Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 2 (2016) 113-122

Alzheimer's & Dementia

Cognitive & Behavioral Assessment

Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment

David Glenn Clark[a,b,*], Paula M. McLaughlin[c], Ellen Woo[d], Kristy Hwang[e], Sona Hurtz[f], Leslie Ramirez[f], Jennifer Eastman[g], Reshil-Marie Dukes[a], Puneet Kapur[h], Thomas P. DeRamus[i], Liana G. Apostolova[j]

[a]*Department of Neurology, Medical University of South Carolina, Charleston, SC, USA*
[b]*Department of Neurology, Ralph H. Johnson VA Medical Center, Charleston, SC, USA*
[c]*Ontario Neurodegenerative Disease Research Initiative, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada*
[d]*Department of Neurology, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA*
[e]*Oakland University William Beaumont School of Medicine, Rochester, MI, USA*
[f]*Drexel University College of Medicine, Philadelphia, PA, USA*
[g]*Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
[h]*Department of Neurology, SUNY Upstate Medical University, Syracuse, NY, USA*
[i]*Department of Psychology, University of Alabama at Birmingham, Birmingham, AL, USA*
[j]*Department of Neurology, Indiana University, Indianapolis, IN, USA*

Abstract

Introduction: The objective of this study was to assess the utility of novel verbal fluency scores for predicting conversion from mild cognitive impairment (MCI) to clinical Alzheimer's disease (AD).
Method: Verbal fluency lists (animals, vegetables, F, A, and S) from 107 MCI patients and 51 cognitively normal controls were transcribed into electronic text files and automatically scored with traditional raw scores and five types of novel scores computed using methods from machine learning and natural language processing. Additional scores were derived from structural MRI scans: region of interest measures of hippocampal and ventricular volumes and gray matter scores derived from performing ICA on measures of cortical thickness. Over 4 years of follow-up, 24 MCI patients converted to AD. Using conversion as the outcome variable, ensemble classifiers were constructed by training classifiers on the individual groups of scores and then entering predictions from the primary classifiers into regularized logistic regression models. Receiver operating characteristic curves were plotted, and the area under the curve (AUC) was measured for classifiers trained with five groups of available variables.
Results: Classifiers trained with novel scores outperformed those trained with raw scores (AUC 0.872 vs 0.735; $P < .05$ by DeLong test). Addition of structural brain measurements did not improve performance based on novel scores alone.
Conclusion: The brevity and cost profile of verbal fluency tasks recommends their use for clinical decision making. The word lists generated are a rich source of information for predicting outcomes in MCI. Further work is needed to assess the utility of verbal fluency for early AD.

Keywords: Alzheimer's disease; Cognitive neuropsychology; Dementia; MCI (mild cognitive impairment); Machine learning; MRI (magnetic resonance imaging); Natural language processing

1. Introduction

Alzheimer's disease (AD) is a major socioeconomic crisis for the 20th century, with a projected 14 million cases by the year 2050 [1]. The dominant hypothesis for the

*Corresponding author. Tel.: +1-843-792-7446; Fax: +1-843-792-8626.
E-mail address: clarkda@musc.edu

http://dx.doi.org/10.1016/j.dadm.2016.02.001

pathogenesis of AD involves early deposition of beta-amyloid in the brain, but clinical trials targeting amyloid during the past decade have not met primary endpoints [2–4]. There is now evidence that beta-amyloid accumulates in the brain >10 years before the onset of cognitive symptoms [5,6]. Although cognitive symptoms appear late, studies in autosomal dominant AD suggest that individuals with mutations have measurable changes in cognition several years before onset of symptoms, when compared to mutation-free individuals [5]. These discoveries raise the concern that if treatments targeting beta-amyloid are to work, it might be necessary to implement them before the onset of symptoms. Two practical challenges arise. First, what is the best way to design clinical trials to ensure that pathophysiological changes of AD are occurring, if the individuals we would like to enroll are asymptomatic and cannot be expected to seek medical attention? Second, if a beta-amyloid targeted treatment is proven to work in the asymptomatic or early symptomatic stages of the disease, how can we identify individuals in the general population who will benefit from them?

There is an imminent need for new methods of detecting the earliest changes of AD, as the available biological methods are expensive, invasive, or entail exposure to radiation. Candidate methods under investigation include brief neuropsychological tests, ocular imaging, speech signal analysis, and computerized assessment of gait [7]. Structural MRI has been evaluated for this purpose, but the typical approach requires more than one image over a 6–12-month period, making it relatively expensive for a single patient [8]. Another approach is to use machine learning methods to train a classifier to discern between amyloid-positive and amyloid-negative individuals using available predictor variables, such as demographic data, structural brain imaging, cognitive tests, and blood tests. This approach has shown some success in a recent analysis of individuals with mild cognitive impairment (MCI—a condition thought to be a risk state for dementia [9,10]) who were studied in the AD Neuroimaging Initiative study [11]. The classifier achieved 0.78 area under the receiver operating characteristic curve (AUC) on a test set and 0.76 AUC when predicting conversion from MCI to AD.

The current work focuses on verbal fluency tasks—very brief cognitive measures in which the participant is given 1 minute to generate as many words as possible within a certain category, such as animals, or that start with a specific letter. The traditional method of scoring these tests is to simply count the number of unique, valid items in the list. The raw score obtained has proven clinical value [12–14], and as a result, verbal fluency tasks are performed in many research studies on AD or other cognitive disorders. However, there is strong evidence that careful examination of the words produced during the tasks may have additional clinical value, apart from or in addition to the raw score. The classic method for studying the explicit word content of these lists is to identify clusters of consecutively listed words that are related in some way (e.g., they have similar meaning, rhyme, or start with same two letters) [15–18]. The average length of these clusters is termed the clustering score and is thought to relate to spreading activation in a semantic or lexical network. The number of transitions between clusters is termed the switching score and is thought to relate to an individual's ability to deliberately change the subcategory of items that is currently being searched. Investigators have found that each of these scores has value for predicting dementia in longitudinal studies [19,20]. Some investigators have made use of unsupervised learning methods either on a corpus of fluency word lists [21] or on large English-language corpora [19,22] to improve prognostications.

