Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: https://www.journals.elsevier.com/eclinicalmedicine

Linguistic markers predict onset of Alzheimer's disease

Elif Eyigoz[a,*], Sachin Mathur[b], Mar Santamaria[b], Guillermo Cecchi[a,*], Melissa Naylor[b,*]

[a] *IBM Thomas J. Watson Research Center, IBM Research, Yorktown Heights, NY 10598, United States*
[b] *Pfizer Worldwide Research and Development, Cambridge, MA 02139, United States*

ARTICLE INFO

Article History:
Received 27 April 2020
Revised 19 September 2020
Accepted 22 September 2020
Available online xxx

ABSTRACT

Background: The aim of this study is to use classification methods to predict future onset of Alzheimer's disease in cognitively normal subjects through automated linguistic analysis.
Methods: To study linguistic performance as an early biomarker of AD, we performed predictive modeling of future diagnosis of AD from a cognitively normal baseline of Framingham Heart Study participants. The linguistic variables were derived from written responses to the cookie-theft picture-description task. We compared the predictive performance of linguistic variables with clinical and neuropsychological variables. The study included 703 samples from 270 participants out of which a dataset consisting of a single sample from 80 participants was held out for testing. Half of the participants in the test set developed AD symptoms before 85 years old, while the other half did not. All samples in the test set were collected during the cognitively normal period (before MCI). The mean time to diagnosis of mild AD was 7.59 years.
Findings: Significant predictive power was obtained, with AUC of 0.74 and accuracy of 0.70 when using linguistic variables. The linguistic variables most relevant for predicting onset of AD have been identified in the literature as associated with cognitive decline in dementia.
Interpretation: The results suggest that language performance in naturalistic probes expose subtle early sings of progression to AD in advance of clinical diagnosis of impairment.
Funding: Pfizer, Inc. provided funding to obtain data from the Framingham Heart Study Consortium, and to support the involvement of IBM Research in the initial phase of the study.

1. Introduction

A key priority in Alzheimer's disease (AD) research is the identification of early intervention strategies that will decrease the risk, delay the onset, or slow the progression of disease. Early interventions can only be effectively tested and implemented if the population that stands to benefit can be identified. While many variables have been associated with risk of AD, there is still a great need for the development of cheap, reliable biomarkers of preclinical AD. Aging-related cognitive decline manifests itself in almost all aspects of language comprehension and production. Even seemingly mundane linguistic abilities, such as object naming, engage extensive brain networks [1]. As a result, these linguistic abilities can easily be disrupted, which makes language competence a sensitive indicator of mental dysfunction. The influential Nun Study [2] provided initial evidence of a correlation between lower linguistic performance early in life and higher incidence of cognitive decline and conversion rates to AD late in life.

The aim of this study is to test to what extent linguistic performance at a single time point can be utilized as a prognostic marker of conversion to AD. We used data from the Framingham Heart Study (FHS) [3], a large cohort longitudinal study spanning several decades. As a part of FHS, qualifying participants were administered a neuropsychological (NP) test battery in successive visits [4–6], which included the cookie-theft picture description task (CTT) from the Boston Aphasia Diagnostic Examination [7]. Picture description tasks are used to assess discourse in subjects with disorders such as aphasia and dementia, and CTT has become the most frequently used picture description task in clinical settings [8]. We applied computational techniques to extract linguistic variables from written responses to the CTT and compared their prognostic value with that of more traditional clinical variables that could easily be obtained in the screening period of a clinical trial, including NP test scores, demographic and genetic information, and medical history. Using the variables obtained when the participants were assessed *to be cognitively normal*, we developed models to *predict* whether or not a particular participant will *develop MCI due to AD on or before 85 years old*.

Our work significantly differs from the current literature on predicting future onset of AD in the following ways: First, our prediction

[*] Corresponding authors.
E-mail addresses: ekeyigoz@us.ibm.com (E. Eyigoz), gcecchi@us.ibm.com (G. Cecchi), melissa.naylor@takeda.com (M. Naylor).

https://doi.org/10.1016/j.eclinm.2020.100583

is based on data collected while the participants were cognitively healthy. Second, we focus exclusively on variables readily attainable as part of the screening phase of an early-intervention trial and assess predictive performance using *only* linguistic metrics derived from a *single* administration of the Cookie Theft Task, a relatively simple and naturalistic language probe. Third, we utilized a machine learning approach to deal with a multivariate representation of linguistic performance. Finally, we compare the predictive ability of language features with that of more traditional variables associated with identification of high risk for AD, e.g., for inclusion in a clinical trial of potentially disease-modifying therapy).

2. Methods

2.1. Cognitive assessment in the Framingham heart study

The FHS is a well-documented, community-based cohort study initiated in 1948, with the purpose of longitudinal monitoring of participants' health [3,9]. Cognitive status monitoring of the original cohort began in 1975, and since 1981 the participants' cognitive status has been assessed with the Mini–Mental State Examination (MMSE) [10] at examinations taking place every 4 years [4,11]. Participants in the offspring cohort have undergone MMSEs since 1991, and have undergone NP examinations every 5 or 6 years since 1999 [6]. Annual neurologic and neuropsychological examinations were performed when cognitive decline was reported by a family member of the participant, upon referral by a physician or by the investigators of the FHS, or after review of the participant's medical records [11]. Cognitive status monitoring of the participants was reviewed by the Institutional Review Board of Boston University, and informed consent was obtained from the participants.

The neuropsychological test battery resulted in a dementia rating, which represents the impression of the examiner who administered the test battery [11]. The test battery included the cookie-theft picture description task (CTT) from the Boston Aphasia Diagnostic Examination, in which participants were asked to write down the description of the cookie-theft picture. As highlighted above, picture description tasks are commonly used to assess discourse in subjects with disorders such as aphasia and dementia, and, given its sensitivity to cognitive impairments, CTT has become the most frequently used picture description task in clinical settings [8]. The FHS study participants who qualified for inclusion in our study were among the oldest participants of the FHS, mostly from the original cohort, which was limited in its representativeness of the wider population [3].

A dementia-review panel with at least one neurologist and one neuropsychologist reviewed possible cognitive decline and dementia cases documented in the FHS [12,13]. Diagnosis of dementia was based on criteria from DSM-IV [14], and diagnosis of Alzheimer's disease was based on criteria from NINCDS–ADRDA [15,16].

