The Journal of Prevention of Alzheimer's Disease 13 (2026) 100689

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Original Article

Linguistic features from constrained spontaneous speech as potential markers of preclinical Alzheimer's disease

Haochun Huang [a], Zexu Li [b], Christina B. Young [c,d], Viktorija Smith [e], Ting Fang Alvin Ang [b,f,g], Vijaya B. Kolachalama [h,i], Rhoda Au [b,f,g,h,j,k], Jinying Chen [l,m,*]

[a] Bioinformatics Program, Faculty of Computing and Data Sciences, Boston University, Boston, MA, USA
[b] Department of Anatomy and Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[c] Department of Radiology, University of California Davis, Davis, CA, USA
[d] Department of Neurology, University of California Davis, Davis, CA, USA
[e] MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK
[f] Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[g] Slone Epidemiology Center, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[h] Department of Medicine/Section of Computational Biomedicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[i] Department of Computer Science and Faculty of Computing & Data Sciences, Boston University, Boston, MA, USA
[j] Department of Neurology and Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[k] Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA
[l] Department of Medicine/Section of Preventive Medicine and Epidemiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[m] Data Science Core, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA

ARTICLE INFO

Keywords:
Digital biomarker
Preclinical Alzheimer's disease
Natural language processing
Speech analysis
Neuropsychological test
tau
Amyloid

ABSTRACT

Background: Language impairment is an early sign of Alzheimer's disease (AD) that may emerge before more obvious cognitive decline. Linguistic features derived from verbal responses to cognitive tests have shown promise for identifying mild cognitive impairment and predicting AD risk; however, their relationships with established AD biomarkers remain insufficiently characterized.
Objectives: To examine relationships between a comprehensive set of linguistic features derived from logical memory delayed recall (LMDR) responses and amyloid status as well as regional tau burden among cognitively normal participants.
Design: In this cross-sectional study, we evaluated associations between two types of AD biomarkers (amyloid status and regional tau burden) and linguistic features using covariate-adjusted regression models tailored to each feature's distribution (e.g., linear, negative binomial, zero-inflated beta, and gamma regression). We then used LASSO regression and the Boruta feature-selection algorithm to identify features most informative of biomarker status.
Setting: A community-based subcohort from the Framingham Heart Study.
Participants: A total of 238 cognitively unimpaired participants who completed the LMDR test within 1 year before or after tau positron emission tomography (PET) were included. Of these, 228 also underwent amyloid PET and 227 had LMDR voice recordings available.
Measurements: Verbal responses to the LMDR test were digitally recorded and manually transcribed. We extracted 54 linguistic and paralinguistic features from transcribed responses using natural language processing, capturing lexical diversity and density, syntactic complexity, semantic/information density, and speech fluency and rate. Tau SUVR values were quantified from PET imaging in 5 brain regions (amygdala, entorhinal cortex, inferior temporal [IT], inferior parietal [IP], and precuneus). Amyloid status was classified as Aβ− or Aβ+ using Gaussian mixture modeling of PiB PET DVR values. LMDR total score and participant demographics (age, sex, education) were included as covariates.

Short running page headline: Linguistic Markers of Preclinical AD

* Corresponding author at: Boston University Chobanian & Avedisian School of Medicine, 71 E Concord St. Boston, MA, 02118, USA.
E-mail address: jinychen@bu.edu (J. Chen).

https://doi.org/10.1016/j.tjpad.2026.100689
Received 6 March 2026; Received in revised form 25 May 2026; Accepted 16 October 2026
Available online 5 November 2026
2274-5807/© 2026 The Authors. Published by Elsevier Masson SAS on behalf of SERDI Publisher. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Results: After adjusting for age, sex, education, amyloid status, and LMDR score, 23 linguistic features were associated with tau burden in at least one brain region. For example, reduced semantic richness, captured by lower uniqueIdeaDensity, was associated with higher tau burden across multiple regions, including the amygdala (average marginal effect (AME)=-0.007, 95% CI: (-0.013, -0.001)), entorhinal (AME=-0.007, 95% CI: (-0.013, -0.002)), IT (AME=-0.008, 95% CI: (-0.014, -0.002)), and precuneus (AME=-0.006, 95% CI: (-0.012, -1.51e^-5)). More between-utterance pauses were associated with higher tau burden in the entorhinal (AME=0.259, 95% CI: (0.026, 0.493)), IT (AME=0.272, 95% CI: (0.025, 0.518)), and IP (AME=0.404, 95% CI: (0.160, 0.648)) regions. No linguistic features were associated with amyloid status, and LMDR total score was not associated with either tau burden or amyloid status. LASSO regression selected 3–7 features per brain region, capturing multiple aspects of language use (e.g., speech disfluency and semantic density) associated with tau burden. The nonlinear Boruta algorithm selected a larger set of biomarker-informative features, including predictors of amyloid positivity, and identified additional syntactic measures (e.g., Yngve score, mean dependency distance) as important predictors of regional tau burden.
Conclusions: Linguistic features derived from neuropsychological (NP) tests capture subtle language changes associated with regional tau pathology in cognitively normal individuals, beyond conventional NP test scores. These findings highlight the potential of language-based measures as non-invasive digital biomarkers for early AD risk stratification and for supporting clinical decision-making about biomarker testing for possible preclinical AD.

1. Introduction

Language impairment is one of the early signs associated with Alzheimer's disease (AD), often appearing well before clinical diagnosis [1]. Difficulties in lexical retrieval, reduced semantic fluency, and declines in grammatical complexity and vocabulary diversity have been observed in mild cognitive impairment (MCI) and early AD [2–4]. Automated speech analysis offers promising tools for quantifying language impairment and supporting early diagnosis of AD [5,6]. Prior studies have shown that linguistic features from spontaneous speech, such as syntactic complexity, lexical diversity, disfluencies, and word categories reflecting psychological processes (e.g., positive and negative emotions) and relativity (e.g., time, space, motion), are informative for distinguishing between normal cognition and MCI [2,4,7–12]. However, most studies have focused on classifying MCI versus normal cognition, which is limited by the heterogeneity of MCI, including its diverse etiologies and clinical trajectories. Evaluating linguistic features against established AD biomarkers may yield more AD-specific insights; however, research remains limited [13,14], and preclinical AD studies have examined only a narrow set of lexical and paralinguistic features [15,16].

A recent study investigated speech patterns observed in cognitively normal participants of the Framingham Heart Study (FHS) during a delayed recall (of a story) memory task [14]. It examined five paralinguistic speech features and found that longer and more between-utterance pauses and a slower speech rate were associated with elevated tau signals in the medial temporal and early neocortical regions [14]. These findings highlight the potential of speech features for monitoring AD risk in the early stage of tau pathology. Building on this work and using the same study setting [14], this exploratory study aimed to systematically assess a wide range of linguistically motivated features derived from delayed recall responses (Fig. 1). We leveraged natural language processing (NLP) techniques to automate the feature generation process. We then assessed the associations between linguistic features and the global amyloid status and regional tau signal using both statistical and machine learning methods. This study aimed to answer the following questions:

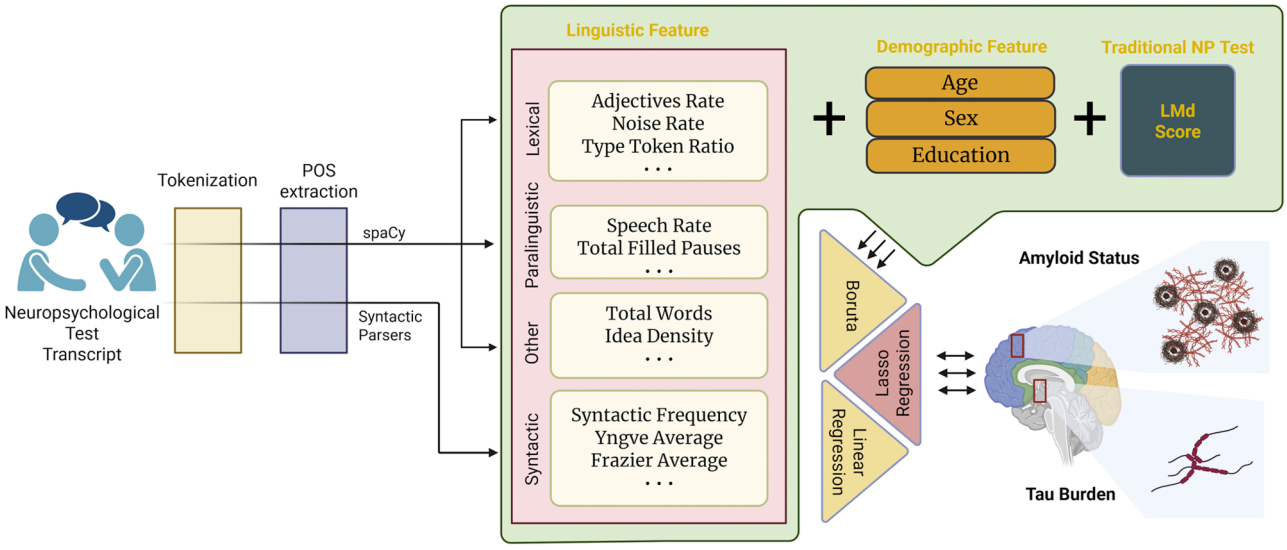

Fig. 1. Overview of the analytical pipeline for investigating associations between linguistic features with two types of AD biomarkers. After preprocessing (data cleaning), neuropsychological test transcripts were tokenized and processed for part-of-speech (POS) tagging using spaCy. Syntactic features were extracted using the spaCy dependency parser and two constituency parsers. These procedures generated linguistic features across lexical (e.g., adjectives rate, idea density, type-token ratio), paralinguistic (e.g., speech rate, filled pauses), and syntactic (e.g., Yngve score) domains. These features are combined with demographic data (age, sex, and education) and LMd (i.e., LMDR) score. Statistical and machine learning models, including Boruta, regression models (linear regression etc.) and Lasso regression, are used to explore associations with biomarkers such as amyloid status and tau burden.