In the present study, we develop new models for estimating risk of dementia conversion in MCI patients using measures derived from structural brain images and novel verbal fluency scores. In the statistical sub-discipline of machine learning, classification is often enhanced through expansion of the set of predictive features. This approach differs from the traditional inferential statistical approach, in which the primary goal is to identify statistically significant relationships between the independent and dependent variables. A potential weakness of the traditional approach is that one may develop a model with very poor predictive accuracy although it contains only statistically significant predictors. In machine learning, there is no immediate ambition to explain the relationship between the outcome and individual predictors—instead, the model is justified through the quality of predictions it yields. With this goal in mind, we developed a large set of novel predictive scores for five fluency tasks (categories animals and vegetables, and letters F, A, and S). Some of these novel predictive scores have roots in previous work (e.g., based on clustering, switching, or independent components analysis [ICA]), whereas others were developed specifically for this project. Some of the new scores are based on fundamental lexical qualities, such as syllable counts or frequencies of the words generated. Both of these quantities have good face validity and are easy to obtain, and the machine learning approach permits us to consider several possible ways of using them, such as (for a given fluency word list) taking the average, taking the sum, or subtracting the minimum value from the maximum value (i.e., metric range).

We based several novel scores on graph theory, a branch of discrete mathematics that provides techniques for analyzing networks. For this approach, we viewed the words in each list as nodes in a network and created weighted graphs by assigning numerical values to the edges or connections between the nodes. These values corresponded to the semantic, orthographic, or phonologic similarity between the two words being connected. Several scores were derived directly from these weighted graphs. The computation of other measures depended on conversion of each weighted graph into an unweighted graph by first identifying a threshold of the similarity metric and then creating a new

graph containing only the edges that met the threshold. For further details and rationale behind the predictor variables, see the Supplementary Methods (available online). In addition, we derived several cerebral measurements from structural MRI scans, including hippocampal atrophy and patterns of cortical thinning.

Following trends in machine learning research [23], we took the approach of developing ensemble classifiers. In the case of this work, the ensemble classifiers were constructed by training several initial classifiers and then training a final classifier using estimates of risk from the initial classifiers. Our goal was to compare the major subsets of the predictor variables in terms of their value for predicting conversion to dementia from MCI. This goal will be an important step if detailed verbal fluency word list analyses are to contribute to future efforts for identification of early symptomatic or pre-symptomatic AD.

2. Methods

2.1. Neurocognitive testing and consensus diagnoses

One hundred fifty-eight individuals met the inclusion/exclusion criteria set by the UCLA Imaging and Genetic Biomarkers for AD (ImaGene) study. ImaGene prospectively enrolled and followed individuals recruited from two sources: (1) referring UCLA and outside neurologists and (2) our Alzheimer's Disease Research Center (ADRC) ongoing longitudinal database study. The latter group consists of existing research participants who agreed to be contacted for future research opportunities and met ImaGene inclusion and/or exclusion criteria. All subjects provided informed consent after detailed explanation by a study clinician, and the UCLA Institutional Review Board approved the study.

To be included, subjects had to be aged at least 50 years, able to independently carry out daily activities of living based on interview, and score $\geq$24 on the mini-mental state examination (MMSE) [24]. MCI diagnosis was based on Petersen criteria [25] and required an objective cognitive deficit of at least 1.5 SD below age-adjusted and education-adjusted neuropsychological norms on at least one neuropsychological test, global clinical dementia rating (CDR) score $<$1, preserved general cognitive function, and intact activities of daily living. Cognitively normal participants performed above the $-$1.5 SD cutoff on the neuropsychological tests (adjusting for age and education) and had a global CDR of 0. Exclusionary criteria for both groups were concurrent medical problems of sufficient severity to impact cognition, history of alcohol or drug abuse in the past 2 years, concurrent neurologic or psychiatric illnesses, contraindications to MRI, cortical strokes or significant white matter changes, and visual and hearing impairment that could interfere with cognitive testing.

During each visit, ImaGene participants underwent detailed clinical and cognitive examinations, blood draw, and magnetic resonance imaging (MRI) examination. The neuropsychological battery and average scores by diagnostic group are listed in Table 1. Diagnosis for each subject was based on a consensus by all UCLA ADRC neurologists, neuropsychologists, and other key study personnel.

Seventy (65.4%) of the MCI participants were classified as "amnestic" because of poor performance on memory measures. Eighteen of these individuals had only memory impairment. Among the 52 amnestic individuals with impairment outside memory, 18 had impairment in only one nonmemory domain (14 executive, one language, two visuospatial, and one attention). The other 34 amnestic MCI patients had impairment in more than one additional nonmemory domain. Thirty-seven (34.6%) of the MCI participants were classified as nonamnestic. Thirty of these individuals had impairment in only one nonmemory domain (12 executive, six language, 11 visuospatial, one attention), whereas the other seven had impairment in executive function plus at least one other domain. During $>$4 years of follow-up, 24 of 107 individuals with MCI at baseline were determined by the consensus panel to have converted to dementia. Conversion was noted between 0.98 and 4.08 years after the baseline evaluation (M = 1.83 years, SD = 0.84 years). Converters were predominantly amnestic (21 of 24, or 87.5%). Twenty-two of the converters met clinical criteria for AD. The other two cases were clinically diagnosed as having dementia with Lewy bodies. One of these patients died and at autopsy was found to have hippocampal sclerosis without Lewy bodies. The other DLB patient has undergone positron emission tomography with an amyloid-detecting tracer and is amyloid positive.