2.2. Predictive modeling approach

To fit predictive models of future diagnosis of AD, we had to determine which participants to label as *cases* and which to label as *controls*. FHS participants varied in terms of whether their data was comprehensively reviewed by a panel of experts to determine dementia and AD status. We first identified a clinically defined test set by using these dementia reviews to label cases, selecting one CTT sample from each case, and matching it to a CTT sample collected in a control of the approximately the same age, gender and level of education. Because the FHS data available to us included a dementia review for only 39% of participants, only 80 of the participants qualified for inclusion in this test set. This left a very large number of participants unused. While most of the participants did not have dementia review data allowing for definitive labeling of cases, a dementia rating was available for the majority of the administrations of the neuropsychological test-battery. Using these dementia ratings, we used additional participants to create a training set. In semi-supervised learning terminology, the clinical dementia-review provided the *ground-truth* labels of the test data, whereas the dementia ratings provided the *weak* labels of the training data. This weakly-labeled training set was used *only* for machine learning *training*.

We validated predictive models in two ways: the *hold-out* method and the *cross-validation* method. For the hold-out method, we made use of the weakly-labeled training data by fitting the model to weakly labeled training data and then validating it on the held-out ground-truth test data. For the cross-validation method we implemented 20-fold cross validation on the test data (see the Supplementary Material for details).

2.3. Selection of participants and samples

In this study, the *onset of AD* was defined as the onset of mild cognitive impairment (MCI) in a participant who later received a diagnosis of AD. MCI is a heterogeneous condition; however, for those MCI patients who eventually convert to AD, MCI is considered by many to represent early-stage AD [17–19]. AD patients who developed MCI on or before age 85 (denoted as $\leq$ 85) were defined as *cases*.

We defined the *normal-aging* group as the participants who were recorded to be dementia-free on or after age 85 ($\geq$ 85). The control group was defined as the combination of the normal-aging group and AD patients whose onset of cognitive impairment was after 85 (>85) years old, as depicted in Fig. 1. According to this definition, all cases have already developed cognitive impairment due to AD at 85, and none of the controls have developed cognitive impairment due to AD at 85. Age 85 was chosen as a threshold, because this threshold was the optimum age to provide the largest balanced test set from the FHS data that was available to this study. As the age threshold increases, less participants qualify to be controls, and more participants qualify to be cases. Conversely, as the age threshold decreases, more participants qualify to be controls, and less participants qualify to be cases. In addition to providing the largest test set from FHS, age 85 has been widely used as a threshold to define oldest-old in AD studies [20]. It has been suggested that very-late onset AD (VLOAD), as defined by AD onset after the second half of the ninth decade differs from earlier-onset AD with respect to genetic and environmental patterns: Genetic risk factors for AD are more influential at relatively earlier ages with decreasing influence as age increases; while environmental factors may play a larger role in developing VLOAD [20–22].

The test data set included only *one sample per participant*, and they were matched to a control sample using age (+/- 2 years), gender, and education. As our purpose was to predict conversion in cognitively normal subjects, we included only samples collected *prior to any cognitive impairment* onset. Samples from participants who did not meet criteria for inclusion as ground-truth were used for training; see the Supplementary Material for the details. Fig. 2 shows a diagram summarizing the selection of participants and samples for the test set and the weakly-labeled training set. The demographics of participants in the test and training data sets can be seen in Table 1. For the test cases, the mean time to diagnosis with mild AD from cognitive normality was 7.59 years with standard deviation of 4.91, and the mean time to cognitive impairment onset from cognitive normality was 3.93 years with standard deviation of 3.69.

2.4. Psycholinguistic analyses

In this section, we provide an overview of the psycholinguistic analyses that were performed automatically for this study. See the Supplementary Material for the variables computed for the analyses presented in this section.

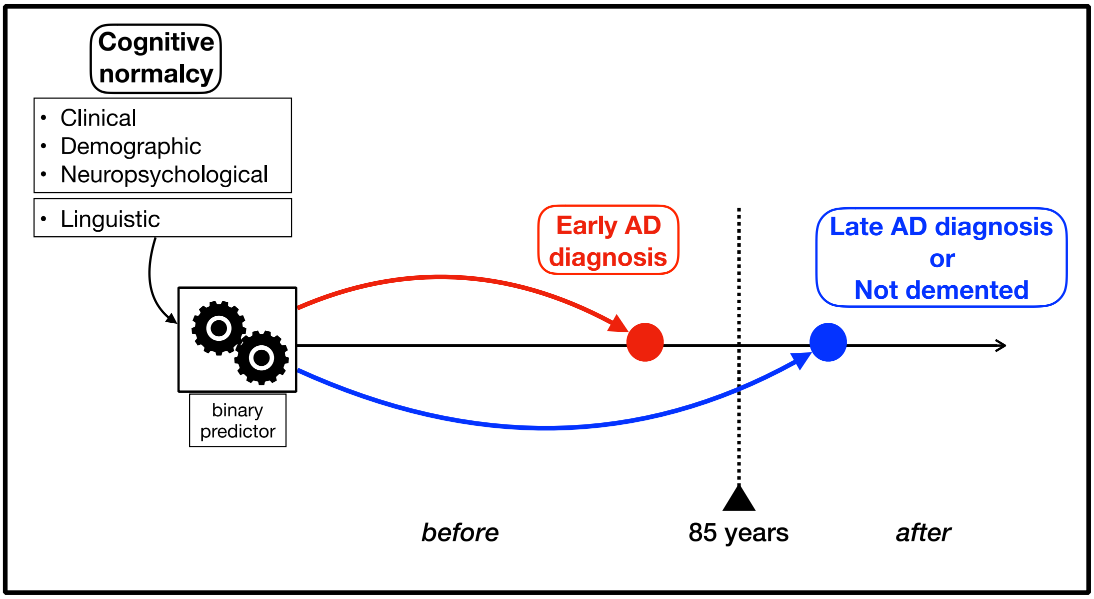


Fig. 1. The diagram depicts method for selection of cases vs controls and predictive model setting. Participants who developed MCI due to AD on or before age 85 were selected as cases, and participants remained dementia-free until age 85 were selected as controls. The three predictive models included only non-linguistic variables, only linguistic variables, or both (see Table 3), collected when participants were considered cognitively normal, and were trained to predict conversion status by age 85 vs. later or no conversion.

