NeuroImage: Clinical 13 (2017) 415–427

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease

Elaheh Moradi[a,1,*], Ilona Hallikainen[b], Tuomo Hänninen[c], Jussi Tohka[d,e,f], Alzheimer's Disease Neuroimaging Initiative[2]

[a] *Institute of Biosciences and Medical Technology, University of Tampere, Tampere, Finland*
[b] *University of Eastern Finland, Institute of Clinical Medicine, Department of Neurology, Kuopio, Finland*
[c] *Neurocenter, Neurology, Kuopio University Hospital, Kuopio, Finland*
[d] *Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Leganes, Spain*
[e] *Instituto de Investigación Sanitaria Gregorio Marañon, Madrid, Spain*
[f] *University of Eastern Finland, AI Virtanen Institute for Molecular Sciences, Kuopio, Finland*

ARTICLE INFO

Article history:
Received 12 July 2016
Received in revised form 25 November 2016
Accepted 11 December 2016
Available online 18 December 2016

Keywords:
Alzheimer's disease
Elastic net
Penalized regression
Magnetic resonance imaging
Rey's Auditory Verbal Learning Test

ABSTRACT

Rey's Auditory Verbal Learning Test (RAVLT) is a powerful neuropsychological tool for testing episodic memory, which is widely used for the cognitive assessment in dementia and pre-dementia conditions. Several studies have shown that an impairment in RAVLT scores reflect well the underlying pathology caused by Alzheimer's disease (AD), thus making RAVLT an effective early marker to detect AD in persons with memory complaints. We investigated the association between RAVLT scores (RAVLT Immediate and RAVLT Percent Forgetting) and the structural brain atrophy caused by AD. The aim was to comprehensively study to what extent the RAVLT scores are predictable based on structural magnetic resonance imaging (MRI) data using machine learning approaches as well as to find the most important brain regions for the estimation of RAVLT scores. For this, we built a predictive model to estimate RAVLT scores from gray matter density via elastic net penalized linear regression model. The proposed approach provided highly significant cross-validated correlation between the estimated and observed RAVLT Immediate ($R = 0.50$) and RAVLT Percent Forgetting ($R = 0.43$) in a dataset consisting of 806 AD, mild cognitive impairment (MCI) or healthy subjects. In addition, the selected machine learning method provided more accurate estimates of RAVLT scores than the relevance vector regression used earlier for the estimation of RAVLT based on MRI data. The top predictors were medial temporal lobe structures and amygdala for the estimation of RAVLT Immediate and angular gyrus, hippocampus and amygdala for the estimation of RAVLT Percent Forgetting. Further, the conversion of MCI subjects to AD in 3-years could be predicted based on either observed or estimated RAVLT scores with an accuracy comparable to MRI-based biomarkers.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by memory deficit, which is followed by problems in other cognitive domains that cause a severe decline in the usual level of functioning. The progressive episodic memory impairment characteristic to AD is best measured by neuropsychological testing. This is evident in recent diagnostic recommendations, which highlight the significance of standardized neuropsychological testing as well as the supportive role of biological evidence for AD pathology (Dubois et al., 2010; Jack et al., 2011; American Psychiatric Association, 2013). Rey's auditory verbal learning test (RAVLT) is a well-known measure of episodic memory, and in previous studies it has had a significant role in early diagnosis of AD (Estévez-González et al., 2003) as well as it has been demonstrated to be useful in differentiating AD from psychiatric disorders (Ricci et al., 2012; Schoenberg et al., 2006; Tierney et al., 1996). In particular, Estévez-González et al. (2003) suggested inclusion of the RAVLT to

[*] Corresponding author.
E-mail address: elaheh.moradi@uta.fi (E. Moradi).