2.2. Overview of machine learning approach

Fluency scores (traditional and novel) and brain imaging measurements were computed and used to create classifiers for discerning individuals with MCI who converted to AD from those that did not convert. All scores were placed in a single data matrix, and missing values for any given score were imputed as the mean of all the nonmissing values for that score. We performed five analyses, each using a different subset of the available scores: raw (traditional scores and counts of intrusions and repetitions), brain (measures derived from structural MRI), raw + brain (the union of the raw set and the brain set), novel (all scores derived from verbal fluency lists, including the raw scores), and novel + brain (the union of the novel and brain sets). Demographic variables of age, sex, and education were included for all analyses.

The quality of each classifier was assessed using leave-one-out cross-validation. This means that a separate classifier was constructed with each MCI participant left out and the classifier was then used to make a prediction about whether the left-out participant converted to dementia. The following three analysis steps were undertaken during each cross-validation loop: variable selection, training of an

Table 1
Participant demographics and selected neuropsychological test scores

	CN (n = 51)	MCI-non (n = 83)	MCI-con (n = 24)
Age (y)	68.9 (7.9)	68.7 (8.6)	73.8 (7.9)***
Sex (M:F)	28:23	37:46	9:15
Education (y)	17.6 (2.2)****	16.0 (3.0)	16.0 (2.9)
Mini-mental state examination	28.9 (1.2)****	27.9 (1.7)	25.1 (3.1)****
Animals	22.0 (4.7)****	18.8 (5.1)	14.5 (4.9)****
Vegetables	15.0 (4.3)***	13.0 (4.1)	9.7 (3.7)***
F	16.8 (4.6)****	13.4 (5.1)	11.3 (5.5)*
A	15.6 (4.6)****	11.1 (5.1)	7.9 (4.7)***
S	16.9 (5.5)****	13.4 (5.1)	10.7 (5.0)**
Boston naming test	58.1 (1.9)****	52.1 (8.1)	47.4 (10.6)*
Digit span forward	10.8 (2.3)	10.2 (2.2)	9.5 (1.9)
Digit span backward	8.2 (2.2)****	6.6 (2.4)	5.6 (1.6)**
Trails A	24.9 (8.9)****	33.8 (14.6)	44.5 (18.1)**
Trails B	70.0 (37.2)****	103.6 (55.5)	161.8 (87.5)***
Stroop A	62.7 (12.2)**	69.0 (20.4)	87.7 (18.7)****
Stroop B	47.5 (8.4)**	51.3 (12.6)	56.5 (10.1)**
Stroop C	114.9 (28.3)****	137.5 (44.0)	181.2 (66.6)***
Wisconsin card sort (categories)	4.3 (0.9)****	3.4 (1.9)	2.4 (1.7)**
Wisconsin card sort (errors)	11.7 (8.9)****	22.7 (13.7)	35.8 (19.0)***
Logical memory I	42.9 (9.6)****	32.9 (11.1)	15.9 (8.1)****
Logical memory II	28.3 (7.1)****	18.5 (9.5)	4.6 (4.7)****
Visual recall I	82.9 (13.5)****	80.0 (16.6)	52.6 (17.4)****
Visual recall II	62.8 (25.2)****	38.1 (23.8)	15.2 (21.8)****
Rey-Osterrieth figure copy	33.4 (2.4)****	30.0 (4.7)	29.7 (4.7)
Rey-Osterrieth delayed recall	20.2 (6.7)****	12.7 (7.2)	7.4 (6.6)***

Abbreviations: CN, cognitively normal group; MCI-non, mild cognitive impairment nonconverter; MCI-con, mild cognitive impairment converter to AD. Numbers in parentheses are standard deviations. All statistical comparisons are made to the MCI-non group.
NOTE. *$P < .1$, **$P < .05$, ***$P < .01$, ****$P < .001$.

ensemble of individual classifiers, and combination of the ensemble predictions through sparse logistic regression.

Variable selection was performed by running the random forests algorithm [26] on the training data set one time with 400 trees and calculating importance for each variable. Importance values were converted to z scores, and separate thresholds were selected for each analysis with iterative search over thresholds between 0 and 2.0.

Each ensemble consisted of classifiers with four different architectures: random forests of conditional trees, support vector machines [27], naïve Bayes [28], and multilayer perceptrons [28]. All analyses were performed in R with the following additional libraries: party, e1071, and RSNNS. One classifier of each architecture was trained using the left-in data, and ten additional classifiers of each architecture were trained using bootstrap samples of the left-in data for a total of 44 classifiers. Predictions from each classifier in the ensemble were obtained on the training data itself and for the left-out data point.

Predictions of the ensemble were combined linearly with sparse logistic regression to yield the final prediction for the left-out data point. The sparse logistic regression model was trained using conversion as the outcome variable and the ensemble predictions on the training data as the predictor variables. Sparse or "LASSO" (least absolute shrinkage and selection operator) regression produces a lower variance model than traditional regression while automatically performing variable selection. Because the ensemble had also generated predictions on the left-out data point, it was then possible to enter these predictions into the sparse logistic regression model to obtain the final prediction.

2.3. Verbal fluency tasks and scoring

Research participants underwent five fluency tasks. During these tasks, they were given 1 minute to generate as many words as possible within certain constraints. For two of the tasks, the constraint was semantic (animals and vegetables), and for three of the tasks, the constraint was orthographic (words had to start with the letters F, A, or S). A psychometrist or neuropsychologist recorded the words generated and the lists were subsequently transcribed into electronic text files by two of the authors (D.G.C. and R.M.D.). Raw and novel scores on verbal fluency word lists were calculated using custom Python software and the NetworkX Python library [29], as described in Table 2. Similarity measurements between words followed methods previously described for analyzing verbal fluency [30].