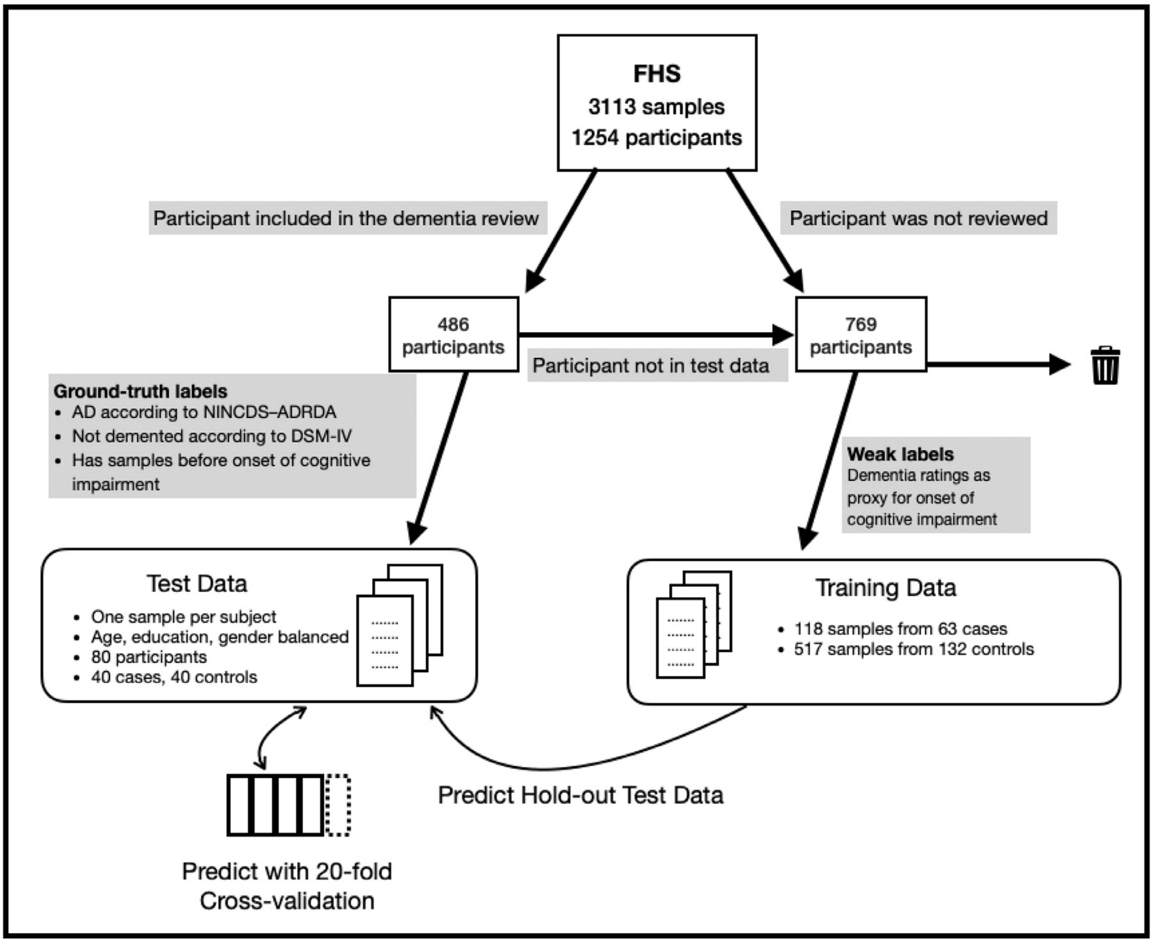
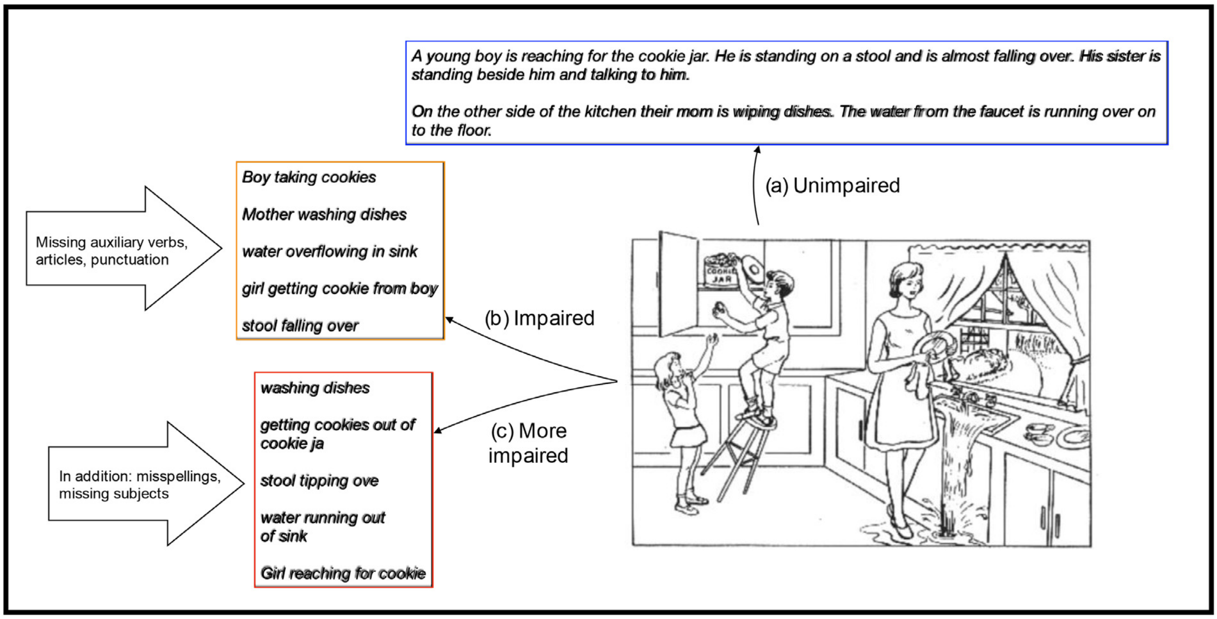

FHS
3113 samples
1254 participants
Participant included in the dementia review
Participant was not reviewed
486 participants
Participant not in test data
769 participants
Ground-truth labels
AD according to NINCDS–ADRDA
Not demented according to DSM-IV
Has samples before onset of cognitive impairment
Weak labels
Dementia ratings as proxy for onset of cognitive impairment
Test Data
One sample per subject
Age, education, gender balanced
80 participants
40 cases, 40 controls
Training Data
118 samples from 63 cases
517 samples from 132 controls
Predict Hold-out Test Data
Predict with 20-fold Cross-validation

Fig. 2. Method of selection of participants for creating a test data set and a training data set from the FHS data. The available data consisted of 3113 samples from 1254 participants, 486 of which have been reviewed by a panel for dementia status. The participants who were reviewed by the panel were candidates for creating a test data set. Their samples were eliminated according to the inclusion criteria, and then the qualifying samples were passed through age, education and gender matching. This resulted in a test set of 80 samples. The participants who were not reviewed by the dementia review panel were used for creating a larger weakly-labeled data set, only for the purpose of machine learning training. Validation of predictive modeling consisted of the hold-out method (train on weak-labels, test on ground-truth), and cross-validation (train on ground-truth, test on ground-truth).
.