- In addition to the few features identified in prior work (e.g., between-utterance pauses and speech rate), are other linguistic features sensitive to subtle language changes linked to early beta-amyloid deposition and tau burden in cognitively normal individuals?
- Which linguistic features are most important to include in a future screening tool for predicting early beta-amyloid deposition and tau burden? Such a tool could support clinical decision-making about whether to order formal AD biomarker testing for an individual.

2. Methods

2.1. Study design and participants

Using a cross-sectional design, we analyzed data collected from participants of the FHS, a community-based longitudinal study initiated in 1948 and further expanded to include 5 additional subcohorts since 1971 [17]. Since 2005, FHS has implemented digital audio recording of neuropsychological (NP) testing to enhance standardized test administration and scoring processes [18]. These digital recordings have evolved into a rich speech dataset and have been used to identify speech and language markers indicative of AD and MCI [11,19,20]. Between 2015 and 2022, a subset of participants from the Offspring, Generation 3, and Omni 1 subcohorts were invited to undergo brain positron emission tomography (PET) scans, allowing for the acquisition of AD biomarkers such as amyloid status and tau burden.

We included participants who underwent tau PET scans and completed at least one logical memory delayed recall (LMDR) test within one year before or after the scan. We excluded one participant with extreme tau PET signals in five brain regions (i.e., exceeding three standard deviations from the mean), yielding a final analytic sample of 238 participants (Supplementary File 1, Figure S1-1). Among these participants, 227 had transcripts of the LMDR test (Anna Thompson story) and 228 had amyloid PET scans. The amyloid PET scans were conducted on the same day or within 3 months of the tau PET. All participants in the final sample were cognitively normal, defined as having no diagnosis of MCI or dementia, at the time they completed the LMDR test and underwent tau and amyloid PET scans.

This study involved secondary analysis of de-identified data and was approved by the Institutional Review Board of Boston University Chobanian & Avedisian School of Medicine, with written informed consent waived per institutional and regulatory guidelines.

2.2. Amyloid-β and tau PET measures

We used the same biomarkers examined in a prior study [14]. Briefly, global amyloid status was classified into negative (Aβ-) and positive (Aβ+) based on Pittsburgh compound B (PiB) data obtained from PET scans using Gaussian mixture modeling and a distribution volume ratio (DVR) cutoff corresponding to a 90% probability threshold [14]. For tau PET, volume-weighted SUVRs were calculated for five target regions of interest (amygdala, entorhinal cortex, inferior temporal [IT], inferior parietal [IP], and precuneus). We also included putamen as a control region with known off-target binding.

2.3. Preprocessing of voice recordings

The transcripts used in this study were derived from the audio recordings of the LMDR test, as described in detail in a prior study [14]. Briefly, a total of 227 recordings were manually transcribed into text using the EUDICO Linguistic Annotator (ELAN) tools based on the Code for the Human Analysis of Transcripts (CHAT) format system. Utterance segmentation followed the protocols outlined in the Language Development Project [21]. Only participants' speech was retained for analysis. Transcript formatting symbols were removed to generate plain text (Supplementary File 1, Table S1-1); pause- and disfluency-related markers were first used to derive paralinguistic features (Supplementary File 1, sMethod 1) and were then removed. The cleaned transcripts were then processed using spaCy (version 3.7.2) for tokenization and part-of-speech (POS) tagging. Linguistic features (Supplementary File 1, Table S1-2) were then automatically extracted from the processed transcripts using natural language processing methods.

2.4. Lexical features

Lexical features measure word use and vocabulary choice in speech, such as lexical diversity, density, and sophistication. We generated 36 lexical features (Supplementary File 1, Table S1-2), including: (1) total unique words used in story retelling; (2) POS rate and unique POS rate: the number of words (or unique words) with a specific POS tag divided by total words; (3) type-token ratio: total unique words divided by total words [8]; (4) content density, unique content density, and unique content density unique: total open words divided by total words, total unique open words divided by total words, and total unique open words divided by total unique words. Open words were defined as those with specific POS tags including nouns, verbs, adjectives, and adverbs; (5) pronoun to proper noun ratio [11]; (6) mean word length: total characters divided by total words; and (7) mean number of unique words per utterance. For any feature whose denominator could be zero, we added 1 to the denominator to avoid division-by-zero. Some previous studies [8] also counted symbols as open-class words. We excluded symbols because most of them carried specific meanings in CHAT feature analysis (Supplementary File 1, Table S1-1) and were not directly relevant to story-retelling content.

2.5. Paralinguistic features

We implemented 7 paralinguistic features that measure speech rate and disfluency, including: (1) speech rate and unique speech rate: defined as words per minute and unique words per minute, respectively; (2) within-utterance pauses: the number of "(.)" in the transcript, which indicates pauses shorter than 2 s during participants' speech; (3) filled pauses: the number of words which start with '&-' in the transcript; (4) log-transformed between-utterance pause time: the log-transformed total duration of between-utterance pauses; (5) the number of between-utterance pauses; (6) standardized pause rate: defined as total words divided by the total number of between-utterance pauses, filled pauses, and within-utterance pauses. Note that, aside from the two consecutive utterances produced by the administrator, all pauses longer than 2 s occurring between the remaining utterance pairs were counted as between-utterance pauses.

2.6. Syntactic features

We used complementary syntactic parsing approaches to characterize linguistic structure in speech transcripts. Features based on phrase-structure parse trees (i.e., constituency parse trees) were derived using the Berkeley Neural Parser through the benepar package (version 0.2.0; benepar_en3 model). Dependency-based features were derived using the spaCy dependency parser (version 3.7.2; en_core_web_sm model). The *Syntax Frequency* feature was derived using a published computational pipeline built on the Stanford CoreNLP constituency parser (Stanford CoreNLP, version 2018-10-05).

2.7. Features derived from phrase-structure parse trees

We implemented three features that quantify utterance-level syntactic complexity and processing load based on the hierarchical structure of syntactic parse trees: Yngve score [22],Frazier score [23], and Frazier-Roark score [24]. Higher values of these scores generally indicate greater syntactic complexity (i.e., higher processing load) in a sentence's structure. The Yngve score quantifies syntactic processing

load by approximating the depth of a pushdown stack during top-down, left-to-right parsing; higher values indicate more pending syntactic structure at each word of a sentence (often due to greater left-branching or embedding in the parse tree) (Supplementary File 1, Figure S1-2a). Lower utterance-level Yngve score in verbal responses to the LMDR task has been associated with MCI [8]. The Frazier score [23] and the Frazier-Roark score [24] also quantify left-to-right syntactic processing load, but they operationalize it differently (Supplementary File 1, Figure S1-2b). The Frazier score has been negatively associated with age [25]. We provide additional details on how the Yngve, Frazier, and Frazier-Roark scores are calculated in Supplementary File 1 (sMethod 2).

2.8. MDD score

Mean dependency distance (MDD) measures the complexity of a sentence's dependency structure [26,27]. Specifically, it calculates the average distance between each word and its center or head word in the dependency tree of a sentence (Supplementary File 1, Figure S1-2c). Longer dependency distances increase the working memory load, making the sentence harder to process. Therefore, a high MDD score reflects greater sentence complexity. MDD is widely used in linguistics and NLP to assess the linguistic complexity of different text types [28,29]. MDD calculated from written text has been associated with cognitive impairment severity [30] and lower MDD in verbal responses to the LMDR task has been associated with MCI [8].

2.9. Nodes

This feature was defined as the number of parse-tree nodes per word in each sentence [24]. Higher values indicate greater hierarchical sentence organization and may also reflect increased cognitive load (Supplementary File 1, sMethod 2).

2.10. Syntax frequency

Syntax frequency measures syntactic complexity based on the specific syntactic rules used in a sentence and their frequencies in an external, large text corpus [31]. The lower the frequency, the more complex the rule is. The syntax frequency of a sentence is calculated as the mean log frequency of the syntactic rules used for this sentence (Supplementary File 1, Figure S1-2d and sMethod 2).

2.11. Semantic/information density features

Idea density was defined as the number of propositions divided by total words in a sentence, where the propositions correspond to verbs, adjectives, adverbs, prepositions, and conjunctions [8]. In addition, we defined the feature *unique idea density* as the number of unique propositions divided by total words in a sentence.