2.3.1. Brain measures

Two types of structural brain measures were incorporated into the predictive models: cortical thickness measurements and volumetric measures of several regions of interest

Table 2
Traditional and novel fluency scores

Score	Description
Traditional	
Raw	Count of unique valid items
Intrusions	Count of nonvalid items
Repetitions	Count of repeated items
Classic and miscellaneous lexical	
Clustering	Automatically calculated as described in Troyer, et al. (1998a) [15] and Clark, et al. (2014) [19]
Switching	Automatically calculated as with clustering
Mean frequency	Lexical frequencies for all words generated were calculated from the Google n-grams corpus and averaged
Mean number of syllables	Syllables for each word generated were quantified as the number of vowel symbols in the pronunciation listed in the Carnegie Mellon University Pronunciation Dictionary
Metric range of frequency	Calculated as the maximum frequency of words within a list minus the minimum frequency of words in the list
Sum of frequencies	Lexical frequencies were added together
Sum of reciprocal of frequencies	The reciprocal of all the lexical frequencies were added together
Independent components analysis (ICA)	
20 scores	ICA was performed on proximity matrices as described in Clark et al. (2014a). Each individual received 20 scores computed as the dot product of the individual's proximity matrix and 20 extracted components
Similarity metric based	
Algebraic connectivity	Second smallest eigen value of the Laplacian of the weighted graph
Average clustering coefficient	Given a vertex in a graph, the clustering coefficient for the vertex is the proportion of edges present among the immediate neighbors of the vertex. This value was calculated for all vertices in the thresholded graph and averaged.
Average degree	Average weight of all edges connected to each vertex in the graph
Diameter	Length of the longest geodesic in the weighted graph
Maximum betweenness centrality	For every pair of distinct vertices in the thresholded graph, the shortest path between the pair was identified. The betweenness centrality for each vertex was calculated as the number of shortest paths passing through that vertex. The score was the maximum of these values.
Radius	Length of the shortest geodesic in the weighted graph
Transitivity	3 times the proportion of triangles in a thresholded graph divided by the number of triads (two edges with a common vertex) in the graph
Coherence	A greedy algorithm was used to discover a short Hamiltonian path through the vertices of the weighted graph. The sum of the similarity weights on the actual path taken by the participant was divided by the sum of the similarities on the optimal path.
P&H clustering	Defined as for traditional clustering, but linkages between words were based on the edges in the thresholded graph, as described by Pakhomov & Hemmy (2014) [17]
P&H switching	Analogous to Pakhomov clustering

NOTE. Similarity metrics included orthographic, phonologic, and semantic similarity measures like those described in Clark et al. (2014b). Thus, there were three versions of each of the similarity-metric based scores.

(ROIs). MRI scans of sufficient quality were available for all but one participant from the MCI-non group (N = 157).

2.3.1.1. Cortical mapping

The cortical mapping measurements were obtained by first preprocessing high-resolution structural T1 brain images using previously described methods [31–33]. For each brain image, these methods yielded a set of 65,280 three-dimensional spatial coordinates and a GM thickness measurement at each point.

Independent components analysis (ICA) was used to reduce the dimensionality of the cortical surface data. To do so, a simple interpolation method was used to map the thickness measurements from each individual into a standard set of coordinates. For each three-dimensional point in the standard coordinates, the three nearest neighbors in the individual cortical surface data file were identified using a Euclidean distance measure. The cortical thickness at the standard point was set to an average of the thickness measurements at these three neighboring points, weighted by each point's proximity to the standard point. The interpolated cortical thickness measurements from the right and left hemispheres were concatenated for each research participant, and ICA was undertaken using the fastICA library for R. Thus, each component represented the entire bihemispheric cortical surface. Thirty components were extracted and reformatted for direct visual inspection. Component scores for individual research participants were computed as the dot product of the actual cortical thickness measures with each component.

2.3.1.2. Region of interest measures

Additional gross measures of the cerebrum included (for each hemisphere) hippocampal volumes, average cortical thickness, and volumes of the superior, inferior, and occipital portions of the lateral ventricles. Hippocampal volumes were derived from manual tracing of the hippocampus proper, dentate gyrus, and subiculum according to a well-established protocol [34,35]. Ventricular volumes were extracted after a semiautomated ventricular segmentation approach, in which lateral ventricles of four MRI scans

were initially manually traced and segmented into three partitions per hemisphere (superior horn, temporal horn, and ventricular body/occipital horn) [36,37]. These traces were then converted into three-dimensional parametric ventricular mesh models (termed "atlases") and were used to fluidly register each unsegmented study image, yielding one segmentation per atlas. The four ventricular segmentations thus derived for each participant were averaged to minimize automated labeling errors. Volumetric measurements were made on the averaged ventricular segmentation.

2.4. Assessment of classifier performance

Receiver operating characteristic (ROC) curves were plotted for each of the five ensemble classifiers. The AUC was reported for each classifier. In addition, for each ROC curve, the optimal cut point was defined as the threshold that maximized the F-measure (the harmonic mean of sensitivity and positive predictive value). Accuracy, sensitivity, specificity, negative predictive value, and positive predictive value were measured at this cut point.

3. Results

3.1. Similarity metrics

For examples of words within each fluency task that were judged to have high orthographic, phonological, or semantic similarity, see Supplementary Table 1. Percentages of edges meeting the similarity threshold of 1.0 standard deviations (for each fluency task and similarity measure) are shown in Supplementary Table 2.

3.2. Selected variables

Variables selected from the novel fluency score subset are listed in Table 3, along with the sum of the importance values assigned to each variable from all cross-validation iterations. Fig. 1 shows the gray matter independent component with highest estimated importance, which loads most heavily on points in the superior parietal lobe. Supplementary Fig. 1 shows the independent component with second highest importance. This component loads most heavily on points in the anterior mesial temporal lobe, an area likely to include the entorhinal and perirhinal cortices.

For the novel score analysis, scores from all five fluency tasks achieved sufficient importance to be selected, although S words figured less prominently. Focusing on the top 10 scores, measures of coherence occupied four of the slots (including the top 2), measures of lexical frequency occupied three slots, and the remaining three slots were occupied by the graph theoretical measures algebraic connectivity, radius, and transitivity. Scores based on ICA, clustering, and switching appeared in the list but with lower importance scores. Among raw scores, only animals achieved sufficient importance to be included. Age was the only demographic variable selected and was selected in only two cross-validation loops.

