Alzheimer's & Dementia: Translational Research & Clinical Interventions 5 (2019) 964-973

Alzheimer's & Dementia

Featured Article

Temporal association of neuropsychological test performance using unsupervised learning reveals a distinct signature of Alzheimer's disease status

Prajakta S. Joshi[a,b], Megan Heydari[c], Shruti Kannan[c], Ting Fang Alvin Ang[a,d,e], Qiuyuan Qin[c], Xue Liu[a], Jesse Mez[f,g], Sherral Devine[a,e], Rhoda Au[a,d,e,f,g], Vijaya B. Kolachalama[c,f,h,i,*]

[a]*Department of Anatomy and Neurobiology, Boston University School of Medicine, Boston, MA, USA*
[b]*Department of General Dentistry, Boston University Henry M. Goldman School of Dental Medicine, Boston, MA, USA*
[c]*Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA*
[d]*Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA*
[e]*The Framingham Heart Study, Boston University School of Medicine, Boston, MA, USA*
[f]*Boston University Alzheimer's Disease Center, Boston, MA, USA*
[g]*Department of Neurology, Boston University School of Medicine, Boston, MA, USA*
[h]*Whitaker Cardiovascular Institute, Boston University School of Medicine, Boston, MA, USA*
[i]*Hariri Institute for Computing and Computational Science & Engineering, Boston University, Boston, MA, USA*

Abstract

Introduction: Subtle cognitive alterations that precede clinical evidence of cognitive impairment may help predict the progression to Alzheimer's disease (AD). Neuropsychological (NP) testing is an attractive modality for screening early evidence of AD.

Methods: Longitudinal NP and demographic data from the Framingham Heart Study (FHS; N = 1696) and the National Alzheimer's Coordinating Center (NACC; N = 689) were analyzed using an unsupervised machine learning framework. Features, including age, logical memory-immediate and delayed recall, visual reproduction-immediate and delayed recall, the Boston naming tests, and Trails B, were identified using feature selection, and processed further to predict the risk of development of AD.

Results: Our model yielded 83.07 ± 3.52% accuracy in FHS and 87.57 ± 1.19% accuracy in NACC, 80.52 ± 3.93%, 86.74 ± 1.63% sensitivity in FHS and NACC respectively, and 85.63 ± 4.71%, 88.41 ± 1.38% specificity in FHS and NACC, respectively.

Discussion: Our results suggest that a subset of NP tests, when analyzed using unsupervised machine learning, may help distinguish between high- and low-risk individuals in the context of subsequent development of AD within 5 years. This approach could be a viable option for early AD screening in clinical practice and clinical trials.

Keywords: Alzheimer's disease; Neuropsychological testing; Machine learning; Framingham Heart Study; National Alzheimer's Coordinating Center

1. Introduction

The underlying pathophysiological processes of Alzheimer's disease (AD) begin many years before the clinical diagnosis of AD dementia [1–3]. This early phase of AD provides a critical opportunity where prognostic and therapeutic interventions may be most effective to delay or possibly prevent disease onset [3,4]. The use of cognitive tests has had an arguably profound impact on screening dementia cases in clinical settings [5]. The Folstein

*Corresponding author. Tel: 617-358-7253; Fax: 617-638-6671.
E-mail address: vkola@bu.edu

https://doi.org/10.1016/j.trci.2019.11.006

Mini-Mental State Examination (MMSE), a 30-item questionnaire, is among the most widely applied test for dementia screening [6]. However, a body of evidence suggests poor sensitivity of the MMSE scale for screening for early signs of dementia [7,8]. Another commonly employed method involves the use of a battery of neuropsychological (NP) tests in tandem with the diagnostic criteria for AD [5,9]. This method associates the performance on each test to affected cognitive domains and quantifies the test scores to the degree of cognitive impairment based on population averages [9,10]. Interpretation of test scores using this method can be challenging due to the involved subjectivity. Additionally, the administration of the entire NP battery can be tedious and time-consuming. Thus, a methodology involving fewer preselected NP tests that can accurately detect subtle changes in cognitive function could potentially help predict AD at an early time point.

Using data-driven machine learning approaches, we leveraged the longitudinal cognitive data from the Framingham Heart Study's (FHS) Offspring cohort (Gen2), to identify NP tests that are associated with early signatures of AD and help provide an early and robust prediction of subsequent clinical disease. The results were validated using a harmonized clinical dataset created by the National Alzheimer's Coordinating Center (NACC).

2. Study subjects and methods

2.1. Data collection and study sample

Three thousand and twenty-one participants from the FHS Gen-2 participants underwent the health Exams 7 and 8, the exam cycles used in this investigation (Fig. 1A). Between the years 1999 and 2000, the Gen-2 participants were administered a battery of NP tests as part of the FHS-ancillary study; this was the first NP exam (time point 1 [TP1]). Subsequently, the participants of this cohort underwent the second NP exam (time point 2 [TP2]) between 2005 and 2011 [11,12]. For the purpose of this study, we included 2282 participants who had undergone NP testing at TP1 and TP2, up to five years apart, and had valid scores on the 30- point MMSE scale, taken at Gen-2 FHS health exams 7 and 8 respectively.

Next, six NP tests (included Logical Memory-Immediate and delayed recall (LMi and LMd), Visual Reproduction-Immediate and Delayed recall (VRi and VRd), the Boston Naming Tests (BNT30) and Trails B) that were highly associated with AD, were selected using a method known as Kullback–Leibler (KL) divergence. Gen-2 participants with no missing values on the above-mentioned tests at the first two NP exams were included in the final study sample. Lastly, participants with prevalent Non-AD dementias and clinical stroke were excluded. The final study sample consists of 1696 FHS Gen-2 participants (Fig. 1A).