[1] A part of this work was performed while Elaheh Moradi was with Department of Signal Processing, Tampere University of Technology, Finland.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

http://dx.doi.org/10.1016/j.nicl.2016.12.011

the cognitive test battery used in evaluation and early detection of AD. Moreover, Balthazar et al. (2010) indicated of the importance of RAVLT in a clinical setting for discriminating normally aging subjects from mild cognitive impairment (MCI) and AD subjects.

Recently revised diagnostic criteria and recommendations emphasize the importance of early diagnosis of AD (Dubois et al., 2010; McKhann et al., 2011; American Psychiatric Association, 2013). The disease processes leading to AD are known to start while individuals are still cognitively normal and may precede clinical symptoms by years or decades (Jack et al., 2010; Adaszewski et al., 2013). Reflecting this and the call for the biological evidence for AD diagnosis, several AD specific biomarkers have been identified, including multivariate patterns of structural brain atrophy measured by magnetic resonance imaging (MRI) (Moradi et al., 2015; Bron et al., 2015; Salvatore et al., 2015; Coupé et al., 2015; Eskildsen et al., 2013; Wee et al., 2013). MRI-based biomarkers have the advantages of being non-invasive and widely available.

However, integrating neuropsychological information and brain atrophy biomarkers might be extremely valuable for early diagnosis. In particular, we have previously shown that integrating cognitive and functional measures to brain atrophy pattern from MRI significantly improved the prediction performance of conversion to AD in mild cognitive impairment (MCI) patients as compared to using either modality alone (Moradi et al., 2015). Among cognitive and functional measures considered, RAVLT was the most important measure in the prediction model (as determined by the out-of-bag variable importance score in the Random Forest classifier (Breiman, 2001; Liaw and Wiener, 2002), which, in part, explains our interest towards RAVLT.

In order to enhance possibilities to early detection of AD and tracking disease progression, it is important to explore the association between cognitive functions and the pathological mechanisms of AD. The essential role of medial temporal lobe structures, especially hippocampus, for episodic memory has been known for long (Squire and Wixted, 2011). The studies of recent years have provided data on neurobiology of memory and learning and on the neurobiological changes of AD, but many aspects still remain unclear (Masdeu et al., 2012; Jeong et al., 2015). The great majority of machine learning based AD studies have been focused on either classification of AD and healthy subjects (Magnin et al., 2009; Beheshti et al., 2016) or predicting conversion to AD in MCI patients (Moradi et al., 2015; Eskildsen et al., 2013) using different neuroimaging techniques. However, the relationships between AD related brain atrophy and decline in cognitive abilities are less studied. In the current study, we aim to analyze the relation between AD related structural change within the brain and RAVLT measures. Particularly, we aim to predict RAVLT scores from MRI based gray matter density images by applying elastic net linear regression forming a multivariate brain atrophy pattern predicting the RAVLT score. According to previous studies (Khundrakpam et al., 2015; Bunea et al., 2011; Carroll et al., 2009) elastic net linear regression is well suited for learning predictive patterns among high dimensional neuroimaging data with many relevant predictors that are correlated with each other. Additionally, this approach offers an interpretable model by automatically selecting a sparse pattern of relevant voxels for predicting RAVLT, thus providing the possibility of finding the brain regions most strongly contributing to the prediction of RAVLT scores.

The association between AD related changes in brain structure and various cognitive measures of dementia (Mattis Dementia Rating Scale (DRS), Alzheimer's Disease Assessment Scale-cognitive subtest (ADAS-Cog), Mini-mental state examination (MMSE) and RAVLT-Percent Retention) was previously studied by Stonnington et al. (2010) based on pattern analysis on gray matter voxel-based morphometry maps. Their results indicated that DRS, ADAS-cog and MMSE measures could be well estimated based on brain structure. However, the accuracy of predicting the RAVLT percent retention score based on MRI was much more modest with a dataset that included a continuum of subjects who were cognitively normal and persons with MCI or AD. This could reflect the small number of subjects or the specific nature of the machine learning method used, which might not be the best possible for learning the associations between MRI and a score related to a specific aspect of cognition (episodic memory) rather than to cognitive ability in general. More recently, the relationship between MRI and RAVLT scores was investigated by Wang et al. (2011). However, as they averaged grey matter density, cortical thickness and subcortical volumetry from MRI into the total of 144 regional measures, they did not probe the relationship between a high-dimensional atrophy pattern and RAVLT. Furthermore, these atlas-based averaging strategies of high-dimensional MRI data may be detrimental to the predictive accuracy of machine learning analysis (Khundrakpam et al., 2015). Additionally, as Wang et al. (2011) used root mean square error (RMSE) measure to report the predictive accuracy and provided no p-values for RMSE, it is difficult to put the prediction accuracy into proper context.