2.12. Other features

We designed features to quantify participants' off-topic speech (i.e., not directly related to the story content) during the NP test; for convenience, we refer to these as noise features. Specifically, we defined *total noise* as the number of occurrences of the words *"I," "you," "forget," "remember,"* and *"recall"* in the dialogue (e.g., *"I don't remember the name.", "I forget what it was.", "uh. I'm sorry I can't recall."*), and noiseRate as totalNoise divided by the total number of words. We also defined *total noise sentences* as the number of sentences containing at least one of these words, and noiseSentenceRate as totalNoiseSentences divided by the total number of utterances.

We also included the total number of words produced during storytelling and the mean number of words per utterance as features.

2.13. Statistical analysis

We reported demographics, the LMDR score, and tau burden in 5 brain regions of study participants across the Aβ+ and Aβ- groups. In addition, we grouped linguistic features into clusters based on their correlations to facilitate result interpretation.

We examined the relationship between each linguistic feature and (1) the LMDR score, (2) the beta-amyloid status, and (3) the regional tau burden as follows.

First, we assessed the association between each linguistic feature and the LMDR score by using (1) Spearman's correlation coefficient and (2) a multi-variable regression model, with the LMDR score as the primary independent variable and a linguistic feature as the dependent variable, to adjust for age, sex, and education.

Given the heterogeneous distributional characteristics of linguistic feature values (Supplementary File 1, Table S1-3), we applied regression models appropriate for each feature. For features bounded between 0 and 1 and exhibiting a large number of zeros, we used a zero-inflated beta regression model implemented in *glmmTMB* [32]. For positively skewed continuous features, we applied Gamma regression with a log link function from package *stats* [33]. Discrete count-based features with many zeros were modeled using negative binomial regression from package *MASS* [34]. Features with approximately normal distributions were analyzed using linear regression. For each model, we computed average marginal effects using the *marginaleffects* [35] to quantify the expected change in each linguistic feature associated with a change in LMDR score.

Second, we assessed the associations between each linguistic feature and beta-amyloid status, with the beta-amyloid status as the primary independent variable and a linguistic feature as the dependent variable, using regression models (as described above) to adjust for age, sex, and education. We repeated the above analysis by using the LMDR score as dependent variables, for comparisons.

Third, we assessed the associations between each linguistic feature and tau burden in five brain regions. For each region, we fit regression models (as described above) with the regional tau burden as the primary independent variable and the linguistic feature as the dependent variable, adjusting for age, sex, education, beta-amyloid status, and LMDR score. For comparison, we examined the associations between LMDR score (dependent variable) and regional tau burdens using linear regression, adjusting for age, sex, education, and beta-amyloid status.

In addition, we conducted two sensitivity analyses. In the first analysis, we assessed whether excluding off-topic sentences affected associations between each linguistic feature and the two types of AD biomarkers by repeating the regression analyses after removing these sentences from transcripts. Off-topic sentences were defined using the same criteria as for the *total noise sentences* feature. In the second analysis, we repeated the primary analysis assessing the associations between each linguistic feature and amyloid status, using the continuous amyloid measure instead of the binary variable.

2.14. Feature selection by machine learning

We utilized two machine learning methods—LASSO regression and Boruta algorithm—to identify linguistic features most important for inclusion in a future screening tool that predicts early changes in amyloid status and regional tau burden.

In the first approach, we began with linguistic features that were associated with beta-amyloid status or regional tau burden in the statistical analyses and then used lasso regression to identify the subset most indicative of each AD biomarker. Lasso regression is a regularized regression method that uses an L1 penalty to shrink coefficients of less relevant features (predictors) to zero. We implemented LASSO models using the *glmnet* [36] R package, setting the parameter alpha to 1 and the model type (i.e., parameter *family*) to gaussian for tau SUVR and binomial for amyloid status. We adjusted the models for (1) age, sex,

education, and the LMDR score when predicting the amyloid status and (2) age, sex, education, the LMDR score, and the amyloid status when predicting the regional tau burdens.

Lasso regression assumes independence between features. To better capture the joint effects and account for potentially non-linear relationships, our second approach used the Boruta algorithm to identify linguistic features indicative of each AD biomarker. Boruta wraps a machine learning model (here, a random forest model) and identifies important features by comparing each feature's importance with that of permuted "shadow" features (as detailed below) [37]. We input all linguistic features along with age, sex, education, beta-amyloid status, and the LMDR score into Boruta. Boruta was then used to evaluate feature importance for predicting amyloid status and tau burden across five brain regions, separately.

To ensure the robustness of Boruta feature selection, we ran Boruta 20 times for each AD biomarker using different random seeds. In each trial, features were iteratively compared against randomly permuted "shadow" features, and classified as either "important" or "unimportant" based on whether their importance consistently exceeded that of the shadow features [37]. For each feature, we counted the number of trials in which it was identified as "important," labeling features as relatively important if they were selected $\geq$10 times and highly important if selected $\geq$18 times.

3. Results

3.1. Characteristics of participants

A total of 238 participants (age: 54.9 ± 8.3 years, 51.3% females, 89.1% had college or higher education, 7.1% were amyloid positive) were included in this study (Table 1). Participants in the Aβ+ group were older than participants in the Aβ- group (65.2 ± 5.6 vs. 54.3 ± 7.8, $P < 0.001$). The LMDR total scores did not differ between the Aβ+ and Aβ- groups. Compared with the Aβ- group, the Aβ+ group showed higher tau burden in all regions except the amygdala after adjusting for age (Table 1; Supplementary File 2, Table S2-1).

3.2. Clusters of linguistic features

Dendrogram analysis revealed four distinct clusters of linguistic features (Fig. 2). Cluster 1 (orange) comprises features reflecting speech rate and disfluency (e.g., pauses, fillers, hesitations). Cluster 2 (green) are mostly features representing syntactic complexity and discourse-planning demands; the latter (e.g., between-utterance pauses) often co-occur with shorter, more fragmented sentences. Cluster 3 (red) captures lexical (e.g., unique nouns, adjectives, or verbs) and semantic richness (e.g., unique ideas) and reflects how efficiently speakers convey meaning. Cluster 4 (purple) is a broad cluster primarily characterized by features capturing overall lexical diversity (e.g., type-token ratio), function-word richness (e.g., unique auxiliary words, pronouns, or prepositions), and off-topic speech. This cluster reflects broad patterns of lexical choice, usage, and diversity in speech.

3.3. Speech features and logical memory delayed recall

Among the 54 features examined by Spearman's correlation analysis, 35 features showed significant correlations (16 negative and 19 positive) with the LMDR score (Supplementary File 2, Table S2-2, Columns B and C). After adjusting for demographic variables (Supplementary File 2, Table S2-2, Columns D-G), five features (uniquePropernounsRate, propernounsRate, totalNoiseSentences, withinUtterancePause, standardizedPauseRate) were not associated with the LMDR score anymore, whereas logBetweenUtterancePause became significantly associated with LMDR (average marginal effect (AME)=0.031, 95% CI: (0.002, 0.060), $P = 0.03$). The direction and magnitude of associations remained relatively stable before and after adjusting for covariates.

Table 1
Characteristics of participants, by amyloid-beta (Aβ) status.

	All (n = 238)	Aβ− (n = 211, 88.7%)	Aβ+ (n = 17, 7.1%)	Unknown Aβ status (n = 10, 4.2%)	*P* value [a]
Age at tau PET scan, mean (std)	54.9 (8.3)	54.3 (7.8)	65.2 (5.6)	51 (12.4)	<0.001 *
Sex, n (%)					0.04*
Male	116 (48.7)	106 (50.2)	4 (23.5)	6 (60)	
Female	122 (51.3)	105 (49.8)	13 (76.5)	4 (40)	
Education, n (%)					0.31
High school or lower	26 (10.9)	23 (10.9)	1 (5.8)	2 (20)	
Some college	70 (29.4)	60 (28.4)	8 (47.1)	2 (20)	
College graduate	142 (59.7)	128 (60.7)	8 (47.1)	6 (60)	
LMDR score, mean (std)	12.2 (3.8)	12.4 (3.8)	11.1 (3.6)	12.0 (4.7)	0.18
Tau SUVR, mean (std) [b]					
Amygdala	1.15 (0.11)	1.14 (0.11)	1.18 (0.12)	1.13 (0.13)	0.26
Entorhinal	1.05 (0.08)	1.05 (0.08)	1.08 (0.09)	1.04 (0.07)	0.01*
Inferior parietal	1.05 (0.08)	1.05 (0.08)	1.13 (0.07)	1.02 (0.05)	0.002*
Inferior temporal	1.12 (0.07)	1.11 (0.07)	1.19 (0.07)	1.10 (0.05)	0.004*
Precuneus	1.08 (0.07)	1.08 (0.07)	1.11 (0.07)	1.08 (0.07)	0.01*

[a] The difference of participants' characteristics between the Aβ− group and the Aβ+ group were assessed by *t*-tests (for age, LMDR score), chi-squared test (for sex), and Fisher's exact test (for education).
[b] Tau burden was measured as tau standardized uptake value ratio (SUVR). Tau burden was compared between the Aβ− and Aβ+ groups using age-adjusted linear regression (Supplementary File 2, Table S2–1).
* indicates statistically significant ($P < 0.05$).