For variables selected during the other four analyses, see Supplementary Tables 3 and 4. Boxplots of 10 scores are shown in Fig. 2.

3.3. Prediction accuracy

ROC curves for the five classifiers are shown in Fig. 3. As shown in Table 4, the ensemble classifier trained only

Table 3
Cumulative importance values of variables selected from novel scores

Coherence semantic (A)	1589.10	ICA10 (S)	598.70	Raw (animal)	281.51
Coherence–ortho (veg)	1066.63	ICA17 (F)	590.24	ICA2 (veg)	191.70
Frequency–metric range (F)	939.63	Frequency–metric range (animal)	585.68	Coherence–phono (S)	156.36
Coherence–ortho (animal)	878.30	Algebraic connectivity–phono (animal)	585.45	Average degree–semantic (veg)	154.91
Algebraic connectivity–ortho (animal)	777.68	Frequency–sum (animal)	564.00	Average clustering coefficient–phono (veg)	136.50
Radius–ortho (A)	772.28	Coherence–semantic (veg)	534.27	Clustering–classic (animal)	122.58
Frequency–mean (animal)	766.87	Switching–phono (veg)	529.87	Diameter–ortho (A)	36.03
Frequency–sum reciprocal (animal)	745.18	Maximum betweenness–phono (animal)	521.75	Algebraic connectivity–ortho (A)	31.80
Transitivity–phono (veg)	690.30	Transitivity–semantic (animal)	495.77	ICA13 (A)	31.22
Coherence–ortho (A)	684.37	Algebraic connectivity–semantic (A)	475.37	Metric range of similarity–semantic (A)	18.30
Coherence–phono (A)	681.66	Average clustering coefficient–ortho (animal)	439.13	Frequency–mean (veg)	17.83
Maximum betweenness–phono (veg)	676.82	Maximum betweenness–semantic (animal)	420.87	Switching–semantic (animal)	9.14
Transitivity–semantic (S)	673.94	Switching–phono (animal)	407.99	Metric range of similarity–ortho (A)	9.09
Average clustering coefficient–semantic (S)	663.10	Frequency–sum (veg)	400.33	Age	8.86
Frequency–sum reciprocal (veg)	656.95	ICA4 (animal)	305.44	Diameter–semantic (animal)	4.57

Abbreviations: veg, Vegetable; ICA, independent components analysis.

NOTE. Each score (apart from age) originated from one of the five fluency tasks (A, animal, F, S, or veg). For scores dependent on measurements of lexical similarity, the type of similarity measure is included (orthographic, phonological, or semantic). Each importance value listed here represents the sum of the importance measurements across all cross-validation loops.

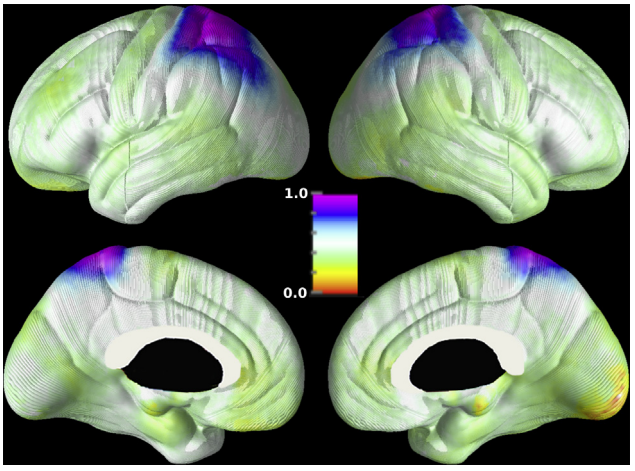

Fig. 1. Component 15 derived from independent components analysis of cortical thickness measures. The values of the component have been normalized to the interval [0,1]. Individuals with a higher gray matter thickness in the parietal lobes and lower gray matter thickness in the right mesial occipital region would achieve the highest scores for this component.

with novel scores performed the best on all measures apart from specificity, with AUC 0.872, and was significantly better than the classifier trained only with raw scores (AUC 0.719, $P < .05$ by DeLong test). The raw + brain ensemble showed the highest specificity (0.916). If our goal, however, is to develop an inexpensive screening test then we may place greater emphasis on sensitivity. The novel ensemble shows a clear advantage here, providing 100% sensitivity with 67.5% specificity (Fig. 3).

4. Discussion

The ability to rapidly, noninvasively, and inexpensively identify individuals at high risk for AD is crucial for the application of effective disease-modifying therapies for the general population [7]. Language is an abundant and readily collectible product of human cognition that is associated with neurological function and may be assessed at