A young boy is reaching for the cookie jar. He is standing on a stool and is almost falling over. His sister is standing beside him and talking to him.

On the other side of the kitchen their mom is wiping dishes. The water from the faucet is running over on to the floor.

Boy taking cookies
Mother washing dishes
water overflowing in sink
girl getting cookie from boy
stool falling over

washing dishes
getting cookies out of cookie ja
stool tipping ove
water running out of sink
Girl reaching for cookie

(a) Unimpaired
(b) Impaired
(c) More impaired
Missing auxiliary verbs, articles, punctuation
In addition: misspellings, missing subjects
COOKIE JAR

Fig. 3. CTT examples from FHS, including an unimpaired sample (a), an impaired sample showing telegraphic speech and lack of punctuation (b), and an even more impaired sample showing in addition significant misspellings and minimal grammatic complexity, e.g. lack of subjects (c).

Table 1
Age demographic and the education level of the ground-truth labeled and weakly-labeled data sets. The number of samples and the number of participants are the same in the ground-truth labeled set, as only one sample per participant was included.

		Ground-truth labels		Weak labels		
		Age (mean +/- SD)	Participants/			
Samples	Age (mean +/- SD)	Samples	Participants			
Control	Female	78.86 ± 6.01	22	84.34 ± 5.18	326	86
	Male	79.0 ± 4.39	18	83.72 ± 4.99	191	46
Case	Female	78.45 ± 5.36	22	71.79 ± 5.1	61	29
	Male	79.17 ± 4.29	18	73.76 ± 5.43	45	29
Control	No college	78.74 ± 5.89	23	83.83 ± 5.14	204	60
	College	79.18 ± 4.48	17	84.36 ± 5.08	313	72
Case	No college	78.22 ± 5.19	23	72.59 ± 5.71	22	16
	College	79.53 ± 4.42	17	72.63 ± 5.23	84	42
Total		80	80		623	190

Verbosity, lexical richness, and repetitiveness was assessed by using metrics such as number of words, number of unique words, and frequencies of repetitions (Fig. 3). Misspellings, use of punctuation, and uppercasing were analyzed to assess writing performance and style. Language-modeling analyses were performed to model the distributions of word sequences. Syntactic complexity was assessed through analysis of parse trees. Semantic content was assessed through analysis of participants' mention of information content units. Finally, propositional idea density analysis was used to quantify syntactic and semantic complexity.

2.5. Non-linguistic variables

The non-linguistic variables are age, gender, education (dichotomized as college degree vs. no college degree, and high-school vs. no high-school degree), number of *APOE* ε4 alleles, two binary indicator variables capturing evidence of hypertension or diabetes, and variables resulting from the NP tests. The NP tests used in this study include assessment of visuospatial and executive reasoning, object naming, memory, attention, abstraction, and language skills. A total of 13 NP tests, as listed in Table 2, resulting in 32 NP variables (see Supplementary Table 6) were used in this study. Consequently, the comprehensiveness of neuropsychological assessment used in this study surpass concise assessments, such as the Montreal cognitive assessment MoCA [23]. The clinical measures MMSE and the dementia ratings were not included in the models, as all samples in the ground-truth labeled test set, for both controls and cases, were collected during the periods of cognitive normality and had no significant variance.

2.6. Variable selection and training of predictive models

In total, 87 linguistic variables were computed (see Supplementary Table 6). Two NP test scores were excluded, because they were missing for more than half of the samples, leaving 31 NP test scores, three clinical and two demographic variables as non-linguistic variables. Before training predictive models, variable selection was performed strictly *on the training data* by using a univariate test between the preclinical AD cases and the control groups for each variable and eliminating variables that were not statistically significant ($p >= 0.05$). The *t*-test was used in the cross-validation experiments and the Wilcoxon signed rank test was used in the hold-out experiments. The use of different univariate tests for different experiment conditions was justified due to difference in data size and the noise in the weak labels. See Supplementary Material for details of the variable selection for the hold-out and the cross-validation methods. For

Please cite this article as: E. Eyigoz et al., Linguistic markers predict onset of Alzheimer's disease, EClinicalMedicine (2020), https://doi.org/10.1016/j.eclinm.2020.100583

Table 2
Neuropsychological tests (NP) included in the predictive models along with clinical and demographic variables, separately and in conjunction with linguistic variables. See Supplementary Table 1 for the full list of NP variables obtained through these NP tests.

Cognitive Domain	Description
Word retrieval	Boston Naming Test
Learning	The paired associate learning subtest from the Wechsler Memory Scale (WMS)
Attention and concentration	Wechsler Adult Intelligence Scale (WAIS) score for digit span
Verbal memory	The logical memory subtest from the Wechsler Memory Scale (WMS)
Premorbid intelligence	The reading subset of the Wide range achievement test WRAT-3
Verbal ability and executive control	Verbal fluency
Attention and concentration	Trail making tests A and B
Abstract reasoning	The similarities test from Wechsler Adult Intelligence Scale (WAIS)
Visuoperceptual organization	Hooper Visual Organization Test
Visual memory	The visual reproduction subtest from the Wechsler Memory Scale (WMS-R)
Verbal Comprehension	The information subset of the Wechsler Adult Intelligence Scale (WAIS-R)
Spatial visualization	Block design test
Psychomotor speed	Finger tapping test

training predictive models, we experimented with linear SVM, logistic regression and Naïve Bayes classifiers. The hyperparameters of the classifiers were set using nested cross-validation.