3.4. Speech features and amyloid status

After adjusting for age, sex, and education, neither the linguistic features nor the LMDR score were associated with amyloid status (Supplementary File 2, Table S2-3). We also did not identify any linguistic features associated with amyloid status in either sensitivity analysis: excluding off-topic sentences or using the continuous amyloid measure (Supplementary File 2, Tables S2-4 and S2-5).

3.5. Speech features and regional tau burden

After adjusting for age, sex, education, and amyloid status, the LMDR score was not associated with regional tau burdens (Table 2, Supplementary File 2, Table S2-6).

After adjusting for age, sex, education, LMDR score, and amyloid status, 23 linguistic features were associated with tau signals in at least one brain region studied (Table 3; detailed in Supplementary File 2, Table S2-7). For example, more between-utterance pauses (betweenUtterancePauseNumber) were associated with higher tau burden in the entorhinal (AME=0.259, 95% CI: (0.026, 0.493)), IT (AME=0.272, 95% CI: (0.025, 0.518)), and IP (AME=0.404, 95% CI: (0.160, 0.648)) regions, indicating that a 1-SD increase in regional tau burden was associated with 0.259–0.404 additional pauses. Lower unique idea density (fewer unique ideas per word) was associated with higher tau burden in four regions (Amygdala: AME=−0.007, 95% CI: (−0.013, −0.001); Entorhinal: AME=−0.007, 95% CI: (−0.013, −0.002); IT: AME=−0.008, 95% CI: (−0.014, −0.002); Precuneus: AME=−0.006,

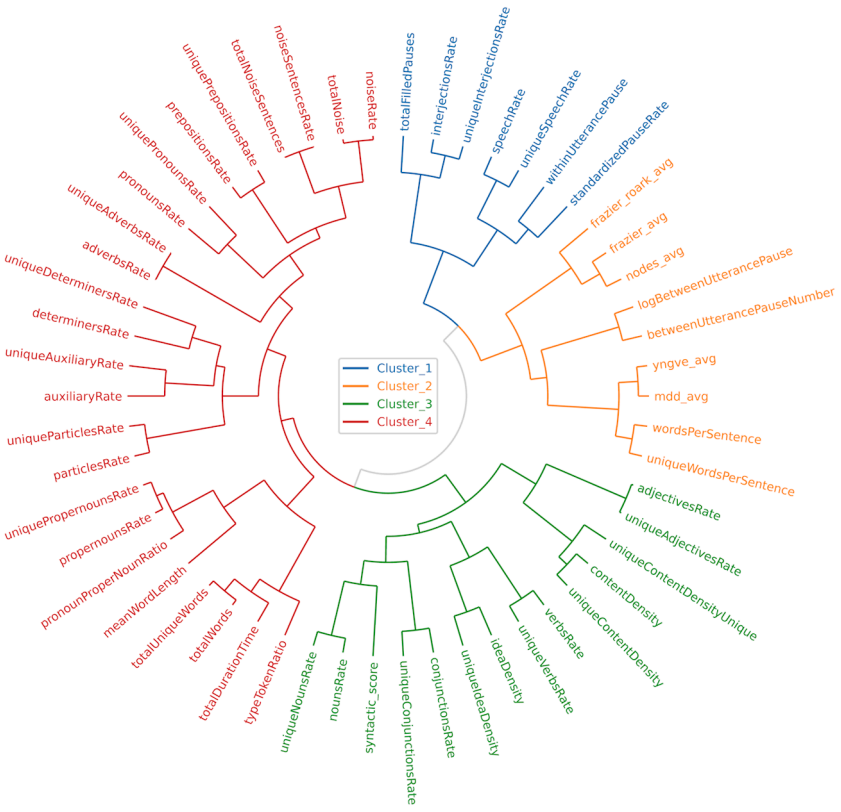

Fig. 2. Hierarchical clustering of linguistic features. A Spearman correlation matrix was computed across all linguistic features, symmetrized, and transformed into a distance matrix using $1 - |\rho|$. Hierarchical clustering was performed using Ward's linkage to group features with similar correlation patterns. The resulting dendrogram is displayed in a circular layout, with branch lengths representing dissimilarity based on the correlation-derived distance. Branch colors denote distinct feature clusters identified at a cut height of 1.7.

95% CI: (−0.012, −1.5eˆ−5), indicating that a 1-SD increase in regional tau burden was associated with 0.6–0.8 fewer unique ideas per 100 words.

The sensitivity analysis excluding off-topic sentences yielded more statistically significant associations between features representing syntactic complexity (Cluster 2; e.g., mdd_avg, nodes_avg, frazier_avg, and wordsPerSentence) and regional tau burden (Supplementary File 2, Table S2–8). Despite this difference, the overall patterns were similar in both analyses: (1) no linguistic features were associated with amyloid status in either analysis; (2) 5 associations significant in the primary analysis became non-significant in the sensitivity analysis, but the AMEs were similar in magnitude and direction (Supplementary File 2, Table S2–8); (3) the directions of all associations were consistent between the two analyses.

InterjectionsRate and adverbsRate were associated with putamen (control) tau signal in both primary and sensitivity analyses, while uniqueAdverbsRate was associated with putamen tau signal only in the sensitivity analysis (Supplementary File 2, Tables S2–6 & S2–7).

3.6. Feature selection by LASSO regression

As shown in Fig. 3 (detailed in Supplementary File 2, Table S2–9), Lasso regression removed several speech features identified by individual regression models (Panel A vs. Panel B), most of which reflected speech disfluency (e.g., pauses). For example, for tau burden in the IT region, LASSO selected 2 positive predictors (e.g., logBetweenUtterancePause, β=0.088, uniqueAuxiliaryRate, β=0.066) and 4 negative predictors (e.g., uniqueConjunctionsRate,β=−0.074, uniqueIdeaDensity, β=−0.059, and uniquePrepositionsRate, β=−0.072). In contrast, for the IP and Amygdala regions, Lasso did not remove any linguistic features.

3.7. Feature selection by Boruta

In predicting amyloid status (Supplementary File 2, Table S2-10), age was the only highly important feature (selected in n = 20 trials) and no linguistic features were identified as important.

For tau burden prediction (Panel C in Fig. 3; Supplementary File 2, Table S2-11), Boruta identified several features as highly important across multiple regions, including uniquePrepositionsRate (n = 20 for IT, IP, and entorhinal), logBetweenUtterancePause (n = 20 for IT and IP), uniqueIdeaDensity (IT: n = 20; entorhinal: n = 18), yngve_avg (amygdala: n = 20; entorhinal: n = 18), mdd_avg (n = 18 for amygdala and precuneus), and standardizedPauseRate (amygdala: n = 20; precuneus: n = 19).

4. Discussion

4.1. Principal findings

Research evaluating spontaneous-speech-derived linguistic features against established AD biomarkers remains limited. The few preclinical AD studies have examined only a small number of features [14–16] and only one has evaluated associations with regional tau [14]. This study analyzed 54 linguistic features automatically extracted from transcribed LMDR voice responses in cognitive normal individuals. We identified 23 features associated with tau burden in at least one of the five brain regions examined; only interjectionsRate was also associated with tau signal in the control region, supporting the specificity of our findings. The associated features span multiple dimensions of language use, including lexical diversity and density, information density, syntactic complexity, and speech fluency. Most features remained indicative of regional tau burden after feature selection using lasso regression. The

Table 2
The associations between regional tau signals and linguistic features, adjusted for age, sex, education level, LMDR score, and Amyloid status.