In this report, we used whole brain gray matter density maps for predicting different RAVLT measures. We analyzed the relationship between RAVLT measures and AD related structural changes within the brain by considering a large ADNI dataset of over 800 subjects ranging from severe AD to age-matched healthy subjects. We also investigated the relationship between AD conversion prediction and the observed and MRI-estimated RAVLT measures to highlight the potential clinical implications of the method. We studied two RAVLT summaries - RAVLT Immediate and RAVLT Percent Forgetting. These summary scores highlight different aspects of episodic memory, namely learning (immediate) and delayed memory (percent forgetting), which both are essential aspects of AD.

2. Materials and methods

2.1. ADNI data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

We used the same dataset as Moradi et al. (2015), but excluded subjects with missing RAVLT scores; the subject demographics are presented in Table 1. For RAVLT Immediate (Percent forgetting), the dataset consisted of 186 (180) AD subjects, 226 (226) NC (normal control) subjects and 394 (393) MCI subjects. The diagnostic and inclusion/exclusion criteria is specified in Petersen et al. (2010) and roster IDs of the subjects are listed in Supplementary material. Of the 394 (393) MCI subjects, 164 subjects were grouped as progressive MCI (pMCI) if diagnosis was MCI at baseline but conversion to AD was reported after baseline within 1, 2 or 3 years, and without reversion to MCI or NC at any available follow-up (0–96 months). 100 subjects were grouped as stable MCI (sMCI) if diagnosis was MCI at all available time points (0–96 months), but at least for 36 months. The remaining 130 (129) MCI subjects were grouped as unknown MCI (uMCI), if diagnosis was MCI at baseline but the subjects were missing a diagnosis at 36 months from the baseline or the diagnosis was not stable at all available time points. The labeling of MCI patients was based on the 3-year cut-off period that was decided based on the length of follow-up for the original ADNI-1 project (Moradi et al., 2015). For estimating the RAVLT Percent Forgetting score, we

Table 1
Subject demographics. RAVLT-Immediate is abbreviated as RAVLT-IR and RAVLT-Percent Forgetting is abbreviated as RAVLT-PF.

Diagnosis	No of subjects IR/PF	Age, mean (std) IR/PF	RAVLT IR mean (std)	RAVLT PF mean (std)
AD	186/180	75.28 (7.53)/75.39 (7.52)	23.20 (7.74) Range: 0–42	90.30 (18.86) Range: 10–100
MCI	394/393	74.91 (7.33)/74.90 (7.34)	30.58 (9.11) Range: 11–68	68.15 (30.83) Range: 0–100
NC	226/226	75.97 (5.05)/75.97 (5.05)	43.32 (9.11) Range: 16–69	35.04 (33.65) Range: 0–100

excluded 3 AD subjects with the score of zero as outliers (roster IDs of these three were 724, 1184, and 1253). In addition, there are many subjects (129 AD, 77 pMCI, 17 sMCI, 38 uMCI and 8 NC subjects) with percent forgetting score of 100%, who did not recall any words during the delayed trial. However, these subjects cannot be considered as outliers. The RAVLT Percent Forgetting of 100% can be considered typical for AD and pMCI subjects and, while not typical, this is not unusual for sMCI subjects. For 8 normal controls, this is an unusual score, which, however, could be explained by a number of factors such as nervousness in the testing situation.