Written informed consent was obtained from all FHS participants. The FHS sample of this study was approved by the Institutional Review Board of Boston University Medical Campus and was monitored by a National Heart, Lung, and Blood Institute Observational Study Monitoring Board and followed their guidelines.

An independent validation analysis was conducted using the Uniform Data Set (UDS) created by NACC. The UDS is a standardized dataset comprising of harmonized clinical data collected from 29 Alzheimer's Disease Centers across the United States of America. Out of the 1955 individuals from UDS version 2 dataset (Fig. 1B) provided by the NACC data center, 689 had NP tests comparable to FHS and met all the above-mentioned study criteria (Fig. 1B); these individuals constituted the NACC study sample.

2.2. Study sub-sample to balance AD cases and normal controls

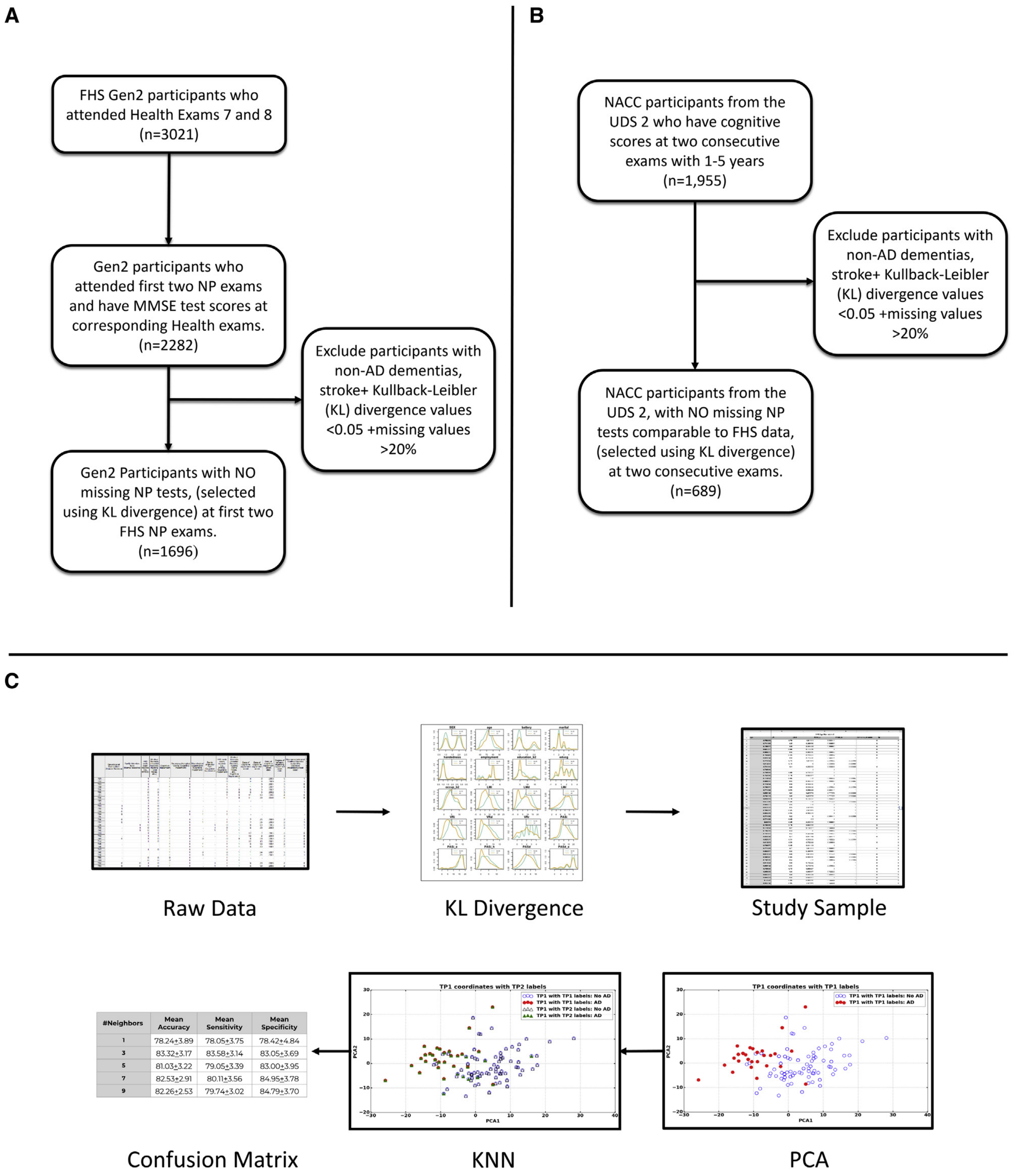
In the general population, the ratio of cognitively normal to AD individuals may be skewed. For the algorithm to effectively distinguish between the two groups, it is important to balance the ratio of AD to cognitively normal individuals. We thus kept the AD cases constant and performed cluster sampling within the healthy controls to avoid excluding the AD cases by complete random sampling of the entire study sample. Sixty-four unique control participants were randomly selected at each iteration, with replacement, and the new sample of 128 participants was used for further analysis. This control selection procedure was independently repeated 25 times for both FHS and NACC, thus, maximally contrasting the data from the controls with cases across the multiple iterations.

2.3. Data analysis

A single clinical parameter in machine learning terminology is called a "feature." Some features are highly associated with the disease. In order to distinguish these features and eliminate those that were superfluous, we first performed a data pre-processing step (the Kullback–Leibler (KL) divergence). This simplified the model, shortened training time, and enhanced the generalization of the pipeline.