Linguistic feature	Average Marginal Effects (AME), (95% CI)				
	Amygdala	Entorhinal	Inferior parietal	Inferior temporal	Precuneus
LMDR	−0.280 (−0.738,0.179)	−0.161 (−0.604, 0.282)	0.051 (−0.428, 0.529)	−0.218 (−0.685, 0.248)	−0.075 (−0.520, 0.371)
Cluster 1 feature: speech rate and disfluency					
interjectionsRate	**0.006 (4e^−4, 0.011)**	—	—	—	**0.006 (0.001, 0.012)**
speechRate	—	—	—	—	**−0.095 (−0.166, −0.023)**
totalFilledPauses	—	—	—	—	**0.322 (0.012, 0.631)**
uniqueSpeechRate	—	**−0.071 (−0.132, −0.011)**	—	—	**−0.087 (−0.147, −0.026)**
withinUtterancePause	—	—	—	—	**0.321 (0.007, 0.635)**
Cluster 2 feature: syntactic complexity and discourse-planning demands					
betweenUtterancePauseNumber	—	**0.259 (0.026, 0.493)**	**0.404 (0.160, 0.648)**	**0.272 (0.025, 0.518)**	—
logBetweenUtterancePause	—	**0.100 (0.003, 0.197)**	**0.185 (0.082, 0.287)**	**0.140 (0.039, 0.242)**	—
uniqueWordsPerSentence	—	**−0.162 (−0.313, −0.011)**	—	**−0.164 (−0.324, −0.005)**	—
yngve_avg	—	—	—	**−0.025 (−0.049, −0.002)**	—
Cluster 3 feature: lexical and semantic richness					
adjectivesRate	—	—	—	—	**−0.003 (−0.005, −4e^−4)**
ideaDensity	**−0.007 (−0.013, −3e^−4)**	—	—	—	—
uniqueAdjectivesRate	—	—	—	—	**−0.003 (−0.005, −2e^−4)**
uniqueConjunctionsRate	—	—	—	**−0.003 (−0.005, −3e^−4)**	—
uniqueIdeaDensity	**−0.007 (−0.013, −0.001)**	**−0.007 (−0.013, −0.002)**	—	**−0.008 (−0.014, −0.002)**	**−0.006 (−0.012, −1.5e^−5)**
Cluster 4 feature: lexical diversity, function-word richness and off-topic speech					
prepositionsRate	—	—	**−0.006 (−0.011, −0.001)**	—	—
totalDurationTime	—	**1.137 (0.052, 2.221)**	—	—	—
totalNoiseSentences	—	—	—	—	**0.136 (0.016, 0.255)**
totalWords	—	—	**3.536 (0.466, 6.606)**	—	—
uniqueAuxiliaryRate	—	—	**0.003 (2e^−4, 0.006)**	**0.003 (3e^−4, 0.006)**	—
uniqueDeterminersRate	—	—	**−0.002 (−0.004, −0.001)**	—	—
uniquePrepositionsRate	—	—	**−0.006 (−0.010, −0.002)**	**−0.005 (−0.009, −0.001)**	—
uniquePronounsRate	—	—	—	—	**−0.004 (−0.008, −0.001)**

[a]The associations between regional tau signals and linguistic features were assessed using regression models, adjusted for age, sex, education, LMDR score, and amyloid status (detailed in Supplementary File 2, Table S2–7). Given the heterogeneous distributional characteristics of the linguistic features, we used regression models tailored to each feature type and report average marginal effects (AME), which estimate the expected change in each linguistic feature per 1-SD increase in regional tau burden. The associations between regional tau signals and the LMDR score were assessed using linear regression, adjusted for age, sex, education, and amyloid status (detailed in Supplementary File 2, Table S2–6). Only significant associations ($P < 0.05$) were reported.

all-relevant feature selection method Boruta identified additional predictors, such as yngve_avg and mdd_avg (which measured syntactic complexity) for amyloid positivity and tau burden in amygdala, entorhinal, and precuneus.

4.2. Linguistic features and amyloid/tau biomarkers

We found that, although the LMDR score was not associated with any of the AD biomarkers analyzed, several linguistic features were associated with regional tau burden, even after adjusting for the LMDR score and demographic covariates. Correlations between linguistic features and the LMDR score were weak to moderate, indicating that the linguistic features—though derived from LMDR responses—capture information beyond what is summarized by the LMDR score. Together, these findings suggest that speech patterns elicited during the LMDR test may provide information about underlying AD pathology beyond what is reflected in the LMDR score alone.

Aligned with previous work [14], we found that features that measured pauses and speech rate were associated with tau burden in multiple brain regions. In particular, more between-utterance pauses and longer between-utterance pause duration were associated with tau accumulation in the entorhinal, IP, and IT. In addition, reduced speech rate and more filled or within-utterance pauses were associated with higher precuneus tau. The medial temporal lobe (MTL)—especially the transentorhinal/entorhinal region—is among the earliest regions affected by tau pathology in AD [38,39]. As tau pathology spreads beyond the MTL, early involvement is also observed in neocortical regions including the IT, IP, and precuneus [38–40]. The entorhinal

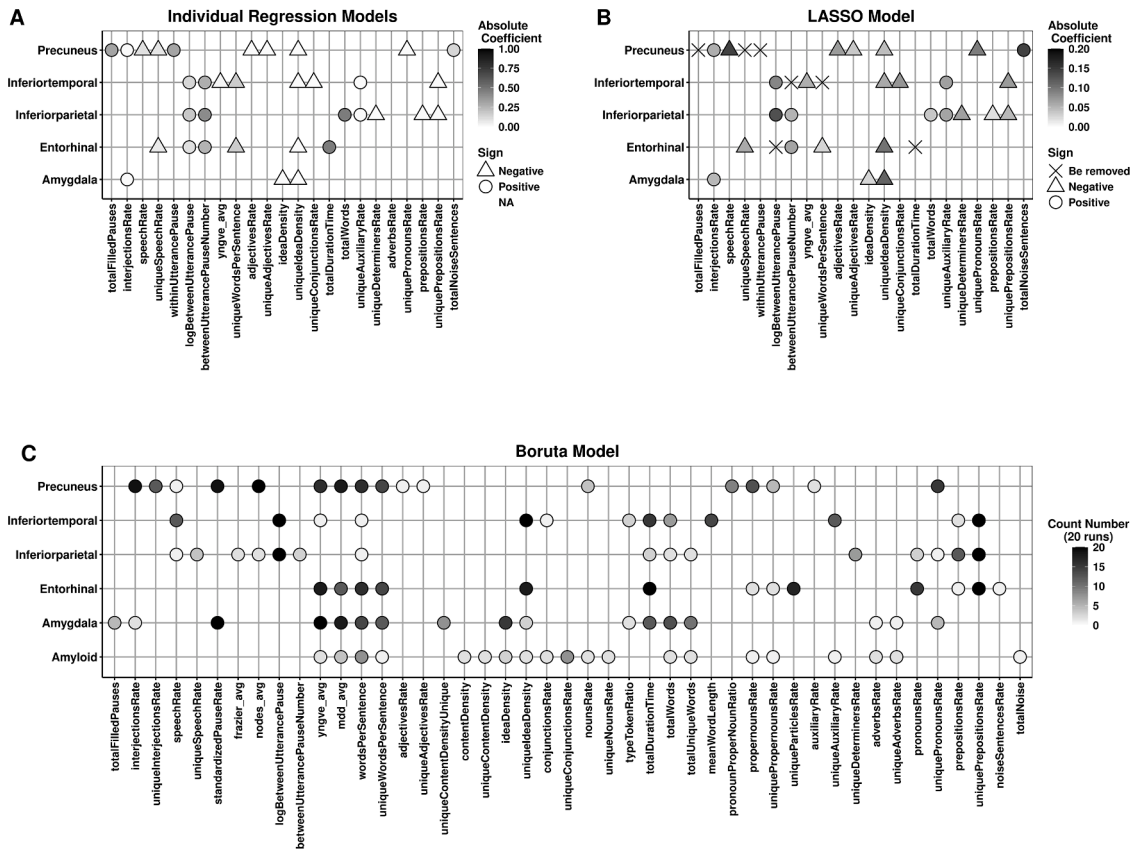

Fig. 3. Comparison of linguistic features' associations with regional tau burden and amyloid status across different models. The three panels illustrate results from individual regression models, LASSO model, and Boruta model: Panel A (Individual Regression Models): Displays the absolute coefficients for linguistic features associated with regional tau burden and amyloid status. Circles and triangles indicate the sign of the coefficients (positive or negative). Panel B (LASSO Model): Shows the selected linguistic features from Lasso regression after applying feature selection by the individual regression models. Circles and triangles represent selected features, while "X" indicates features removed by LASSO. Panel C (Boruta Model): Highlights the count of important linguistic features identified over 20 runs. Darker points represent features more frequently identified as important across the 20 runs.

cortex, IP, and precuneus are key nodes in episodic memory networks [41–43], whereas the IT supports semantic representation and semantically guided lexical retrieval [44,45]. Tau pathology has been associated with impairments in memory retrieval and semantic processing in individuals with AD and MCI [46]. Entorhinal tau pathology has also been linked to memory decline in clinically normal older adults [47]. Because narrative production in the LMDR task involves retrieving story content, accessing semantically rich representations, and generating a coherent verbal response, the observed associations between regional tau burden and pause- and fluency-related features may reflect subtle variation in memory retrieval and speech production during narrative recall. However, the specific cognitive and neurobiological mechanisms underlying these patterns remain to be clarified.

In addition, our analyses revealed novel findings that have not been emphasized in prior work. First, uniqueIdeaDensity emerged as a sensitive and robust marker of tau-related changes. In particular, this feature was associated with tau burden in most brain regions examined (except for the IP), and was retained as a predictor of tau burden in both Lasso- and Boruta-based feature selection. Idea density reflects how efficiently recalled content is translated into coherent, information-bearing statements and may decline with reduced executive control and discourse-level organization [48]. Lower idea density has long been linked to AD risk [49], and studies using spontaneous speech transcripts have also reported lower idea density in AD/MC [8,50]. When shifting the target to AD biomarkers, we found that ideaDensity was associated with tau burden only in the amygdala, whereas uniqueIdeaDensity showed broader sensitivity to tau-related changes. Compared with idea density, unique idea density additionally captures the diversity of lexical items anchoring the propositions, reflecting the ability to access specific semantic representations and avoid generic or repetitive constructions. These findings suggest that speech-derived linguistic features that capture multiple dimensions of cognition may be more sensitive to early tau-related changes. With further validation in larger samples, including more older adults with amyloid positivity, these features may potentially serve as useful digital biomarkers for speech-based tools to screen for neurodegenerative changes, including those associated with AD.