For predicting RAVLT scores all MCI subjects with available RAVLT scores were included regardless of availability of information about the AD conversion as this is not required in predicting RAVLT scores.

2.2. RAVLT score

Rey's Auditory Verbal Learning Test (RAVLT) (Rey, 1964) is a powerful neuropsychological tool that is used for assessing episodic memory by providing scores for evaluating different aspects of memory. The RAVLT is sensitive to verbal memory deficits caused by a variety of neurological diseases such as AD (Schoenberg et al., 2006;Balthazar et al., 2010;Estévez-González et al., 2003). Tierney et al. (1996) and Estévez-González et al. (2003) have shown that the RAVLT score is an effective early marker to detect AD in persons with memory complaints.

Briefly, the RAVLT consists of presenting a list of 15 words across five consecutive trials. The list is read aloud to the participant, and then the participant is immediately asked to recall as many as words as he/she remembers. This procedure is repeated for 5 consecutive trials (Trials 1 to 5). After that, a new list (List B) of 15 new words is read to the participant, who then is immediately asked to recall the words. After the List B trial, the examiner asks participant to recall the words from the first list (Trial 6). After 30-minutes of interpolated testing (timed from the completion of List B recall), the participant is again asked to recall the words from the first list (delayed recall).

Different summary scores are derived from raw RAVLT scores. These include RAVLT Immediate (the sum of scores from 5 first trials (Trials 1 to 5)), RAVLT Learning (the score of Trial 5 minus the score of Trial 1), RAVLT Forgetting (the score of Trial 5 minus score of the delayed recall) and RAVLT Percent Forgetting (RAVLT Forgetting divided by the score of Trial 5). We use naming of the ADNI merge table[3] for these summary measures. We investigated the relationship between MRI measures and RAVLT cognitive test scores by estimating the RAVLT Immediate and RAVLT Percent Forgetting from the gray matter density. These two summary scores were selected since they highlight different aspects of episodic memory, learning (RAVLT Immediate) and delayed memory (RAVLT Percent forgetting), essential to AD and previous studies (Estévez-González et al., 2003; Wang et al., 2011; Gomar et al., 2014; Moradi et al., 2015) have indicated strong relationships between these two RAVLT measures and Alzheimer's disease. For example, Estévez-González et al. (2003) concluded that the most reliable RAVLT measures for AD detection are RAVLT Immediate, a score of zero at the delayed recall and the RAVLT percent forgetting. Particularly, we stress that RAVLT percent forgetting, which is a measure of delayed memory that takes into account the relationship of immediately and delayed recalled words is equivalent of RAVLT percent retention considered by Stonnington et al. (2010).

2.3. MRI and image processing

The downloaded MRIs were acquired with T1-weighted MP-RAGE sequence at 1.5 Tesla, typically with 256 × 256 × 170 voxels with the voxel size of approximately 1 mm × 1 mm × 1.2 mm. The MRIs were downloaded as raw images converted to the NIFTI format. As described by Gaser et al. (2013), Moradi et al. (2015) preprocessing of the T1-weighted images was performed using the SPM8 package[4] and the VBM8 toolbox[5] , running under MATLAB. All T1-weighted images were corrected for bias-field inhomogeneties, then spatially normalized and segmented into gray matter (GM), white matter, and cerebrospinal fluid (CSF) within the same generative model (Ashburner and Friston, 2005). The dimension after the spatial normalization was 181 × 217 × 181 with 1 mm^3 voxels and the template used for the spatial normalization was the SPM8 version of the ICBM152 atlas (the linear registration version) provided by MNI[6] . The segmentation procedure was further extended by accounting for partial volume effects (Tohka et al., 2004), by applying adaptive maximum a posteriori estimations (Rajapakse et al., 1997), and by using an hidden Markov random field model (Cuadra et al., 2005) as described previously (Gaser, 2009). This procedure resulted in maps of tissue fractions of WM and GM. Only the GM images were used in this work. Following the pipeline proposed by (Franke et al., 2010), the GM images were processed with affine registration and smoothed with 8-mm full-width-at-half-maximum smoothing kernels. After smoothing, images were resampled to 4 mm isotropic spatial resolution. This procedure generated, for each subject, 29,852 aligned and smoothed GM density values that were used as MRI features.