2.4. Kullback-Leibler divergence

KL-divergence is a measurement of the difference between two probability distributions. This technique was used to measure the importance of each feature (NP tests, age, gender, education), with respect to the AD status (Fig. 2). For each feature, we plotted the distribution of the feature data corresponding to AD and non-AD (FHS: N = 1696; NACC: N = 689) participants at TP1. The greater the difference between these two distributions, the more information that feature will bring into the model. We calculated KL-divergence for all features in this dataset and selected those with values greater than 0.5 (top 30%). The features selected by this technique from FHS include Age, LMi, LMd, VRi, VRd, PASd, BNT30, and Trails B tests,


A
FHS Gen2 participants who attended Health Exams 7 and 8 (n=3021)
Gen2 participants who attended first two NP exams and have MMSE test scores at corresponding Health exams. (n=2282)
Exclude participants with non-AD dementias, stroke+ Kullback-Leibler (KL) divergence values <0.05 +missing values >20%
Gen2 Participants with NO missing NP tests, (selected using KL divergence) at first two FHS NP exams. (n=1696)
B
NACC participants from the UDS 2 who have cognitive scores at two consecutive exams with 1-5 years (n=1,955)
Exclude participants with non-AD dementias, stroke+ Kullback-Leibler (KL) divergence values <0.05 +missing values >20%
NACC participants from the UDS 2, with NO missing NP tests comparable to FHS data, (selected using KL divergence) at two consecutive exams. (n=689)
C
Raw Data
KL Divergence
Study Sample
TP1 coordinates with TP2 labels
TP1 with TP1 labels: No AD
TP1 with TP1 labels: AD
TP1 with TP2 labels: No AD
TP1 with TP2 labels: AD
PCA1
PCA2
TP1 coordinates with TP1 labels
TP1 with TP1 labels: No AD
TP1 with TP1 labels: AD
Confusion Matrix
KNN
PCA

#Neighbors	Mean Accuracy	Mean Sensitivity	Mean Specificity
1	78.24±3.89	78.05±3.75	78.42±4.84
3	83.32±3.17	83.58±3.14	83.05±3.69
5	81.03±3.22	79.05±3.39	83.00±3.95
7	82.53±2.91	80.11±3.56	84.95±3.78
9	82.26±2.53	79.74±3.02	84.79±3.70

Fig. 1. Criteria for sample selection and overview of the workflow. Flowcharts in (A) and (B) describe the criteria for selecting cases from FHS and NACC studies, respectively. (C) Longitudinal cognitive data from the FHS and NACC were obtained. Kullback-Leibler (KL) divergence was computed on age, education, gender, and each neuropsychological test independently. Individuals with complete data on eight parameters (Logical Memory [immediate and delayed], Visual Reproductions [immediate and delayed], Paired Associate Learning, Boston Naming Test 30 items, Trails B tests and age) with KL >0.5 were selected. Principal component (PC) analysis was performed on selected NP tests and age at time point 1 (TP1) and the first two PCs were plotted. The nearest neighbors approach was used on the first two PCs followed by majority voting and prediction of AD at time point 1 (TP2). This procedure was repeated 25 times, and an average confusion matrix was created. Abbreviations: AD, Alzheimer's disease; FHS, Framingham Heart Study; NACC, National Alzheimer's Coordinating Center.