Second, we identified region-specific associations between tau burden and linguistic features in story retelling. Because evidence on brain-region-specific contributions to memory retrieval and language production is still developing—and may vary across tasks and measurement approaches—we provide only brief interpretations of selected region-specific patterns based on the literature. For example, we found that higher tau burden in the precuneus was associated with reduced adjective density and diversity. Closer inspection of adjective use showed that participants with lower precuneus tau burden produced more descriptive modifiers when referring to story characters (e.g., *young*, *little*, or *hungry* children, *single* mother). Although the functional roles of the precuneus are still being refined, current evidence suggests that it is a key hub of the Default Mode Network (DMN) and supports processes such as self-referential cognition, mental imagery, and episodic memory [41]. The association between higher tau burden in precuneus and fewer descriptive adjectives may reflect subtle differences in the richness or specificity of story reconstruction during retelling. In addition, we observed that elevated tau burden in the IP and IT regions was associated with reduced diversity and/or density of prepositions used. Prepositions play a critical role in encoding relational

and spatiotemporal information, linking entities, actions, locations, and time within a narrative. Inferior/lateral temporal and IP cortices are key components of the distributed semantic memory system, supporting conceptual/semantic representations [51,52]. Inferior/lateral temporal regions are particularly implicated in lexical–semantic representations and meaning-related processing [51,52], whereas IP regions—especially the angular gyrus—support semantic integration and have been proposed to play a special role in representing event concepts (i.e., entities interacting across space and time) [51]. Aligned with this framework, a closer examination of preposition use showed that participants with higher tau burden in IP and IT more often omitted prepositional phrases and instead relied on shorter, fragmented clauses to describe a single event (e.g., "*She got robbed. And fifteen dollars were stolen.*" rather than "*She got robbed of fifteen dollars.*").

Third, we observed that simpler sentence structure (lower yngve score) was associated with higher tau burden in the IT region, and sensitivity analyses excluding off-topic sentences identified additional syntactic features associated with tau burden in the IP and precuneus. Although these regions are not core syntax areas, they support semantic integration, episodic retrieval, and internally guided cognition central to narrative speech [41,51,52]. Thus, in cognitively normal individuals, these associations may reflect early network-level effects on narrative planning and information integration rather than focal syntactic impairment. Future studies are warranted to clarify the underlying mechanisms.

Compared with the other regions examined, we found fewer associations between linguistic features and tau signals in the amygdala. The amygdala is known to be involved in processing affective and social aspects of communication—such as emotional salience, emotional tone, and prosody—rather than directly supporting lexical retrieval, semantic integration, or discourse-level planning [53–55]. Future work is needed to examine the relationship between affective speech markers and tau burden in this region.

Similar as previous work [14], we found that neither the LMDR score nor linguistic features were associated with Aβ status. This is consistent with the view of amyloid pathology as an upstream process that may be present before measurable cognitive and language changes emerge in largely asymptomatic individuals [56]. In addition, the small number of amyloid-positive individuals in our sample may have limited power to detect significant associations. Note that a picture-description study in cognitively unimpaired adults reported that higher amyloid burden was associated with reduced content-word use (e.g., a lower open-to-closed class word ratio) [16]. However, we did not observe significant associations between comparable features (e.g., content density) and amyloid positivity in our sample. One possible explanation is that the LMDR task differs from picture description and imposes tighter constraints on the content participants can produce during story retelling.

4.3. Two types of linguistic features

We employed two methods to generate features from POS tags and words. The first method does not account for repeated word use (e.g., adjectivesRate), which can inflate frequency without adding semantic information. In contrast, the second uses unique word counts (e.g., uniqueAdjectivesRate), which improves the signal-to-noise ratio by emphasizing semantic variability (i.e., distinctive word use).

We found that features derived using the second method were generally more sensitive to regional tau burden. For example, six features based on unique words per POS tag (e.g., uniqueConjunctionsRate) were associated with tau burden in one or more regions, compared with three features based on words per POS tag. In addition, uniqueIdeaDensity was associated with tau burden in four brain regions, whereas ideaDensity was associated with tau burden in only the amygdala. Similarly, features based on unique words (e.g., uniqueWordsPerSentence, uniqueSpeechRate) were associated with tau burden in more brain regions than their counterparts based on overall word count (wordsPerSentence, speech rate).

Taken together, our findings suggest that features capturing lexical diversity and semantic richness may better preserve behaviorally relevant signals linked to tau pathology.

4.4. Feature selection: LASSO vs. Boruta

Both LASSO and Boruta serve as feature selection methods for outcome prediction, but they operate differently.

LASSO shrinks the coefficients of weaker or correlated predictors toward zero, forcing some to become exactly zero. This makes it a powerful tool for selecting the most relevant predictors to improve the compactness and interpretability of the model. In this study, LASSO removed a few features selected by individual regression models (Panel A vs. Panel B, Fig. 3), most of which are highly correlated with the retained features. For example, words prefixed with "&-" were included in the calculation of filled pauses and were also classified as interjections. Consequently, the totalFilledPauses was highly correlated with the interjectionsRate (Cluster 1) and was removed by LASSO.

Unlike LASSO, Boruta retains all relevant predictors rather than enforcing sparsity. It identified more linguistic predictors of amyloid status and regional tau burden than the individual regression models (Panel C vs. Panel A, Fig. 3). For example, while the regression models found no linguistic features associated with amyloid status, Boruta selected predictors across multiple language dimensions (except Cluster 1), though most were selected infrequently (<10 runs).

Additionally, although LASSO removed redundant features, the retained predictors for each brain region still spanned multiple dimensions of language use. Boruta selected a broader set of features, suggesting that non-linear effects and interactions may add predictive value beyond what linear models capture. Future work should focus on identifying a compact feature set that balances predictive performance with model interpretability. Moreover, integrating linguistic features with other modalities, such as measurements on daily activities, changes in behaviors (e.g., sleep patterns, gaits), and electrophysiological measurements, may enhance the sensitivity and accuracy of automated methods for early AD detection.

4.5. Implications for AD clinical trials

This study identified several linguistic features associated with regional tau burden measured by PET imaging in cognitively normal adults. With further validation, these measures may support participant identification and selection for early-intervention AD trials. Such trials increasingly focus on preclinical or prodromal AD [57,58], but typically require confirmation of AD pathology using amyloid PET or CSF biomarkers, which are expensive and difficult to scale [57,59]. In this context, speech-derived linguistic measures could potentially serve as a low-burden screening or enrichment tool to prioritize individuals for confirmatory biomarker testing.

Their low-burden and scalable nature may also make them useful for repeated assessment, cognitive safety monitoring, and exploratory or secondary outcome measures. Early-intervention AD trials typically require sensitive cognitive or composite measures as primary outcomes to demonstrate clinically meaningful therapeutic effects [58,60,61]. However, detecting subtle cognitive changes remains a major challenge in these trials [61,62]. In this study, speech-derived linguistic measures detected subtle differences in language use associated with regional tau burden, even when conventional LMDR scores did not. These findings suggest that language-based measures may be more sensitive than traditional neuropsychological tests for detecting early disease-related changes. With further validation in larger samples and longitudinal datasets, these measures could serve as scalable repeated measures or secondary outcomes in AD clinical trials, particularly those targeting preclinical or prodromal AD populations. They may also support cognitive safety monitoring by helping detect unexpected or accelerated

changes in language and cognition during these trials.

Note that, although this study provides preliminary evidence on the potential use of speech-derived linguistic measures as markers of preclinical AD, additional validation studies, including both analytical and clinical validation, are needed to strengthen the evidence base supporting the incorporation of these linguistic measures into clinical trials or routine screening workflows. For example, to serve as preclinical AD biomarkers, these features should be clinically validated in larger samples of cognitively normal older adults, with greater representation of amyloid-positive participants and individuals from diverse sociodemographic backgrounds (e.g., education level, race/ethnicity). Analytical validation is also important for real-world applications, including assessment of whether the same features can be reliably derived from speech transcripts across software versions, recording devices, and study sites. Additional methodological development will also be important for scalability and real-world usability, such as validation of features extracted from automated transcriptions and from more naturalistic, open-ended speech settings. Finally, prospective longitudinal studies and multi-center validation will be important for determining whether these linguistic measures are sufficiently stable, sensitive, and clinically meaningful to serve as endpoints, enrichment tools, or monitoring biomarkers in AD clinical trials. Because speech data can contain identifiable information, clinical trials incorporating these measures would also need clear consent language and additional safeguards for data storage, processing, and sharing.