2.4. Machine learning framework

We applied elastic net linear regression (ENLR) (Zou and Hastie, 2005) for the estimation of RAVLT score (RAVLT Immediate and RAVLT Percent forgetting) from MRI measurements. Due to the high dimensionality of MRI data, the number of predictor variables (voxels) is greater than the number of subjects. Therefore, the ordinary least squares linear regression cannot be applied. However, regularization approaches are effective in solving underconstrained

[3] http://adni.bitbucket.org/adnimerge.html.

[4] http://www.l.ion.ucl.ac.uk/spm.

[5] http://dbm.neuro.uni-jena.de.

[6] http://nist.mni.mcgill.ca/?p=798.

problem like this in a statistically principled manner. In particular, we used the elastic net penalty as regularizer. The ENLR provides spatially sparse model by performing simultaneously variable selection and model estimation, thus providing a subset of voxels relevant to predict RAVLT scores. Further, ENLR possesses so called grouping effect meaning that correlated predictors are selected simultaneously. The number of voxels that are included in the regression model is controlled by a regularization parameter λ, which is typically, and also in this work, selected by cross-validation. A more detailed description of ENLR is provided in Appendix A.

To compare the performance of ENLR approach, we additionally applied relevance vector regression (RVR) for estimation of RAVLT scores as this was the machine learning approach used by Stonnington et al. (2010). The RVR (Tipping, 2001) is a pattern recognition method that uses Bayesian inference to obtain sparse regression models. We used kernelized RVR with the linear kernel as Stonnington et al. (2010) and also RVR without kernelization. Similarly to ENLR, RVR provides a sparse solution with only a subset of predictors contributing to the final model. However, having a sparse predictive model in a kernel space does not provide easily interpretable prediction model in a voxel space, since enforcing sparsity in the kernel space does not result on a sparse solution in the original feature space (Khundrakpam et al., 2015).

We considered different datasets of subjects in our experiments. The main dataset included all subjects, i.e., AD and MCI patients and NC subjects. In this way, the dataset included a contiguous range of RAVLT scores. The range of RAVLT Immediate in this dataset was from 0 to 69 and the range of RAVLT Percent Forgetting was from 0 to 100. Secondarily, we included only two groups of subjects for learning the regression model and predicting RAVLT scores. This resulted in 3 distinct datasets with different subject characteristics (1. AD and NC subjects, 2. AD and MCI subjects and 3. NC and MCI subjects). Finally, we included only one group of subjects (only for AD and MCI groups) and repeated the experiments.

