope for the lower-level variable involved in a cross-level interaction. *Eur Sociol Rev* 2019; 35: 258–279.

52. Lieberman MD and Cunningham WA. Type I and type II error concerns in fMRI research: re-balancing the scale. *Soc Cogn Affect Neurosci* 2009; 4: 423–428.
53. Shea O, Cohen A, Porges RA, et al. Cognitive aging and the hippocampus in older adults. *Front Aging Neurosci* 2016; 8: 298.
54. Dawe RJ, Yu L, Arfanakis K, et al. Late-life cognitive decline is associated with hippocampal volume, above and beyond its associations with traditional neuropathologic indices. *Alzheimers Dement* 2020; 16: 209–218.
55. Seyedi S, Jafari R, Talaei A, et al. Comparing VBM and ROI analyses for detection of gray matter abnormalities in patients with bipolar disorder using MRI. *Middle East Curr Psychiatry* 2020; 27: 69.
56. Venneri A, Mitolo M, Beltrachini L, et al. Beyond episodic memory: semantic processing as independent predictor of hippocampal/perirhinal volume in aging and mild cognitive impairment due to Alzheimer's disease. *Neuropsychology* 2019; 33: 523–533.
57. Whitwell JL, Shiung MM, Przybelski SA, et al. MRI Patterns of atrophy associated with progression to AD in amnestic mild cognitive impairment. *Neurology* 2008; 70: 512–520.
58. Hong YJ, Lee JH, Choi EJ, et al. Efficacies of cognitive interventions in the elderly with subjective cognitive decline: a prospective, three-arm, controlled trial. *J Clin Neurol* 2020; 16: 304–313.
59. Balota DA, Burgess GC, Cortese MJ, et al. The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: evidence for two processes in episodic recognition performance. *J Mem Lang* 2002; 46: 199–226.
60. Troyer AK, Moscovitch M, Winocur G, et al. Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia* 1998; 36: 499–504.
61. Hoyo LD, Xicota L, Sánchez-Benavides G, et al. Semantic verbal fluency pattern, dementia rating scores and adaptive behavior correlate with plasma Aβ42 concentrations in down syndrome young adults. *Front Behav Neurosci* 2015; 9: 301.
62. Bamford AR, Adams JN, Kim S, et al. The amyloid beta 42/38 ratio as a plasma biomarker of early memory deficits in cognitively unimpaired older adults. *Neurobiol Aging* 2024; 144: 12–18.
63. Ambrosini E, Giangregorio C, Lomurno E, et al. Automatic spontaneous speech analysis for the detection of cognitive functional decline in older adults: multilanguage cross-sectional study. *JMIR Aging* 2024; 7: e50537.