4.6. Limitations and future work

This study has several limitations. First, as an observational study, the associations we observed between regional tau burden and linguistic features do not imply causality. Our interpretations are therefore grounded in existing evidence from functional neuroanatomy and language research. Future external replications and biological validation studies will be important for corroborating these results and clarifying their underlying mechanisms. Second, the dataset we analyzed is relatively biased toward younger participants, with only 17 amyloid positive cases, which limited statistical power for analyses involving amyloid status and may affect the generalizability of the findings to older populations with AD-specific pathology. Therefore, the observed associations between speech-derived linguistic features and tau-related biomarkers should be interpreted as reflecting tau-related or neurodegenerative changes more broadly, rather than AD-specific pathology. Accordingly, our findings should be viewed as hypothesis-generating evidence for potential linguistic markers of preclinical AD, rather than as definitive evidence supporting their use as AD biomarkers. Future studies including older participants and larger numbers of amyloid-positive individuals are needed to clarify the associations between linguistic features and AD biomarker staging. Third, this study extracted linguistic features from manually transcribed voice recordings for analysis. While manual transcription ensures high quality of data, its scalability remains a challenge. Studies evaluating fully automated linguistic features (i.e. those extracted from automatic transcripts) against AD biomarkers are warranted to inform their future applications in early AD detection. Lastly, this study focused on linguistic features that can be automatically extracted from text using popular NLP methods, such as those related to POS tags and syntactic or dependency parsing. More complex features, such as self-corrections, repetitions, and changes in voice intonation during NP tests, were not included in this analysis. Future studies may develop methods to generate and evaluate these features for AD detection.

Future studies incorporating fluid-based AD biomarkers would also be an important extension of this work. Plasma-based biomarkers, such as p-tau217, p-tau181, and the Aβ42/Aβ40 ratio, have shown promise for detecting AD-related pathology and disease progression [63,64]. These biomarkers could serve as additional outcomes for assessing speech-derived linguistic features. Consistent associations with both imaging- and fluid-based biomarkers would further support the robustness and biological relevance of the language-derived measures identified in this study. Because plasma-based biomarkers are more accessible and scalable than PET imaging or CSF collection, they could also enable larger-scale validation and support future multimodal risk prediction frameworks.

5. Conclusion

This study examines linguistic features from constrained spontaneous speech as potential markers of preclinical AD. We identified 23 features—reflecting reduced lexical diversity, semantic richness, and sentence complexity, along with slower speech and more pauses—associated with regional tau burden across five brain regions. Future work should leverage more diverse datasets, expanded feature sets, and multimodal data to improve early AD detection.

Funding

This project was supported by funding from the National Institute on Aging (NIA) Artificial Intelligence and Technology Collaboratories (P30-AG073104, P30-AG073105), the National Institutes of Health (U19-AG068753, R01-AG062109, R01-HL159620, R01-AG083735, R01-NS142076, R00-AG071837, R03-DC023134), the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institute of Health and Boston University School of Medicine under Contract No. 75N92019D00031, Gates Ventures, the American Heart Association (20SFRN35460031), the Alzheimer's Drug Discovery Foundation (201902-2017835), and the Alzheimer's Association (AARFD-21-849349). HC was supported by the Boston University Bioinformatics Masters Summer Internship Program (BMSIP).

Data and code availability

The datasets analyzed in this study are available to qualified investigators through a formal research application to the Framingham Heart Study (https://www.framinghamheartstudy.org/fhs-for-researchers/). Population-level, de-identified outputs generated from the analyses are provided in Supplemental File 2. Computer code developed for feature derivation and data analysis is available through the GitHub repository: https://github.com/jchen-BUlab/Linguistic_markers_for_preclinical_AD/.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, OpenAI's ChatGPT was used to assist with language editing and text refinement. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Haochun Huang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Zexu Li:** Writing – review & editing, Methodology, Investigation. **Christina B. Young:** Writing – review & editing, Investigation. **Viktorija Smith:** Writing – review & editing, Investigation, Data curation. **Ting Fang Alvin Ang:** Writing – review & editing, Investigation, Data curation. **Vijaya B. Kolachalama:** Writing – review & editing, Investigation, Funding acquisition. **Rhoda Au:** Writing – review & editing, Investigation, Funding acquisition. **Jinying Chen:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests Christina B. Young reports a relationship with Medidata Solutions that includes: consulting or advisory. Rhoda Au reports a relationship with Signant Health that includes: consulting or advisory. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work used data from the Framingham Heart Study (National Heart, Lung, and Blood Institute; Boston University School of Medicine). We thank the Framingham Heart Study participants and staff for their invaluable contributions. We also thank the Data Curation and Management Core of the Framingham Heart Study Brain Aging Program (FHS-BAP) for technical support in data acquisition.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tjpad.2026.100689.

References

[1] Bature F, Guinn B-A, Pang D, Pappas Y. Signs and symptoms preceding the diagnosis of Alzheimer's disease: a systematic scoping review of literature from 1937 to 2016. BMJ Open 2017;7(8):e015746. https://doi.org/10.1136/bmjopen-2016-015746.
[2] Reeves SM, Williams V, Blacker D, Woods RL. Further evaluation of narrative description as a measure of cognitive function in Alzheimer's disease. Neuropsychology 2023;37(7):801–12. https://doi.org/10.1037/neu0000884.
[3] Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci 2015;7:195. https://doi.org/10.3389/fnagi.2015.00195.
[4] Kaltsa M, Tsolaki A, Lazarou I, Mittas I, Papageorgiou M, Papadopoulou D, et al. Language markers of dementia and their role in early diagnosis of Alzheimer's disease: exploring grammatical and syntactic competence via sentence repetition. J Alzheimer's Dis Rep 2024;8(1):1115–32. https://doi.org/10.3233/ADR-230204.
[5] Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inf Assoc 2020;27(11):1784–97. https://doi.org/10.1093/jamia/ocaa174.
[6] 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 Alzheimer 19s Dis 2020;78(4):1547–74. https://doi.org/10.3233/JAD-200888.
[7] Asgari M, Kaye J, Dodge H. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimer's Dement: Transl Res Clin Interv 2017;3(2):219–28. https://doi.org/10.1016/j.trci.2017.01.006.
[8] Roark B, Mitchell M, Hosom J-P, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans Audio Speech Lang Process 2011;19(7):2081–90. https://doi.org/10.1109/TASL.2011.2112351.
[9] Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calzà L. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? Front Aging Neurosci 2018;10:369. https://doi.org/10.3389/fnagi.2018.00369.
[10] Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I. Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Comput Speech Lang 2019;53:181–97.
[11] Thomas JA, Burkhardt HA, Chaudhry S, Ngo AD, Sharma S, Zhang L, et al. Assessing the utility of language and voice biomarkers to predict cognitive impairment in the framingham heart study cognitive aging cohort data. J Alzheimer's Dis 2020;76(3):905–22. https://doi.org/10.3233/JAD-190783.
[12] Nyongesa CA, Hogarth M, Pa J. Artificial intelligence-driven natural language processing for identifying linguistic patterns in Alzheimer's disease and mild cognitive impairment: a study of lexical, syntactic, and cohesive features of speech through picture description tasks. J Alzheimer 19s Dis 2025;106(1):120–38. https://doi.org/10.1177/13872877251339756.
[13] Cho S, Cousins KAQ, Shellikeri S, Ash S, Irwin DJ, Liberman MY, et al. Lexical and acoustic speech features relating to Alzheimer disease pathology. Neurology 2022;99(4):e313–e22. https://doi.org/10.1212/WNL.0000000000200581.
[14] Young CB, Smith V, Karjadi C, Grogan SM, Ang TFA, Insel PS, et al. Speech patterns during memory recall relates to early tau burden across adulthood. Alzheimers Dement 2024;20:2552–63. https://doi.org/10.1002/alz.13731.
[15] Verfaillie SC, Witteman J, Slot RE, Pruis IJ, Vermaat LE, Prins ND, et al. High amyloid burden is associated with fewer specific words during spontaneous speech in individuals with subjective cognitive decline. Neuropsychologia 2019;131:184–92. https://doi.org/10.1016/j.neuropsychologia.2019.05.006.
[16] Mueller KD, Van Hulle CA, Koscik RL, Jonaitis E, Peters CC, Betthauser TJ, et al. Amyloid beta associations with connected speech in cognitively unimpaired adults. Alzheimer's Dement: Diagn Assess Dis Monit 2021;13(1):e12203. https://doi.org/10.1002/dad2.12203.
[17] Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the framingham heart study. Nat Rev Cardiol 2019;16(11):687–98. https://doi.org/10.1038/s41569-019-0202-5.
[18] Au R, Piers RJ, Devine S. How technology is reshaping cognitive assessment: lessons from the framingham heart study. Neuropsychology 2017;31(8):846–61. https://doi.org/10.1037/neu0000411.
[19] Alhanai T, Au R, Glass J. Spoken language biomarkers for detecting cognitive impairment. 2017 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE; 2017.
[20] Tavabi N, Stück D, Signorini A, Karjadi C, Al Hanai T, Sandoval M, et al. Cognitive digital biomarkers from automated transcription of spoken language. J Prev Alzheimer's Dis 2022;9(4):791–800. https://doi.org/10.14283/jpad.2022.66.
[21] Project ULD. Transcription guide: utterance segmentation. 2025 [Available from, https://ldp-uchicago.github.io/docs/guides/transcription/sect_4.html.
[22] Yngve VH. A model and an hypothesis for language structure. Proc Am Philos Soc 1960;104(5):444–66.
[23] Frazier L. Syntactic complexity. Natural language parsing: psychological, computational, and theoretical perspectives. 1985. p. 129–89.
[24] Roark B, Mitchell M, Hollingshead K. editors. Syntactic complexity measures for detecting mild cognitive impairment. Biological, translational, and clinical language processing. 2007.
[25] Cheung H, Kemper S. Competing complexity metrics and adults' production of complex sentences. Appl Psycholinguist 1992;13(1):53–76.
[26] Liu H. Dependency distance as a metric of language comprehension difficulty. J Cogn Sci 2008;9(2):159–91.
[27] Agmon G, Pradhan S, Ash S, Nevler N, Liberman M, Grossman M, et al. Automated measures of syntactic complexity in natural speech production: older and younger adults as a case study. J Speech Lang Hear Res 2024;67(2):545–61. https://doi.org/10.1044/2023_JSLHR-23-00009.
[28] Wang L, Liu H. Syntactic variations in Chinese–english code-switching. Lingua 2013;123:58–73. https://doi.org/10.1016/j.lingua.2012.10.003.
[29] Gao N, He Q. A dependency distance approach to the syntactic complexity variation in the connected speech of Alzheimer's disease. Humanit Soc Sci Commun 2024;11(1):1–12. https://doi.org/10.1057/s41599-024-03509-0.
[30] Sand Aronsson F, Kuhlmann M, Jelic V, Östberg P. Is cognitive impairment associated with reduced syntactic complexity in writing? evidence from automated text analysis. Aphasiology 2021;35(7):900–13.
[31] Rezaii N, Mahowald K, Ryskin R, Dickerson B, Gibson E. A syntax–lexicon trade-off in language production. Proc Natl Acad Sci 2022;119(25):e2120203119. https://doi.org/10.1073/pnas.2120203119.
[32] McGillycuddy M, Popovic G, Bolker BM, Warton DI. Parsimoniously fitting large multivariate random effects in glmmTMB. J Stat Softw 2025;112:1–19. https://doi.org/10.18637/jss.v112.i01.
[33] R Core Team. R: a language and environment for statistical computing. 2023.
[34] Venables WN, Ripley BD. Modern applied statistics with s. Springer Science & Business Media; 2013.
[35] Arel-Bundock V, Greifer N, Heiss A. How to interpret statistical models using marginaleffects for R and Python. J Stat Softw 2024;111:1–32.
[36] Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1–22.
[37] Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 2010;36:1–13.
[38] Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol 1991;82(4):239–59. https://doi.org/10.1007/BF00308809.
[39] Berron D, Vogel JW, Insel PS, Pereira JB, Xie L, Wisse LEM, et al. Early stages of tau pathology and its associations with functional connectivity, atrophy and memory. Brain 2021;144(9):2771–83. https://doi.org/10.1093/brain/awab114.
[40] Therriault J, Pascoal TA, Lussier FZ, Tissot C, Chamoun M, Bezgin G, et al. Biomarker modeling of Alzheimer's disease using PET-based Braak staging. Nat Aging 2022;2(6):526–35. https://doi.org/10.1038/s43587-022-00204-0.
[41] Dadario NB, Sughrue ME. The functional role of the precuneus. Brain 2023;146(9):3598–607. https://doi.org/10.1093/brain/awad181.
[42] Cabeza R, Ciaramelli E, Olson IR, Moscovitch M. The parietal cortex and episodic memory: an attentional account. Nat Rev Neurosci 2008;9(8):613–25. https://doi.org/10.1038/nrn2459.
[43] Maass A, Lockhart SN, Harrison TM, Bell RK, Mellinger T, Swinnerton K, et al. Entorhinal tau pathology, episodic memory decline, and neurodegeneration in aging. J Neurosci 2018;38(3):530–43. https://doi.org/10.1523/JNEUROSCI.2028-17.2017.
[44] Shimotake A, Matsumoto R, Ueno T, Kunieda T, Saito S, Hoffman P, et al. Direct exploration of the role of the ventral anterior temporal lobe in semantic memory: cortical stimulation and local field potential evidence from subdural grid electrodes. Cereb Cortex 2015;25(10):3802–17. https://doi.org/10.1093/cercor/bhu262.
[45] Antonucci SM, Beeson PM, Labiner DM, Rapcsak SZ. Lexical retrieval and semantic knowledge in patients with left inferior temporal lobe lesions. Aphasiology 2008;22(3):281–304. https://doi.org/10.1080/02687030701294491.