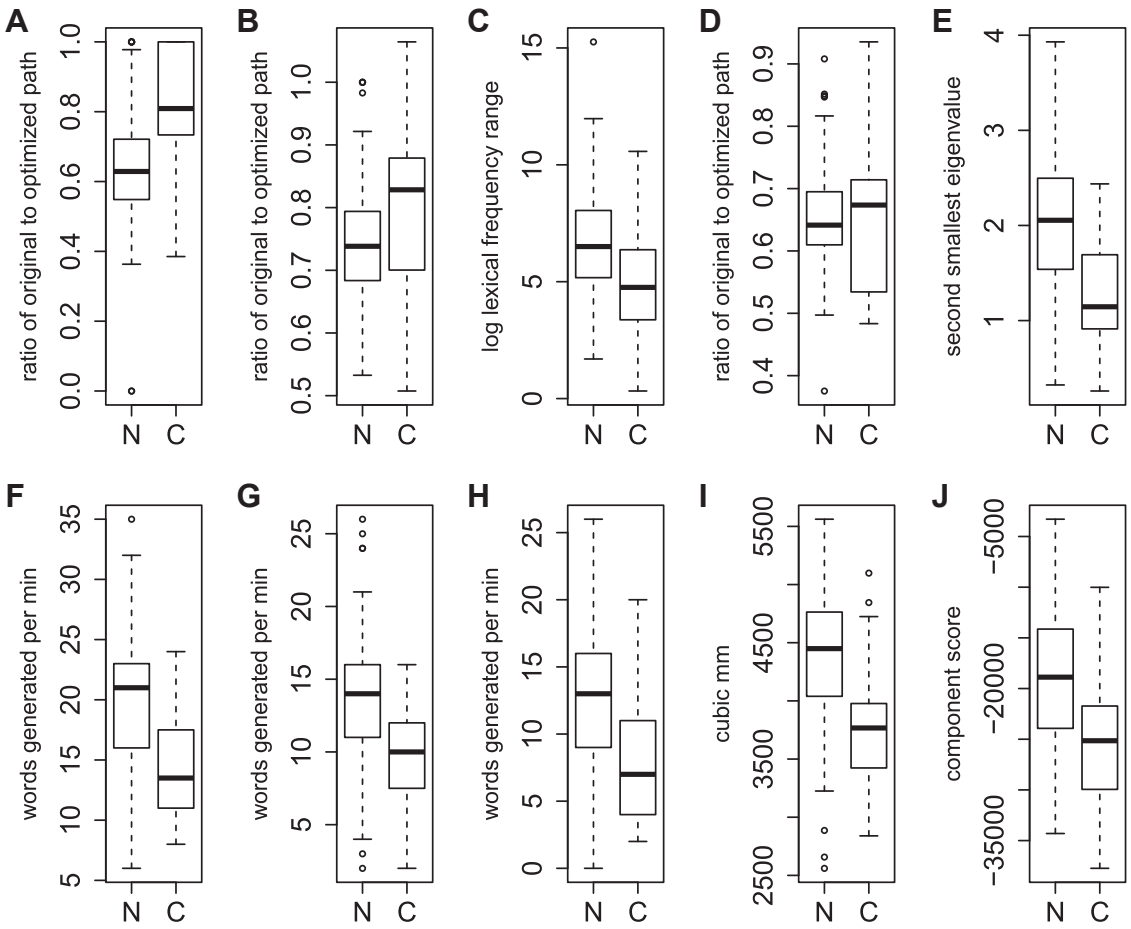

Fig. 2. Boxplots of 10 selected variables. The top row includes the five top-ranked variables from the analysis including only novel scores. The bottom row includes the three raw scores selected for the "raw" analysis and the two imaging scores selected for the "brain" analysis. Differences between the MCI-non (N) and MCI-con (C) groups are apparent for all variables shown. Factors that may be relevant but cannot be readily depicted include potential interactions among several variables and nonlinear relationships between an individual variable and conversion risk. (A) semantic coherence letter A; (B) orthographic coherence vegetables; (C) metric range of frequency letter F; (D) orthographic coherence animals; (E) orthographic algebraic connectivity animals; (F) raw score for animals; (G) raw score for vegetables; (H) raw score for letter A; (I) volume of right hippocampus; (J) gray matter volume score for independent component 15.

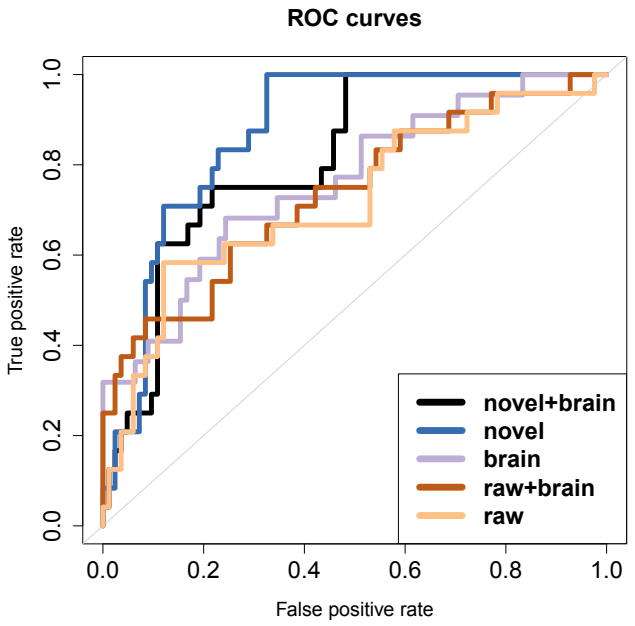

Fig. 3. ROC curves for the five ensemble classifiers. Novel verbal fluency scores yield the best AUC (0.872). This classifier may be thresholded to have sensitivity 1.00 with a specificity of 0.675. Abbreviations: ROC, receiver operating characteristic curve; AUC, area under the receiver operating characteristic curve.

multiple levels of representation. Language is often noted to be disrupted early during the course of AD [38–40], suggesting that measurements of language may play an important role in predictive models of AD. Verbal fluency tasks are a simple method for quickly obtaining a constrained linguistic sample and have proven value for differentiating causes of dementia [12–14]. In this study, we examined the value of an assortment of novel verbal fluency scoring methods for predicting subsequent cognitive and functional decline in MCI patients and evaluated the potential contribution of measures from structural brain imaging for the predictive models. We achieved good prediction results using cross-validated ensembles combined linearly with LASSO logistic regression models. These findings contribute to the literature on machine learning for predicting outcomes in MCI because they highlight the value of brief, information-dense cognitive tests from which many potential predictor variables may be extracted.

4.1. Predictor variables

Despite the well-known utility of traditional raw scores for diagnosis of cognitive disorders, these scores did not make large contributions to the best classifier. Other easily quantified measures, such as counts of intrusions and repetitions, demonstrated little capacity to predict MCI conversion.