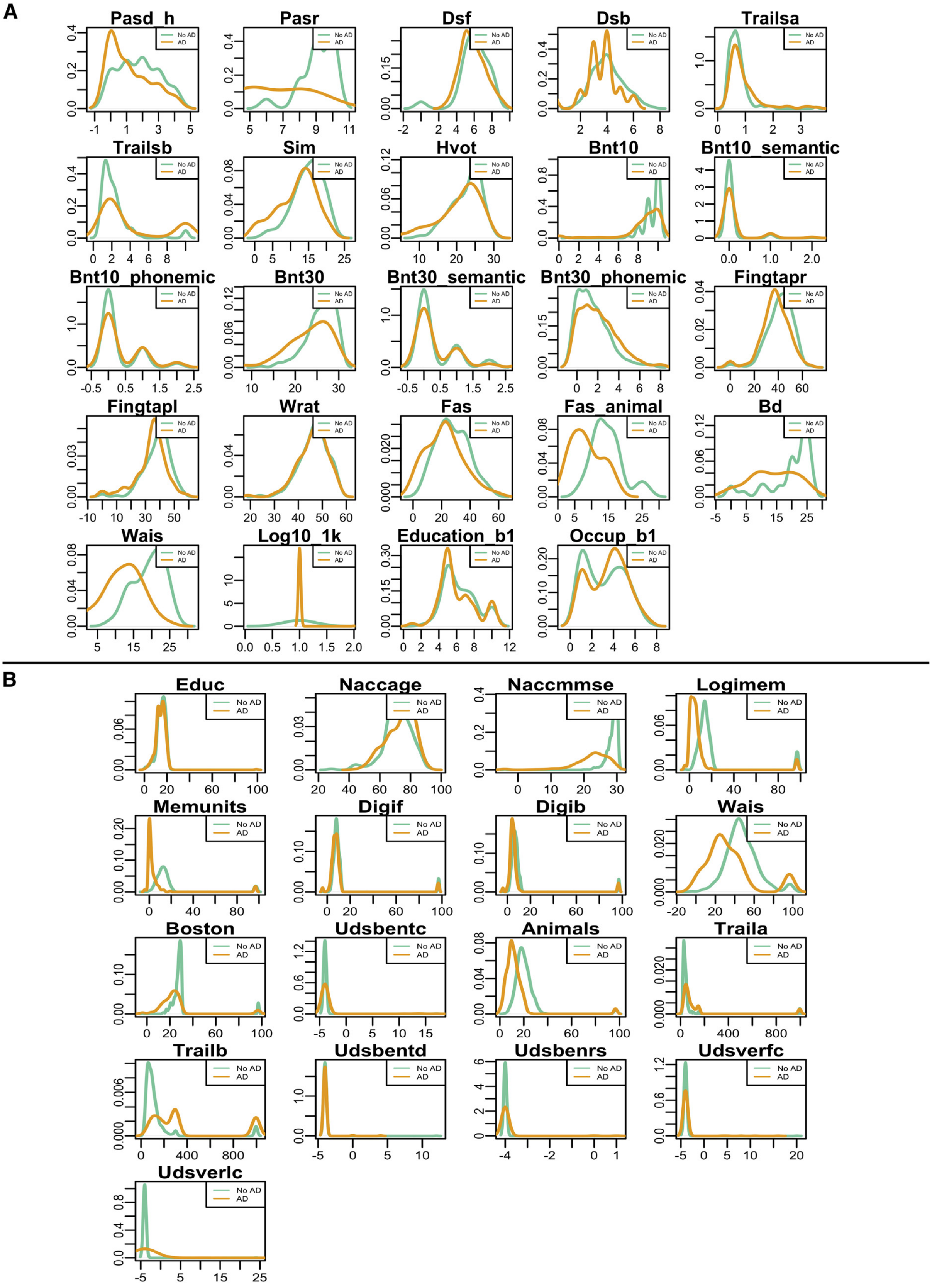


Fig. 2. Kullback-Leibler (KL) divergence distribution plots on (A) FHS and (B) NACC data sets. The KL divergence was calculated between the probability distributions (Y-axis) of Alzheimer's and no-Alzheimer's status at time point 1 (TP1) for various parameters including neuropsychological (NP) tests, age, gender, education (X-axis). Seven NP tests (Logical Memory Immediate and Delayed Recall, Visual Reproduction Immediate and Delayed recall, Paired Associate Learning, Boston Naming Tests 30 items, Trails B) and age had KL values greater than 0.5 and were selected for further analysis. Abbreviations: FHS, Framingham Heart Study; NACC, National Alzheimer's Coordinating Center.

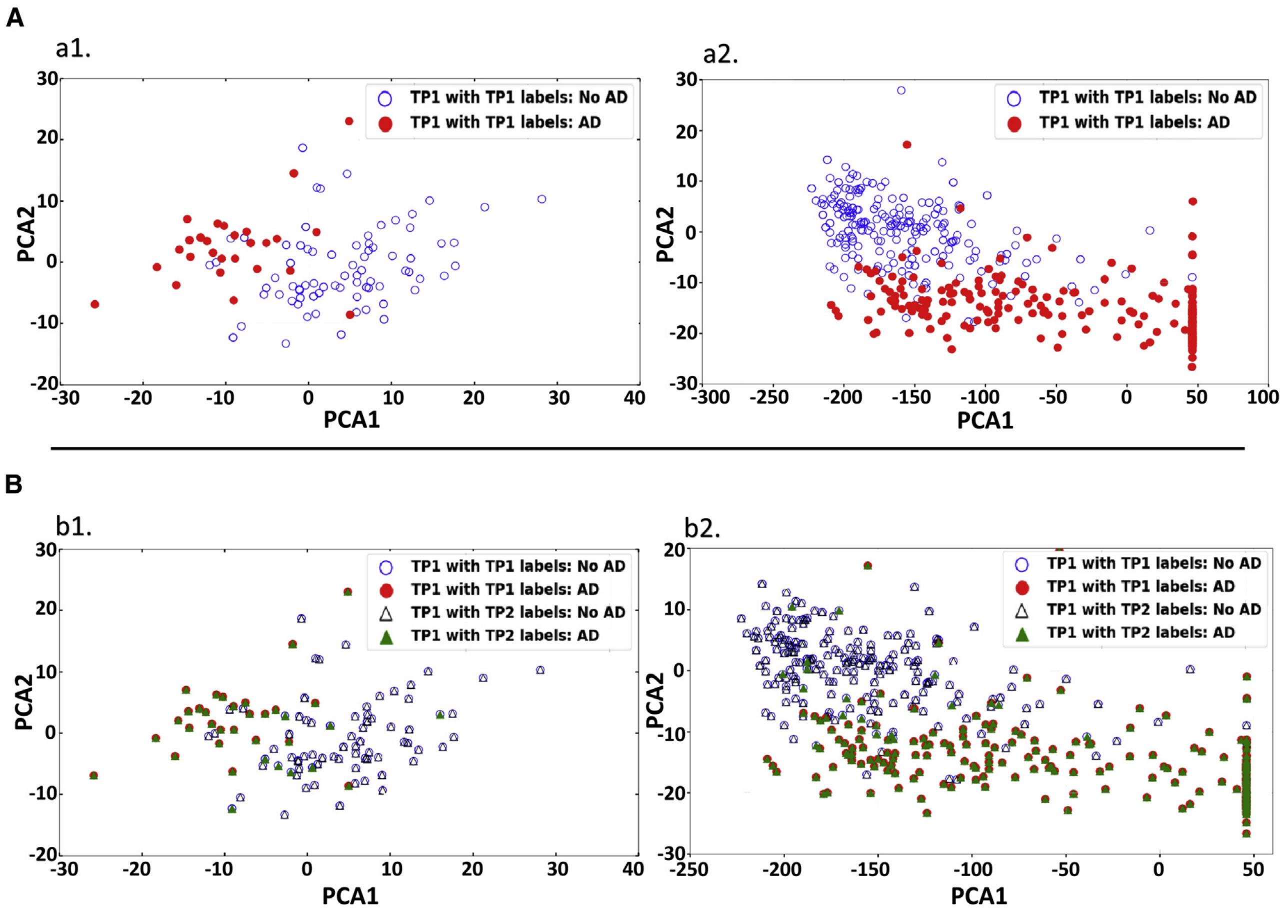


Fig. 3. PCA and K-NN plots. (A) The first two principal components (PCA1, PCA2) were plotted in (a1) FHS and (a2) NACC studies. They represent the maximum variance of the cumulative neuropsychological (NP) test scores and age. AD (red) and no-AD (blue ring) labels were added to represent the observed clustering. (B) Using the k-nearest neighbors approach, AD or no-AD predictions at time point 2 (TP2) (triangles) were superimposed on status at time point 1 (TP1) in (b1) FHS and (b2) NACC studies. Abbreviations: AD, Alzheimer's disease; FHS, Framingham Heart Study; NACC, National Alzheimer's Coordinating Center; PCA, principal component analysis.

and these features were used for further analysis (Fig. 2A). The above-selected features derived from the NACC study were used for model validation (Fig. 2B).