[46] Bejanin A, Schonhaut DR, La Joie R, Kramer JH, Baker SL, Sosa N, et al. Tau pathology and neurodegeneration contribute to cognitive impairment in Alzheimer's disease. Brain 2017;140(12):3286–300. https://doi.org/10.1093/brain/awx243.
[47] Berron D, Vogel JW, Insel PS, Pereira JB, Xie L, Wisse LE, et al. Early stages of tau pathology and its associations with functional connectivity, atrophy and memory. Brain 2021;144(9):2771–83. https://doi.org/10.1093/brain/awab114.
[48] Harrag C, Sabil A, Conceição MC, Radvansky GA. Propositional density: cognitive impairment and aging. Front Psychol 2024;15:1434506. https://doi.org/10.3389/fpsyg.2024.1434506.
[49] Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, Markesbery WR. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. Jama 1996;275(7):528. https://doi.org/10.1001/jama.1996.03530310034029.
[50] Sirts K, Piguet O, Johnson M. Idea density for predicting Alzheimer's disease from transcribed speech. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017); 2017.
[51] Binder JR, Desai RH. The neurobiology of semantic memory. Trends Cogn Sci 2011;15(11):527–36. https://doi.org/10.1016/j.tics.2011.10.001.
[52] Friederici AD. The brain basis of language processing: from structure to function. Physiol Rev 2011;91(4):1357–92. https://doi.org/10.1152/physrev.00006.2011.
[53] Du Y, Zhang S, Qiu Q, Fang Y, Zhao L, Yue L, et al. The mediating effect of the amygdala-frontal circuit on the association between depressive symptoms and cognitive function in Alzheimer's disease. Transl Psychiatry 2024;14(1):301. https://doi.org/10.1038/s41398-024-03026-3.
[54] Fecteau S, Belin P, Joanette Y, Armony JL. Amygdala responses to nonlinguistic emotional vocalizations. Neuroimage 2007;36(2):480–7. https://doi.org/10.1016/j.neuroimage.2007.02.043.
[55] Liebenthal E, Silbersweig DA, Stern E. The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. Front Neurosci 2016;10:506. https://doi.org/10.3389/fnins.2016.00506.
[56] Aizenstein HJ, Nebes RD, Saxton JA, Price JC, Mathis CA, Tsopelas ND, et al. Frequent amyloid deposition without significant cognitive impairment among the elderly. Arch Neurol 2008;65(11):1509. https://doi.org/10.1001/archneur.65.11.1509.
[57] Watson JL, Ryan L, Silverberg N, Cahan V, Bernard MA. Obstacles and opportunities in Alzheimer's clinical trial recruitment. Health Aff 2014;33(4):574–9. https://doi.org/10.1377/hlthaff.2013.1314.
[58] Jutten RJ, Papp KV, Hendrix S, Ellison N, Langbaum JB, Donohue MC, et al. Why a clinical trial is as good as its outcome measure: a framework for the selection and use of cognitive outcome measures for clinical trials of Alzheimer's disease. Alzheimer's Dement 2023;19(2):708–20. https://doi.org/10.1002/alz.12773.
[59] Yigamawano FK, Odom AR, Xue C, Pandey HK, Kolachalama VB. AI-augmented frameworks for enhancing Alzheimer's disease clinical trials: a memory clinic perspective. J Prev Alzheimer's Dis 2026;13(1):100396.
[60] Cohen S, Cummings J, Knox S, Potashman M, Harrison J. Clinical trial endpoints and their clinical meaningfulness in early stages of Alzheimer's disease. J Prev Alzheimer's Dis 2022;9(3):507–22. https://doi.org/10.14283/jpad.2022.41.
[61] Aisen P. Cognitive/clinical endpoints for pre-dementia AD trials. J Prev Alzheimer's Dis 2015;2(2):1–3. https://doi.org/10.14283/jpad.2015.62.
[62] Petersen RC, Aisen PS, Andrews JS, Atri A, Matthews BR, Rentz DM, et al. Expectations and clinical meaningfulness of randomized controlled trials. Alzheimer's Dement 2023;19(6):2730–6. https://doi.org/10.1002/alz.12959.
[63] Ashton NJ, Brum WS, Di Molfetta G, Benedet AL, Arslan B, Jonaitis E, et al. Diagnostic accuracy of a plasma phosphorylated tau 217 immunoassay for Alzheimer disease pathology. JAMA Neurol 2024;81(3):255. https://doi.org/10.1001/jamaneurol.2023.5319.
[64] Hansson O, Edelmayer RM, Boxer AL, Carrillo MC, Mielke MM, Rabinovici GD, et al. The Alzheimer's Association appropriate use recommendations for blood biomarkers in Alzheimer's disease. Alzheimer's Dement 2022;18(12):2669–86. https://doi.org/10.1002/alz.12756.