Table 4
Quality of predictions made by the five ensemble classifiers

	AUC	F	Sensitivity	Specificity	NPV	PPV	Accuracy
Raw	0.719	0.583	0.583	0.880	0.880	0.583	0.813
Brain	0.760	0.536	0.682	0.756	0.894	0.441	0.740
Raw + Brain	0.735	0.524	0.458	0.916	0.854	0.611	0.813
Novel	0.872*	0.667	0.708	0.880	0.913	0.630	0.841
Novel + Brain	0.814	0.625	0.625	0.892	0.892	0.625	0.832

Abbreviations: AUC, area under the receiver operating characteristic curve; F, F-measure (harmonic mean of sensitivity and positive predictive value); NPV, negative predictive value; PPV, positive predictive value.
NOTE. *$P < .05$ compared to AUC for Raw classifier using DeLong test.

Several of the novel scores based on lexical similarity metrics made significant contributions to the final classifiers. Among the various tasks and types of lexical similarity, scores derived through the measurement of coherence appeared to have good utility. We note that three of the top five predictors in Table 3 were coherence measures in which the similarity metric did not coincide with the task demands (e.g., semantic similarity in the letter A task). In each case (as shown in Fig. 2), MCI converters had higher average scores than nonconverters on these measures. This finding suggests that the converters may have been more likely to be distracted by forms of word similarity that were not relevant to the current task. Despite the strong theoretical neuropsychological basis for clustering and switching scores, they were not found to be prominent among the other novel scores introduced here. However, previous work in which switching or clustering was found to have prognostic value was conducted on data from longitudinal studies with larger sample sizes and was not focused on individuals with MCI [19,20]. Thus, future work should continue to consider the potential value of clustering and switching scores, whether calculated by the classic method or with methods based on similarity scores.

4.2. Imaging

Novel verbal fluency scores outperformed structural MRI measures for predicting MCI conversion. The brain-only classifier achieved an AUC 0.760 (Table 4). This score is on par with AUC measures reported by other investigators using structural MRI and other biological measures, which have ranged from 0.734 (MRI + CSF in [41]) up to 0.843 (MRI only in [42]).

The inconsistent improvements we observed could not justify the expense of undertaking an MRI scan only for use in these types of classifiers. However, other MRI-based measurements, such as resting state functional MRI, diffusion kurtosis imaging, magnetic resonance spectroscopy, or arterial spin labeling, may provide better predictive features.

4.3. Limitations

A few limitations of this work should be noted, as they point the way for future research along these lines. First, owing to the small sample size, it was not possible to test the final classifiers on a held-out test set. This shortcoming was mitigated by using a rigorous cross-validation method. Second, the techniques for extracting predictor variables from MRI scans, although state-of-the-art for brain-mapping purposes, are extremely labor intensive. If MRI measures are to be used to achieve our clinical goals, it will be necessary to use an automatic imaging analysis pipeline. Third, the calculation of the predictor variables for this work rests on the availability of accurate electronic transcriptions of fluency word lists. Information about the latencies of the words generated could enhance the quality of the ICA scores and thereby the predictions from them. Moreover, the need for the transcriptions adds to the cost of the technique. Automatic transcription with speech recognition technology may make future work faster, cheaper, and more accurate.

4.4. Conclusions

Using cross-validated ensembles of classifiers trained with a variety of novel verbal fluency scores, we show generally good quality predictions of MCI conversion over approximately 5 years of follow-up, with most conversions fitting the AD phenotype. Many novel scores contribute to the quality of the final classifiers. However, lexical frequency measures and certain graph theoretical scores, especially those based on coherence, stand out as having the strongest relationships with conversion risk. Verbal fluency word lists contain a great deal of information, and detailed analysis of their contents may lead to the development of a rapid, inexpensive, and noninvasive method for detecting the earliest pathophysiological changes of AD.

Acknowledgments

Funding sources: VA (E6553W), NIH (R01 AG040770, K02 AG048240, P50 AG16570), and the Easton Consortium for Alzheimer's Drug Discovery and Biomarker Development. The authors are grateful to Glenn L. Clark, MD for comments on an earlier version of the article and to Naira Goukasian for assistance with preparation of the data and figures.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.dadm.2016.02.001.

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed the literature using traditional sources, such journal articles, meeting abstracts, and presentations. Recent observations regarding the natural history of Alzheimer's disease (AD) suggest that some treatments (e.g., those targeting amyloid) may be most effective if administered early. There is a need for inexpensive and noninvasive methods for detecting patients likely to benefit from new treatments.
2. Interpretation: Our findings point to the potential value of applying machine learning methods to natural language samples to realize the goal of early AD detection.
3. Future directions: Further research along these lines will seek (1) to validate these findings in larger samples of patients, (2) to integrate the method with other rapid, inexpensive tests, and (3) to apply speech recognition technology for rapid transcription and scoring of natural language samples.

References

1 Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. Am J Public Health 1998;88(9):1337–42.
2 Salloway S, Sperling RA, Fox NC, Blennow K, Klunk WE, Raskind M, et al. Two phase 3 trials of bapineuzimap in mild-to-moderate Alzheimer's disease. N Engl J Med 2014;370:322–33.
3 Doody RS, Thomas RG, Farlow M, Iwatsubo T, Vellas B, Joffe S, et al. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. N Engl J Med 2014;370:311–21.
4 Green RC, Schneider LS, Amato DA, Beelen AP, Wilcock G, Swabb EA. Effect of tarenflurbil on cognitive decline and activities of daily living in patients with mild Alzheimer disease: a randomized controlled trial. JAMA 2009;302:2557–64.
5 Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, et al. Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. N Eng J Med 2012;367:795–804.
6 Langbaum JB, Fleisher AS, Chen K, Ayutyanont N, Lopera F, Quiroz YT, et al. Ushering in the study and treatment of preclinical Alzheimer disease. Nat Rev Neurol 2013;9:371–81.
7 Laske C, Sohrabi HR, Frost SM, Lopez-de-Ipina K, Garrard P, Buscema M, et al. Innovative diagnostic tools for early detection of Alzheimer's disease. Alzheimer's Dement 2014:1–18.
8 Holland D, McEvoy LK, Desikan RS, Dale AM. Alzheimer's Disease Neuroimaging Initiative, Enrichment and Stratification for Predementia Alzheimer Disease Clinical Trials. PLoS ONE 2012;7:e47739.
9 Knopman DS, Boeve BF, Petersen RC. Essentials of the proper diagnoses of mild cognitive impairment, dementia, and major subtypes of dementia. Mayo Clin Proc 2003;78:1290–308.