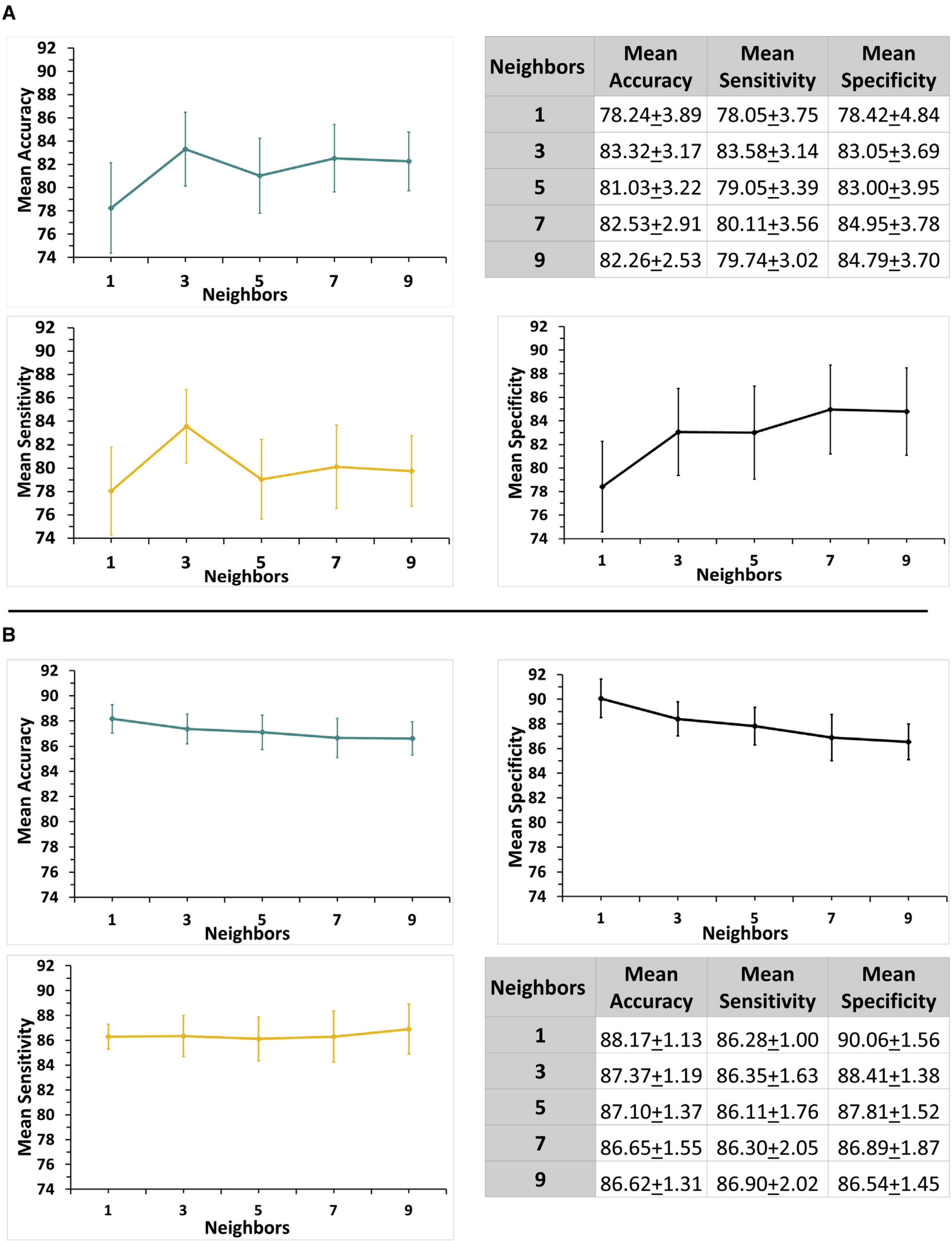
2.5. Principal component analysis

Each unique feature forms a coordinate axis in a multidimensional space. In this case, the eight features selected by using KL divergence would require eight-dimensional space (8D-feature space) to be represented. For analyzing this multidimensional data, the ML algorithms would require high computational cost. Additionally, it is likely that some features are highly correlated or add redundancies to the algorithm. We thus used a dimensionality reduction technique called principal component analysis (PCA) (Fig. 3A). This method enabled us to combine NP tests across multiple cognitive domains without introducing the investigator's subjectivity or simply averaging the NP test scores. Using PCA, we created a new feature space with two dimensions (called PC1 and PC2), whose coordinate values were computed as a linear combination of the existing 8D feature data.

2.6. Analysis using k-nearest neighbor (kNN) algorithm and majority voting

The PCA coordinates at the TP1 exam were plotted on a 2-D graph; each study participant represented a unique point on the coordinate axis (Fig. 3A). Examining one participant at a time from the study subsample of 128 participants, each individual (corresponding to a single point on the plot in Fig. 3A) was considered a "test subject," while the points in its spatial proximity on the plot were termed as its "neighbors." We considered odd numbers of neighbors (k = 1, 3, 5, 7, 9…) to prevent ties during majority voting used in subsequent parts of the algorithm. These "k" nearest neighbors in the 2-D feature space were determined based on the Euclidean distance.