2.7. Longitudinal analysis of linguistic and NP variables

To identify possible longitudinal trends present in our multi-dimensional assessment of cognitive status, we implemented a factorization analysis, using all available samples from each eligible participant taking into account the correlational structure between both linguistic variables and the NP test scores. For this, we used the cases and the normal-aging participants who have a record of cognitive impairment onset. We aligned their samples temporally by their cognitive impairment date. The frequency of administration of the NP exams varied across participants and was on average 2.2 years. In order to normalize this variance across participants, we created synthetic samples by linear interpolation with a frequency of six months. We then used Nonnegative Matrix Factorization (NMF) on the up-sampled dataset. To compare the progression of cases and the normal-aging group, we projected the latter onto the factors learned for the former. The projections of each sample on the first factor were computed, and then averaged over all samples in each six-month interval to obtain the loading of each interval. For this analysis, we used all NP variables, and linguistic variables that were statistically significant on the test set with *t*-test, that were statistically significant on the training set with Wilcoxon signed rank test, and linguistic variables that were statistically significant with the Cox proportional-hazards model analysis, which is described in the following section.

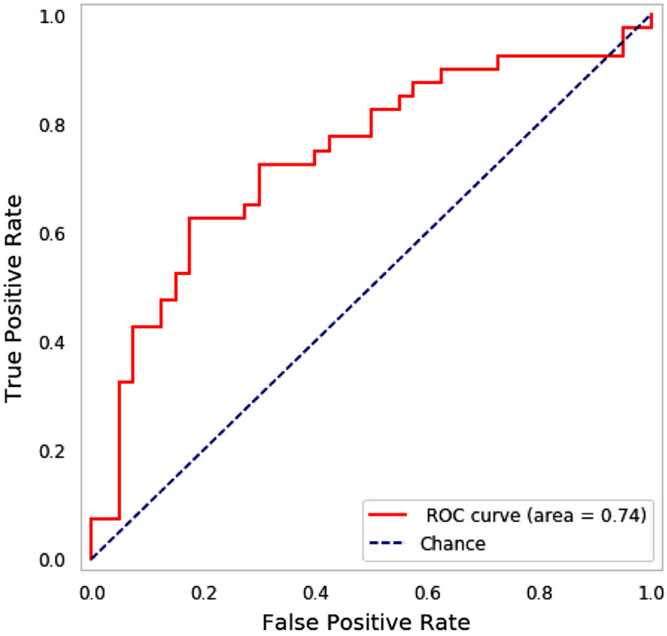


Fig. 4. The ROC curve of the test-set with the hold-out method for the linguistic-based model (see Table 3). This result was obtained by a Logistic Regression classifier.

2.8. Analysis of time to diagnosis with mild AD

To assess whether linguistic variables associated with the *time-to-diagnosis* with mild AD, we used Cox proportional-hazards models. Date of mild AD diagnosis was obtained from the dementia review, and the participants who were recorded as dementia free in their dementia review were censored. If a censored participant was alive at the date of the review, then the review date was used as the censor date. If the participant was no longer alive at the date of the review, then the oldest age the participant is known to be not demented was used as the censor date. Models for each single linguistic variable included as additional covariates age, gender, and education (i.e., college degree vs. no college degree.)

2.9. Role of funding source

Pfizer, Inc. provided funding to obtain the data from the Framingham Heart Study Consortium, and to support the involvement of IBM Research in the initial phase of the study. The funder had no role in data analysis or interpretation, which were the responsibility of the contributing authors.

3. Results

Univariate tests of individual variables between cases and controls showed that future onset of AD was associated with telegraphic speech, repetitiveness, and misspellings (see Supplementary Table 3). Telegraphic speech, as exemplified in Fig. 3, is a common symptom of non-fluent aphasia. In telegraphic speech, grammatical structure is reduced or absent, such that language contains simplified phrases consisting mainly of content words, with morphology and function words largely missing [24,25]. As shown in the examples from cognitively impaired participants in Fig. 3, telegraphic speech is not only simpler in grammatical structure, but also marked by lack of determiners ('the', 'a'), auxiliaries ('is', 'are') and entire subjects. Furthermore, samples from impaired participants further demonstrate misspellings and lack of punctuation.

Prediction performance in each experimental setting obtained by the best performing classifier are shown in Table 3. The plots showing the separation of the test and the training datasets by the best

Table 3
Results of prediction experiments for the three models. AUC stands for the area-under-the-curve statistic. Accuracy is ratio of correctly predicted samples to the total number of samples. Positive predictive value is the ratio of correctly predicted positive samples to the total predicted positive samples. Sensitivity is the ratio of correctly predicted positive samples to the all observations in the patient class. All metrics for each experimental setting were obtained by the same classifier. The best performing classifiers in the hold-out experiments were Logistic Regression in the linguistic and non-linguistic settings, and Naïve Bayes with the combination of linguistic and non-linguistic features. The best performing classifiers in the CV-experiments were Logistic Regression in the linguistic settings, and Naïve Bayes in the non-linguistic setting and the combination of linguistic and non-linguistic features.

		CV method	Hold-out method
	Best classifier	Logistic Regression	Logistic Regression
Linguistic variables from single CTT samples during cognitive normalcy	Accuracy	0.65	0.70
	AUC	0.73	0.74
	Positive predictive value	0.64	0.74
	Sensitivity	0.67	0.62
	Best classifier	Naïve Bayes	Logistic Regression
Non-linguistic variables (age, gender, education, *APOE*, hypertension, diabetes, NP)	Accuracy	0.60	0.59
	AUC	0.64	0.60
	Positive predictive value	0.64	0.61
	Sensitivity	0.44	0.48
	Best classifier	Naïve Bayes	Naïve Bayes
Aggregation of linguistic and non-linguistic variables	Accuracy	0.67	0.69
	AUC	0.72	0.67
	Positive predictive value	0.81	0.71
	Sensitivity	0.44	0.62

performing classifier reported in Table 3 can be seen in Supplementary Figure 2. In Table 3, all metrics for each experimental setting were obtained by the same classifier. The results obtained by other classifiers that are not reported in Table 3 can be found in the Supplementary Table 2.

Supplementary Table 4 shows the weights assigned to the linguistic variables by the best performing classifier reported in Table 3 in the hold-out method. Similarly, Supplementary Table 5 shows the weights assigned to the non-linguistic variables by the best performing classifier reported in Table 3 in the hold-out method. We performed a step-wise classification analysis by ranking the variables with respect to the weights assigned to the them by the best performing classifier, and by incrementally adding variables for classification until all variables with p-value < 0.05 were exhausted. Supplementary Figure 4 shows that the highest AUC of 0.76 was obtained with using the highest ranked 10 linguistic variables, which can be found in Supplementary Table 4.