10 Petersen RC, Aisen PS, Boeve BF, Geda YE, Ivnik RJ, Knopman D, et al. Mild cognitive impairment due to Alzheimer disease in the community. Ann Neurol 2013;74:199–208.
11 Apostolova LG, Hwang KS, Avila D, Elashoff D, Kohannim O, Teng E, et al. Brain amyloidosis ascertainment from cognitive, imaging, and peripheral blood protein measures. Neurology 2015;84:729–37.
12 Canning SJ, Leach L, Stuss D, Ngo L, Black SE. Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. Neurology 2004;62:556–62.
13 Marczinski C, Kertesz A. Category and letter fluency in semantic dementia, primary progressive aphasia, and Alzheimer's disease. Brain Lang 2006;97:258–65.
14 Monsch AU, Bondi MW, Butters N, Salmon DP, Katzman R, Thal LJ. Comparisons of verbal fluency tasks in the detection of dementia of the alzheimer type. Arch Neurol 1992;49:1253–8.
15 Troyer AK. Normative data for clustering and switching on verbal fluency tasks. J Clin Exp Neuropsychol 2000;22:370–8.
16 Troyer AK, Moscovitch M, Winocur G. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology 1997;11:138–46.
17 Troyer AK, Moscovitch M, Winocur G, Alexander MP, Stuss D. Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. Neuropsychologia 1998a;36:499–504.
18 Troyer AK, Moscovitch M, Winocur G, Leach L, Freedman M. Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. J Int Neuropsychol Soc 1998b;4:137–43.
19 Pakhomov SV, Hemmy LS. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. CORTEX 2013.
20 Raoux N, Amieva H, Le Goff M, Auriacombe S, Carcaillon L, Letenneur L, et al. Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study. CORTEX 2008;44:1188–96.
21 Clark DG, Kapur P, Geldmacher DS, Brockington JC, Harrell L, DeRamus TP, et al. Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease. CORTEX 2014;55:202–18.
22 Pakhomov SV, Jones DT, Knopman DS. Language networks associated with semantic indices. Neuroimage 2014;104:125–37.
23 Seni G, Elder J. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Chicago: Morgan & Claypool: University of Illinois; 2010. 126.
24 Folstein MF, Folstein SE. "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.
25 Petersen RC, Doody R, Kurz A, Mohs RC, Morris JC, Rabins PV, et al. Current concepts in mild cognitive impairment. Arch Neurol 2001; 58:1985–92.
26 Breiman L. Random forests. Machine Learning 2001;45:2–32.
27 Cristiannini, N, Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. 2000, Cambridge, UK: Cambridge University Press.
28 Bishop CM. Pattern Recognition and Machine Learning. In: Jordan M, Kleinberg J, Schoelkopf B, eds. Information Science and Statistics. New York: Springer; 2000.
29 Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. in 7th Python in Science Conference (SciPy2008). 2008. Pasadena, CA.
30 Clark DG, Wadley VG, Kapur P, DeRamus TP, Singletary B, Nicholas AP, et al. Lexical factors and cerebral regions influencing verbal fluency performance in MCI. Neuropsychologia 2014; 54:98–111.
31 Apostolova LG, Lu P, Rogers S, Dutton R, Hayashi K, Toga AW, et al. 3D mapping of language networks in clinical and pre-clinical Alzheimer's disease. Brain Lang 2008;104:33–41.
32 Thompson PM, Apostolova LG. Computational anatomical methods as applied to ageing and dementia. Br J Radiol 2007;80(Spec No 2):S78–91.
33 Thompson PM, Hayashi KM, de Zubicaray G, Janke A, Rose SE, Semple J, et al. Dynamics of gray matter loss in Alzheimer's disease. J Neurosci 2003;23:994–1005.
34 Narr KL, Thompson PM, Sharma T, Moussai J, Blanton R, Anvar B, et al. Three-dimensional mapping of temporo-limbic regions and the lateral ventricles in schizophrenia: gender effects. Biol Psychiatry 2001;50:84–97.
35 Pantel J, O'Leary DS, Cretsinger K, Bockholt HJ, Keefe H, Magnotta VA, et al. A new method for in vivo volumetric measurement of the human hippocampus with high neuroanatomical accuracy. Hippocampus 2000;10:752–8.
36 Apostolova LG, Beyer M, Green AE, Hwang KS, Morra JH, Chou Y, et al. Hippocampal, caudate, and ventricular changes in Parkinson's disease with and without dementia. Mov Disord 2010;25:687–95.
37 Apostolova LG, Green AE, Babakchanian S, Hwang KS, Chou Y, Toga AW, et al. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment, and Alzheimer's disease. Alzheimer Dis Assoc Disord 2012;26:17–27.
38 Ahmed S, Haigh AM, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 2013;136:3727–37.
39 Garrard P, Maloney LM, Hodges JR, Patterson K. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain 2005;128:250–60.
40 Garrard P, Lambon Ralph MA, Watson PC, Powis J, Patterson K, Hodges JR. Longitudinal profiles of semantic impairment for living and nonliving concepts in dementia of the Alzheimer's type. J Cog Neurosci 2001;13:892–909.
41 Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging 2011;32:e19–27.
42 Wee CY, Yap PT, Shen D, the Alzheimer's Disease Neuroimaging Initiative. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum Brain Mapp 2013; 34:3411–25.