Next, all points on the plot, excluding the test subject, were assigned their outcome status label (AD/normal) at the TP2 exam (Fig. 3B). Then, the test subjects' outcome



Neighbors	Mean Accuracy	Mean Sensitivity	Mean Specificity
1	78.24±3.89	78.05±3.75	78.42±4.84
3	83.32±3.17	83.58±3.14	83.05±3.69
5	81.03±3.22	79.05±3.39	83.00±3.95
7	82.53±2.91	80.11±3.56	84.95±3.78
9	82.26±2.53	79.74±3.02	84.79±3.70

Neighbors	Mean Accuracy	Mean Sensitivity	Mean Specificity
1	88.17±1.13	86.28±1.00	90.06±1.56
3	87.37±1.19	86.35±1.63	88.41±1.38
5	87.10±1.37	86.11±1.76	87.81±1.52
7	86.65±1.55	86.30±2.05	86.89±1.87
9	86.62±1.31	86.90±2.02	86.54±1.45

Fig. 4. Model performance on (A) FHS and (B) NACC data sets. Mean accuracy, mean sensitivity, and mean specificity are plotted as a function of the number of nearest neighbors. Performance metrics were generated by running the model 25 times and averaging them. Mean values are also summarized in the tables. Abbreviations: FHS, Framingham Heart Study; NACC, National Alzheimer's Coordinating Center.

status at the TP2 exam was predicted based on majority voting of the outcome label of its "k" neighbors. For example, if a test subject has 5 neighbors (i.e., k = 5) and three out of the five neighbors have "AD" as their outcome label at the TP2 exam, then the test subject's outcome at the TP2 exam will be predicted as "AD." This procedure of predicting the outcome label at TP2 was repeated for all the 128 study participants, considering each unique participant as a test subject. Based on the model's ability to predict AD/normal outcome at TP2, the accuracy, sensitivity, specificity, positive predictive values, and negative predictive values were calculated for the "k" neighbors (k = 1,3,5,7, and 9).

For comparison, the kNN algorithm was applied on the 30-point MMSE scores at the nearest NACC and FHS health exams, was used to predict AD/normal outcomes, and the accuracy, sensitivity, and specificity of this model were computed. The PCA and kNN algorithms for the MMSE and NP tests were repeated 25 times, and the accuracy, sensitivity, and specificity for the "k" neighbors, obtained at each iteration, were averaged (Supplementary Figs. 1 and 2).

3. Results

The current study sample consists of 1696 FHS Gen-2 participants and 689 NACC participants, out of which 64 FHS participants had AD, and 1632 were cognitively normal at the baseline NP exam (TP1); 271 NACC participants had AD, and 418 were cognitively normal at TP1.

In FHS, individuals with AD were significantly older, and few had reached a level of education beyond high school, compared to participants without AD (Table 1). Conversely, in NACC, the age and number of years of education were not significantly different between AD and NC participants. On average, AD participants had lower scores on all the selected cognitive tests; while this finding was more evident in the NACC study except the VRi, VRd tests, this difference reached 0.05 level significance only on LMd, VRi, VRd, BNT-30, Trails B and MMSE tests in the FHS study (Tables 1).

Table 1
Descriptive statistics on demographics and cognitive measures of the FHS and NACC study participants at baseline/time-point1 (TP1)

Baseline characteristics of the FHS population–Mean (SD)

	AD (N = 64)	No-AD (N = 1632)	*P* value
Demographic Characteristics			
Age (years)	74.46 ± 4.24	61.20 ± 9.26	.0054*
Sex (Female), N (%)†	35 (54.68)	877 (53.74)	.4034
Education (High school or more), N (%)†	19 (29.68)	81 (04.96)	.0003*
Neuropsychological Measures			
Verbal memory			
LMi	7.54 ± 3.71	10.83 ± 3.26	.4280
LMd	4.31 ± 5.02	10.90 ± 3.42	.0229*
VRi	3.46 ± 3.53	9.39 ± 3.07	.0001*
VRd	2.54 ± 2.51	8.58 ± 3.29	.0001*
Boston Naming Test			
BNT30	21.38 ± 3.69	27.71 ± 2.26	.0031*
Visual scanning and motor speed			
TrailsB	4.22 ± 3.46	1.29 ± 0.75	.0100*
Mini-Mental State Exam			
MMSE score	27.62 ± 1.85	29.02 ± 1.25	.0507*

Baseline characteristics of the NAAC population–Mean (SD)

	AD (N = 271)	No-AD (N = 418)	*P* value
Demographic Characteristics			
Age (years)	73.01 ± 9.37	71.63 ± 9.31	.056
Sex (Female), N (%)†	142 (52.40)	282 (67.46)	<.0001*
Education (Years of education)	14.54 ± 6.53	15.04 ± 5.65	.285
Neuropsychological Measures			
Verbal memory			
LMi	8.17 ± 1.93	19.23 ± 2.41	<.0001*
LMd	6.22 ± 2.05	17.45 ± 2.17	<.0001*
VRi	3.67 ± 2.41	3.95 ± 1.03	.073
VRd	3.90 ± 0.80	3.96 ± 0.78	.293
Boston Naming Test			
BNT30	24.24 ± 18.89	29.70 ± 14.74	<.0001*
Visual scanning and motor speed			
TrailsB	393.13 ± 340.7	135.04 ± 184.96	<.0001*
Mini-Mental State Exam			
MMSE score	21.76 ± 5.95	28.57 ± 2.28	<.0001*

*<.05 significance levels for differences between AD and NC.
†Counts and percentages were calculated for categorical variables.