In order to assess statistical significance, we computed a null distribution of AUCs for chance classification outcomes, and applied z-statistics to estimate the probability of the AUCs obtained by the predictive models [26]. The z-score indicated that AUC of 0.74 (see Fig. 4 for the ROC curve) corresponds to a 4.26-fold increase in

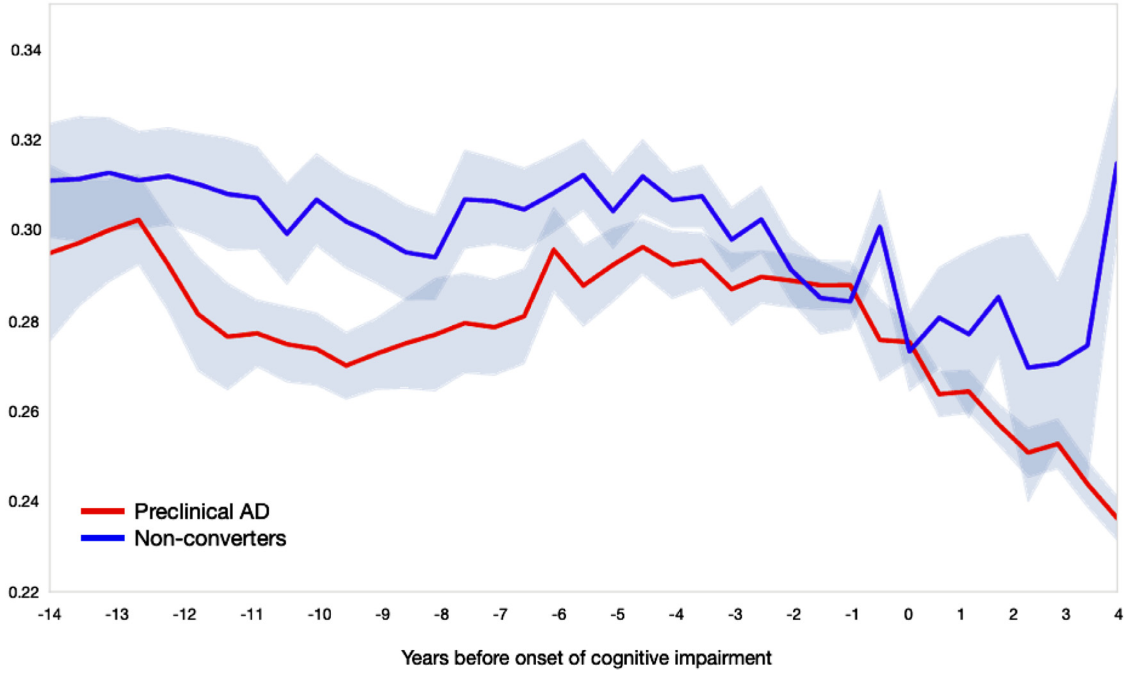


Fig. 5. The results of the non-negative matrix factorization (NMF) analysis of the linguistic and the NP variables on longitudinal data. This plot demonstrates that the factorization of the variables without using time information temporal trend as well as a differentiation between cases and controls, which starts several years before cognitive impairment. The controls' samples are projected onto the factorization learned from the cases' samples and averaged over six-month intervals. Controls are shown in blue, and cases in red. The horizontal axis is years to/from cognitive impairment onset, where 0 stands for the date of cognitive impairment. .

Table 4
Results of the Cox proportional hazards models: HR stands for hazard ratio, CI for lower 95 and upper 95 confidence interval for the hazard ratio. HRs are for 1 SD increase in these measures.

ICU	HR	CI	P-value
falling	1.3148	(1.053–1.6417)	0.0157
dishes	0.8172	(0.6901–0.9677)	0.0193
girl	1.1895	(1.0231–1.3829)	0.024
dishcloth	0.8368	(0.712–0.9835)	0.0307
boy	1.1704	(1.0116–1.354)	0.0344
woman	1.2066	(1.0013–1.4539)	0.0484

predictability over chance ($p<0.001$). We observed a ten-point increase in accuracy obtained by adding linguistic variables to the non-linguistic variables (non-linguistic alone 0.59, combined 0.69). This indicates that linguistic variables offer significant information over the non-linguistic variables in terms of their predictive diagnostic ability. The ratio of z-scores relative to the hull hypothesis indicated that the linguistic variables yielded a classification performance 2.4 times better than non-linguistic variables; the ratio of AUC gains respect to chance, provides a comparable value (0.19/0.09 = 2.11).

To examine the effects of education and sex on performance of the model using linguistic variables and the hold-out method, AUC scores were computed for participants with college degree vs participants without a college degree, and for females vs males. The participants with a college degree were harder to predict than participants without a college degree (AUC of 0.70 for college-degree vs 0.76 for no-college degree, see Supplementary Figure 3 for the ROC curves). The ratio of z-scores indicated that classification of the participants with no college degree was 1.52 times better than for the participants with college degree (as above, the gain ratio is 0.26/0.20 = 1.3). Similarly, females were both more accurately and more confidently predicted than males, and the difference is substantial (AUC of 0.83 for females vs 0.64 for males, see Supplementary Figure 3 for the ROC curves). The ratio of z-scores indicated that the females were classified 2.61 times better than males when compared to chance (the gain ratio is 0.33/0.14 = 2.35).

The longitudinal analysis in Fig. 5 shows the results of NMF factorization of linguistic and NP variables, and demonstrates that an unsupervised grouping of the variables without using time information indeed shows a clear temporal trend, as well as a differentiation between cases and controls which starts several years before cognitive impairment. The plot shows the change in the loading of each time interval on the first component obtained by NMF, where 0 in the horizontal axis stands for the date of cognitive impairment onset. The green line shows the controls' progression in time, whereas the blue line shows the cases's progression in time, with a steeper decline for the cases. Supplementary Figure 5 shows the loading of the factors on the first component from the NMF analysis, which shows the respective contribution of linguistic and NP variables in the computation of the plot in Fig. 5 in the manuscript.

For the Cox proportional-hazards analysis, we used 143 participants, of which 28 were censored, with a total of 1159 person-years, where average was 8.10 years per person. See Table 4 for all statistically significant linguistic variables according to the Wald statistic. Our results show that using the referentially generic terms *boy, girl, woman* instead of the more specific *son, brother, sister, daughter, mother* to refer to the subjects in the picture is associated with higher risk of AD. Our results also show that mentioning the details in the picture, such as the *dishcloth* and the *dishes*, is associated with lower risk of AD. Consequently, this analysis revealed that the strongest prognostic factors of AD involved semantic processing.

4. Discussion

Our results demonstrate that it is possible to predict future onset of Alzheimer's disease using language samples obtained from cognitively normal individuals. Moreover, we showed that using linguistic variables from a *single administration* of the cookie-theft picture description task performed better than predictive models that incorporated *APOE*, demographic variables, and NP test results.

Linguistic competence is a behavioral marker of educational and occupational attainment, both of which have been suggested to increase 'cognitive reserve' by epidemiologic studies. Higher cognitive reserve allows some people to be more resilient to brain pathology than others [27], such that they can compensate the dysfunction and delay diagnosis of AD [28]. In this regard, we found a significant differentiation between participants with and without college education. Furthermore, it is well-known that the prevalence of AD is significantly higher in women as compared to men, and that women show a faster rate of progression after onset of cognitive impairment [29–31]. Similar to what we observed with educational attainment, we found that it is much easier to predict conversion in women than in men, suggesting that prodromal changes are more prominent in females than in males.

The linguistic variables that we identified as most relevant for predicting future onset of AD, prominently agraphia, telegraphic speech and repetitiveness (see Supplementary Table 3), have been consistently identified in the literature as associated with cognitive decline in dementia. Repetitive speech that involves repetitive questioning, repetitive stories/statements, repetitive themes have been reported in patients with dementia [32,33]. Studies on agraphia in dementia and in AD participants have shown that patients made more writing errors compared to controls [34]. Declines in structural complexity of utterances have been extensively investigated in people with Alzheimer's disease and dementia [35,36]. Another linguistic element that has been associated with dementia, referential specificity, was identified as having a strong weight in the survival analysis, which is supported by a large number of studies showing that semantic impairments are the earliest linguistic markers of dementia [37,38].

While the Cox Proportional-Hazards analysis identified semantic/lexical factors, these factors did not prove to be discriminatory in the classification tasks. We believe that this is due to the differences in the design of these analyses. An age threshold was used for inclusion in the control vs case group in the classification task, whereas the Cox analysis treated all participants with a diagnosis of AD *equally* as non-censored participant. As a result, among the 115 non-censored participants in the Cox analysis, 48 of them had MCI onset after 85 years old, which would put them in the control group in the classification task. The contrasting results in these analyses indicate that, in accordance with prior literature, semantic factors are predictive of future diagnosis of AD for all subjects regardless of the age of onset, as opposed to being predictive of AD onset before mid-eighties. Similarly, verbosity and lexical richness metrics, which stand out as strong markers of cognitive impairment in already demented patients [39], were not among the strong predictors of future diagnosis of AD in cognitively normal individuals in our study.

The result of the longitudinal analysis of linguistic and NP variables, depicted in Fig. 5, shows a steeper decline in the trajectory of aging for the AD group as compared with normal aging, which starts during the preclinical phase. Similar clinical trajectories for AD and normal aging were suggested in the literature [40].

The analysis of the written version of the CTT may be considered a limitation of our study. The spoken version of the task may reveal different aspects of linguistic dysfunction. Another limitation of our study is that a thorough analysis of the correlational structure of linguistic features and neuropsychological test scores is outside the scope of the present article. Finally, our definition of the ‘case’ and ‘control’ labels, while designed to be as clinically relevant as possible, is ultimately discretionary and open to interpretation.

Biomarkers such as cerebrospinal fluid or brain imaging [41] and neuropsychological tests [41,42] have been used to predict progression of MCI to AD/dementia. Most recently, very promising results were reported using Neurofilament light chain (NfL) for disease progression at the early pre-symptomatic stages of familial Alzheimer’s disease [43]. However, these are still technologically or logistically demanding, and require significant specialists’ involvement. On the other hand, simple, naturalistic and inexpensive speech probes, as our results suggest, can provide an assistive tool for the early detection and progression monitoring of AD, particularly given that such probes can be easily adapted to remote digital platforms with low patient burden.

Contributors

Elif Eyigoz contributed to the research design, implemented the coding and ran the experiments. She performed the literature review and drafted and edited the manuscript. Finally, she contributed to the interpretation of the results. Melissa Naylor contributed to the research design, reviewed and edited the manuscript, and contributed to the interpretation of the results. Guillermo Cecchi contributed to the research design, reviewed and edited the manuscript, and contributed to the interpretation of the results. Sachin Mathur contributed to the coding, and research design, and reviewed and edited the manuscript. Mar Santamaria contributed to the research design and reviewed the manuscript.

Funding

Pfizer, Inc. provided funding to obtain the data from the Framingham Study Consortium, and funding to IBM Corp. for the initial phase of the study.

Data sharing statement

In order to gain access to the Framingham Heart Study (FHS) data, investigators have to submit a research proposal for review by one or more FHS review committees. Approved study proposals further require a fully executed Data and Materials Distribution Agreement, and an IRB approval. The Data and Materials Distribution Agreement can be accessed from the following link: https://framinghamheartstudy.org/files/2017/08/Data-and-Materials-Distribution-Agreement.pdf

Declaration of Competing Interests

Elif Eyigoz and Guillermo Cecchi has worked as salaried employees of IBM Corp. for the full duration of this project. Melissa Naylor was a salaried employee of Pfizer, Inc. when assigned to this project, until October 2018, and since then has been a salaried employee of Takeda Pharmaceuticals. Sachin Mathur and Mar Santamaria have worked as salaried employees of Pfizer, Inc. for the full duration of this project. Guillermo Cecchi declares that IBM holds a patent (US-9508360-B2) for the extraction of one of the features used in the linguistic model.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eclinm.2020.100583.

References

[1] Roth CR, Helm-Estabrooks N. Boston naming test. Encyclopedia of clinical neuropsychology. Springer International Publishing; 2018. p. 611–5.

[2] Snowdon DA. Linguistic ability in early life and cognitive function and Alzheimer’s disease in late life. Findings from the Nun study. JAMA: J Am Med Assoc 1996;275:528–32.

[3] Dawber TR, Meadors GF, Moore Jr FE. Epidemiological approaches to heart disease: the Framingham study. Am J Public Health Nations Health 1951;41:279–86.

[4] Farmer ME, White LR, Kittner SJ, Kaplan E, Moes E, McNamara P, et al. Neuropsychological test performance in Framingham: a descriptive study. Psychol Rep 1987.

[5] Seshadri S, Wolf PA, Beiser A, Au R, McNulty K, White R, et al. Lifetime risk of dementia and Alzheimer’s disease: the impact of mortality on risk estimates in the Framingham study. Neurology 1997;49:1498–504.

[6] Au R, Seshadri S, Wolf PA, Elias MF, Elias PK, Sullivan L, et al. New norms for a new generation: cognitive performance in the Framingham offspring cohort. Exp Aging Res 2004;30:333–58.