3.1. PCA and classification results

The study subsample of 128 FHS participants was used for each model run using PCA followed by k-NN classification (Appendix S1). On average, across the 25 iterations, the NP test models with 3 (k = 3) and 7 (k = 7) neighbors resulted in the highest accuracies of 83.94 ± 3.69 and 83.07 ± 3.52, respectively. While the k = 7 model had a higher specificity (85.63), the k = 3 model had a higher sensitivity (83.68). Out of the subsample, the iteration with the best prediction with three neighbors yielded 92.11% accuracy, 88.16% sensitivity, and 96.05% specificity (Fig. 4A). The MMSE model using the FHS data, had an accuracy of 56.73 ± 0.60%, sensitivity of 14.47 ± 0.00% and specificity of 98.05 ± 1.20%.

This method of using PCA followed by k-NN classification was validated using the NACC data (Appendix S2). Across the 25 iterations of the NACC data, the NP test models with three and five neighbors had the highest accuracies of 87.57 ± 1.19% and 87.10 ± 1.37%, respectively.

Both the k = 3 and k = 5 models had similar sensitivities 86.74 ± 1.63% and 86.39 ± 1.763% and the k = 3 had a slightly better specificity (88.41 ± 1.38%) compared to the k = 5 model (87.81 ± 1.52%) (Fig .4B). The MMSE model based on the NACC data yielded an average accuracy of 51.34 ± 0.13%, a sensitivity of 2.73 ± 4.44 ×10–16, and a specificity of 99.94 ± 0.26%.

3.2. Secondary analysis

A secondary stratified analysis was performed on the FHS sample. We computed the mean change in scores across each of the 7 selected NP tests, as well as the MMSE test scores, in two groups (1) those who were cognitively intact at TP1 and converted to AD at TP2 (Converters), and (2) those who remained cognitively normal across the two time intervals (Nonconvertors) (Supplementary Table 3). In general, there was a decline in the MMSE, as well as NP test scores, in the convertors as well as the nonconvertors. The average change in scores was 0.96 ± 2.75 on the 8 NP tests and 0.88 ± 1.82 on the MMSE test, for the convertors, and for the nonconvertors, the change in scores was 0.30 ± 2.13 on the 8 NP tests and 0.24 ± 1.59 on the MMSE test. On comparing the change in individual NP test scores from TP1 to TP2, between the convertors and nonconvertors, only LM-DR, VM-DR, BNT30, and Trails B tests were significantly different between the two populations. We further assessed the subpopulation that was predicted correctly versus incorrectly from the iteration that resulted in the highest accuracy of 87.50% (Supplementary Table 3); only the change in scores on LM-DR, VM-DR, and Trails B tests were significantly different.

4. Discussion

We assessed the cumulative potential of a subset of NP tests to identify an early cognitive decline in a population-based cohort and predicted the future outcome (AD/NC) at an individual level. We demonstrate the value of using an unsupervised learning framework to identify a group of individuals who exhibit an early indication of global cognitive deterioration. The results from our analysis across the two datasets also highlight the capacity of unsupervised machine learning in detecting subtle changes on cognitive tests, which may not be identified as accurately using the MMSE test or regular statistical modeling of NP test results. The MMSE test is among the most commonly used tests for dementia screening [6]. However, as demonstrated by this study and findings from the literature, the MMSE test has poor sensitivity for early detection of cognitive impairment; this is likely due to the ease with which the individuals perform on the test prior to the onset of substantial impairment [7,8]. Thus, the use of a battery of NP tests may be a better alternative for dementia screening.

Traditionally, an individual's score on each NP test is compared to the mean score of age and education comparable to a normal reference group [13]. Together with evidence of functional impairment, a drop in the performance at or below the 5th percentile in two domains compared to the reference group is used to diagnose AD dementia [10,13]. The drop in NP scores likely indicates an incident cognitive decline, which may be too late to allow effective interventions [14]. Additionally, the clinical interpretation of test scores using the above cut-off based method may not be uniform due to the involved subjectivity of clinically interpreting the test scores [15]. For example, an individual may have large variability in scores across different NP tests, with very high scores on certain tests and very low scores on others. In addition, the number of high or low scores depends on the number of tests administered and the correlations between these tests. In such cases, averaging scores across all the tests within the domain, as seen in our secondary analysis (Supplementary Table 3), fails to present a clear picture of the underlying disease state. Also, evidence suggests that cognitive domains may not be independent of each other, and deficits across multiple domains may be explained by a global deterioration in cognitive abilities seen in AD [16]. Accurate AD diagnosis mandates the understanding of the interplay between multiple cognitive domains, and none of the NP tests can singularly capture these multidomain changes. As a result, it is important to combine the essence of multiple NP tests to ascertain these subtle cognitive changes in AD [17,18]. This is precisely when techniques, such as PCA, can be useful as they can provide a lower dimensional projection of the combined NP tests (representing global cognition) to depict the largest variance, that is, most informative viewpoint of these data.

In agreement with our study, other investigators have also examined the diagnostic criteria for AD dementia through a series of NP tests to determine the number of factors that best represent AD. Some have used traditional multivariate approaches and explored these associations only in AD cases [10,19,20]. Lowenstein et al. [19] found that a six-factor model provided the best fit to the data, while a study by Davis et al. [20] found that a three-factor model to be the best fit. However, using traditional models to analyze highly related NP test features could lead to the problem of collinearity [21]. This issue would not arise in a PCA model since each principal component is linearly uncorrelated [17].

Previously, investigators used techniques, such as PCA and discriminant analysis, on cross-sectional data, to diagnose AD using NP tests. For example, Chapman et al. [10] created 13 component scores for each subject using weighted combined scores from multiple tests and obtained an accuracy of ~90% in determining the prevalent AD cases; however, this study lacked a longitudinal predictive element, which the present study has provided. By leveraging the longitudinal component of the FHS and NACC data, we complimented the PCA-based clustering with the kNN algorithm to predict cognitive decline due to AD. This methodology yielded 83.07% accuracy in detecting AD cases tested within 1–5 years of baseline testing, which is an encouraging result considering that the model used only age and a subset of NP testing scores. However, this study has some limitations since the number of confirmed AD cases is fairly limited. Furthermore, it is important to acknowledge that different batteries of NP tests may be used in various clinics and research centers. Nevertheless, it is plausible that the changes in global cognitive function represented by the NP tests are more important than the specific tests themselves. It may be possible to use different NP tests that encompass the same cognitive domains as ours and replicate our findings. Additionally, although this study used an extensive dementia review procedure to diagnose AD cases, there may, however, be some non-differential misclassification of controls.

In conclusion, our study highlights the capacity of unsupervised machine learning approaches to capture the heterogeneity of cognitive profiles at an individual level and identify groups of individuals who exhibited similar patterns of global cognitive deterioration. Such tools could serve as efficient frameworks to concomitantly process multiple NP tests and generate an overall signature, even when there is some variability in the individual's performance on various NP tests. Further validation of this framework is needed to enable it as a screening tool for AD cases or for recruiting participants for AD clinical trials.

Acknowledgments

This project was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through BU-CTSI Grant (1UL1TR001430), a Scientist Development Grant (17SDG33670323) from the American Heart Association, and a Hariri Research Award from the Hariri Institute for Computing and Computational Science & Engineering at Boston University, Framingham Heart Study's National Heart, Lung and Blood Institute contract (N01-HC-25195; HHSN268201500001I) and NIH grants (R56-AG062109, AG008122, R01-AG016495, and R01-AG033040). Additional support was provided by Boston University's Affinity Research Collaboratives program and Boston University Alzheimer's Disease Center (P30-AG013846).

Supplementary Data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.trci.2019.11.006.

RESEARCH IN CONTEXT

1. Systematic review: Research has shown that subtle cognitive trends preceding clinical evidence of cognitive impairment may help predict conversion to Alzheimer's disease [10, 19, 20].
2. Interpretation: In this study, an unsupervised machine learning approach was used to combine scores across select neuropsychological (NP) tests to identify groups of individuals exhibiting similar signatures of global cognitive deterioration. This signature was computed on the Framingham Heart Study Offspring cohort and validated on the National Alzheimer's Coordinating Center data. This framework also predicted an individual level progression to AD in five years.
3. Future directions: An NP test-based signature for early identification of individual AD risk is feasible and could be useful for early cognitive screening.

References

[1] Vickers JC, Mitew S, Woodhouse A, Fernandez-Martos CM, Kirkcaldie MT, Canty AJ, et al. Defining the earliest pathological changes of Alzheimer's disease. Curr Alzheimer Res 2016;13:281–7.

[2] 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's Dement 2011;7:280–92.

[3] Sperling R, Mormino E, Johnson K. The evolution of preclinical Alzheimer's disease: implications for prevention trials. Neuron 2014; 84:608–22.

[4] Yassine HN. Targeting prodromal Alzheimer's disease: too late for prevention? Lancet Neurol 2017;16:946–7.

[5] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement 2011;7:263–9.

[6] 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.

[7] Spencer RJ, Wendell CR, Giggey PP, Katzel LI, Lefkowitz DM, Siegel EL, et al. Psychometric limitations of the Mini-Mental State Examination among nondemented older adults: an evaluation of neurocognitive and magnetic resonance imaging correlates. Exp Aging Res 2013;39:382–97.

[8] Arevalo-Rodriguez I, Smailagic N, Roqué I Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, et al. Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev 2015:CD010783.

[9] Khachaturian ZS. Diagnosis of Alzheimer's disease. Arch Neurol 1985;42:1097–105.

[10] Chapman RM, Mapstone M, Porsteinsson AP, Gardner MN, McCrary JW, DeGrush E, et al. Diagnosis of Alzheimer's disease using neuropsychological testing improved by multivariate analyses. J Clin Exp Neuropsychol 2010;32:793–808.

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

[12] Satizabal C, Beiser AS, Seshadri S. Incidence of dementia over three decades in the Framingham heart study. N Engl J Med 2016;375:93–4.

[13] Brooks BL, Iverson GL. Comparing actual to estimated base rates of "abnormal" scores on neuropsychological test batteries: implications for interpretation. Arch Clin Neuropsychol 2010;25:14–21.

[14] Cavedo E, Grothe MJ, Colliot O, Lista S, Chupin M, Dormont D, et al. Reduced basal forebrain atrophy progression in a randomized Donepezil trial in prodromal Alzheimer's disease. Scientific Rep 2017;7:11706.

[15] Patrick RE, Hobbs K, Mathias L, Harper DG, Forester BP. The limitations of using cognitive cutoff scores for enrollment in Alzheimer's trials. Am J Geriatr Psychiatry 2019;27:1153–8.

[16] Alkadhi K, Eriksen J. The complex and multifactorial nature of Alzheimer's disease. Curr Neuropharmacol 2011;9:586.

[17] Lever J, Krzywinski M, Altman N. Principal component analysis. Nat Methods 2017;14:641–2.

[18] Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research - commentary. BioMed Eng Online 2014;13:94.

[19] Loewenstein DA, Ownby R, Schram L, Acevedo A, Rubert M, Argüelles T. An evaluation of the NINCDS-ADRDA neuropsychological criteria for the assessment of Alzheimers disease: a confirmatory factor analysis of single versus multi-factor models: J Clin Exp Neuropsychol 2001;23:274-84.

[20] Davis RN, Massman PJ, Doody RS. WAIS-R factor structure in Alzheimer's disease patients: a comparison of alternative models and an assessment of their generalizability. Psychol Aging 2003;18:836–43.

[21] The problem of multicollinearity. In: Allen MP, editor. Understanding Regression Analysis. Boston, MA: Springer US; 1997. p. 176–80.