[7] Goodglass H, Kaplan E. The assessment of aphasia and related disorders. Lea & Febiger; 1972.

[8] Cummings L. Describing the cookie theft picture: sources of breakdown in Alzheimer’s dementia. Pragm Soc 2019;10:153–76.

[9] Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families: the Framingham offspring Study. Am J Epidemiol 1979;110:281–90.

[10] Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.

[11] Seshadri S, Wolf P, Beiser A, Au R, McNulty K, White R, et al. Lifetime risk of dementia and Alzheimer’s disease: the impact of mortality on risk estimates in the Framingham Study. Neurology 1997;49:1498–504.

[12] Seshadri S, Beiser A, Au R, Wolf PA, Evans DA, Wilson RS, et al. Operationalizing diagnostic criteria for Alzheimer’s disease and other age-related cognitive impairment—part 2. Alzheimer’s Dement 2011;7:35–52.

[13] Au R, Seshadri S, Knox K, Beiser A, Himali JJ, Cabral HJ, et al. The Framingham brain donation program: neuropathology along the cognitive continuum. Curr Alzheimer Res 2012;9:673–86.

[14] Association AP, others. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub; 2013.

[15] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer’s disease: report of the NINCDS-ADRDA Work Group under the auspices of department of health and human services task force on Alzheimer’s disease. Neurology 1984;34 939−939.

[16] Bachman DL, Wolf PA, Linn RT, Knoefel JE, Cobb JL, Belanger AJ, et al. Incidence of dementia and probable Alzheimer’s disease in a general population the Framingham study. Neurology 1993;43 515−515.

[17] Morris JC, Storandt M, Miller JP, McKeel DW, Price JL, Rubin EH, et al. Mild cognitive impairment represents early-stage Alzheimer disease. Arch Neurol 2001:58.

[18] Morris JC. Mild cognitive impairment is early-stage Alzheimer disease: time to revise diagnostic criteria. Arch Neurol 2006;63:15–6.

[19] Stephan B, Hunter S, Harris D, Llewellyn D, Siervo M, Matthews F, et al. The neuropathological profile of mild cognitive impairment (MCI): a systematic review. Mol Psychiatry 2012;17:1056.

[20] Silverman JM, Smith CJ, Marin DB, Mohs RC, Propper CB. Familial patterns of risk in very late-onset Alzheimer disease. Arch Gen Psychiatry 2003;60:190–7.

[21] Silverman JM, Li G, Zaccario ML, Smith CJ, Schmeidler J, Mohs RC, et al. Patterns of risk in first-degree relatives of patients with Alzheimer’s disease. Arch Gen Psychiatry 1994;51:577–86.

[22] Silverman JM, Ciresi G, Smith CJ, Marin DB, Schnaider-Beeri M. Variability of familial risk of Alzheimer disease across the late life span. Arch Gen Psychiatry 2005;62:565–73.

[23] Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53:695–9.

[24] Goodglass H. Understanding aphasia. Academic Press; 1993.

[25] Thompson CK. Treatment of syntactic and morphologic deficits in agrammatic aphasia: treatment of underlying forms. Language intervention strategies in aphasia and related neurogenic communication disorders: fifth edition. Wolters Kluwer Health Adis (ESP); 2012. p. 735–55.

[26] Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. Quart J R Meteorol Soc: J Atmos Sci Appl Meteorol Phys Oceanogr 2002;128:2145–66.

[27] Stern Y. Influence of education and occupation on the incidence of Alzheimer’s disease. JAMA 1994;271:1004–10.

[28] Katzman R, Terry R, DeTeresa R, Brown T, Davies P, Fuld P, et al. Clinical, pathological, and neurochemical changes in dementia: a subgroup with preserved mental status and numerous neocortical plaques. Ann Neurol: Off J Am Neurol Assoc Child Neurol Soc 1988;23:138–44.

[29] Andersen K, Launer LJ, Dewey ME, Letenneur L, Ott A, Copeland JRM, et al. Gender differences in the incidence of AD and vascular dementia: the EURODEM studies. Neurology 1999;53 1992−1992.

[30] Viña J, Lloret A. Why women have more Alzheimer’s disease than men: gender and mitochondrial toxicity of amyloid-β peptide. J Alzheimer’s Dis 2010;20: S527–33.

[31] Mielke M, Vemuri P, Rocca W. Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. Clin Epidemiol 2014:37.

[32] Barton S, Findlay D, Blake RA. The management of inappropriate vocalisation in dementia: a hierarchical approach. Int J Geriatr Psychiatry 2005;20:1180–6.

[33] de Lira JO, Ortiz KZ, Campanha AC, Bertolucci PHF, Minett TSC. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. Int Psychogeriatr 2010;23:404–12.

[34] Lambert J, Eustache F, Viader F, Dary M, Rioux P, Lechevalier B, et al. Agraphia in Alzheimer's disease: an independent lexical impairment. Brain Lang 1996;53:222–33.

[35] Kempler D, Almor A, Tyler LK, Andersen ES, MacDonald MC. Sentence comprehension deficits in Alzheimer's disease: a comparison of off-line vs. on-line sentence processing. Brain Lang 1998;64:297–316.

[36] Lyons K, Kemper S, Labarge E, Ferraro FR, Balota D, Storandt M. Oral language and Alzheimer's disease: a reduction in syntactic complexity. Aging, Neuropsychol Cogn 1994;1:271–81.

[37] Martin A, Fedio P. Word production and comprehension in Alzheimer's disease: the breakdown of semantic knowledge. Brain Lang 1983;19:124–41.

[38] Appell J, Kertesz A, Fisman M. A study of language functioning in Alzheimer patients. Brain Lang 1982;17:73–91.

[39] Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology 2000;14:71–91.

[40] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer\Textquotesingles Dement 2011;7:280–92.

[41] Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T, et al. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PLoS ONE 2011;6:e21896.

[42] Pereira T, Lemos L, Cardoso S, Silva D, Rodrigues A, Santana I, et al. Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows. BMC Med Inform Decis Mak 2017:17.

[43] Preische O, Schultz S, Apel A, Kuhle J, Kaeser S, Barro C, et al. Dominantly inherited Alzheimer network. serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic Alzheimer's disease. Nat Med 2019;25:277–83.

Please cite this article as: E. Eyigoz et al., Linguistic markers predict onset of Alzheimer's disease, EClinicalMedicine (2020), https://doi.org/10.1016/j.eclinm.2020.100583