2.5. Implementation and performance evaluation

For the performance evaluation of the model and estimation of the regularization parameter λ, we used two nested and stratified cross-validation loops (10-fold for each loop) (Ambroise and McLachlan, 2002; Huttunen et al., 2012)[7] . The number of folds was selected to be 10 because this is typically recommended compromise (Hastie et al., 2011; Arlot et al., 2010). First, an external 10-fold cross-validation was implemented in which the dataset were randomly divided into 10 subsets. At each step, a single subset was used for testing and remaining subsets were used for training. The training set was used to train the elastic net regression model. We re-divided the training set into 10-folds for finding the optimal λ for the model. The optimal λ was selected according to the mean absolute error (*MAE*) across the inner 10-fold cross-validation loop. Note that the test sets in the external cross-validation loop were used only for evaluating the model. The performance of the model was characterized using the (cross-validated) Pearson correlation coefficient (R), mean absolute error (*MAE*) and the coefficient of determination[8] (Q^2) between estimated and true RAVLT scores in the test set. Three different metrics are reported to provide complementary information. Cross-validated correlation is simple to interpret, but it can hide the bias in the predictions, which are made apparent by Q^2-value. MAE provides the prediction errors in the equal scale with the original scale of the RAVLT scores. The reported metrics in the Results section are the averages over 100 nested 10-fold CV runs in order to minimize the effect of the random variation in the division of the data into different folds. To compare the performance of two learning algorithms, we computed a p-value for the 100 correlation scores with a permutation test. For computing p-values associated with the correlation coefficient between the observed and estimated values, we used a permutation test (Anderson and Robinson, 2001) and, for computing the 95% confidence intervals of the correlation coefficient, we used bootstrap on the run with the median correlation score across 100 cross-validation runs. For evaluating the power of RAVLT scores in discriminating between pMCI (progressive MCI) and sMCI (stable MCI) subjects, we used AUC (area under the receiver operating characteristic curve) measure (Hanley and McNeil, 1982) and for comparing AUCs we used StaR tool (Vergara et al., 2008).

The ENLR was implemented with the GLMNET library (Friedman et al., 2010)[9] , and the RVR was implemented with the "SparseBayes" package (Tipping et al., 2003)[10] .

3. Results

3.1. Prediction of RAVLT scores

We estimated RAVLT scores, both RAVLT Immediate and RAVLT Percent Forgetting, from MRI data. The cross-validated accuracies of these estimations with different methods (ENLR, KRVR, RVR) and different subject sets are listed in Table 2.

3.1.1. Accuracy of estimated RAVLT scores with all subjects

As shown in Table 2, the RAVLT scores estimated by ENLR were the most accurate ones. The correlation score (R) of ENLR was significantly better compared to KRVR ($p < 0.0001$) and RVR ($p < 0.0001$) approaches when using the whole dataset. In addition, R was highly significant using all three approaches and for both summary scores as revealed by the permutation test on the run with the median correlation score across 100 cross-validation runs ($p < 0.0001$ in all cases). The 95% bootstrap confidence intervals (CIs) for the correlation score for the estimation of RAVLT Immediate were as follows: ENLR: [0.45, 0.55], KRVR: [0.41,0.51], RVR: [0.21,0.33]; and, for the estimation of RAVLT Percent Forgetting, the 95% bootstrap CIs were as follows: ENLR: [0.37,0.48], KRVR: [0.35, 0.47], RVR: [0.23, 0.35]. The scatter plots between the estimated and observed RAVLT scores based on ENLR and KRVR approaches are illustrated in Fig. 1. The scatter plots corresponding to the estimated values by using RVR approach are provided in the supplement.

We investigated the effect of age-correction on the performance of the prediction model by estimating normal aging effects on MRI data in NC subjects of the training set and removing it from MRI data of all subjects as proposed in (Moradi et al., 2015). With the age correction step for the estimation of RAVLT Immediate using the ENLR approach, the average correlation score increased from 0.50 to 0.51 ($p < 0.001$), the average MAE decreased from 7.86 to 7.80 and the average Q^2 increased from 0.25 to 0.26. For estimation of RAVLT Percent Forgetting with age corrected MRI data, the average correlation score increased from 0.43 to 0.46 ($p < 0.001$), the average MAE decreased from 25.53 to 25.18 and the average Q^2 increased from 0.185 to 0.21.

[7] The Matlab code used for constructing stratified cross-validation folds for regression is available at https://github.com/jussitohka/general_matlab.

[8] The Q^2 provides a measure of how well out-of-training set RAVLT scores are predictable by the learned model (http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics). It is defined as $Q^2 = 1 - \frac{\sum_{i=1}^{N}(s_i - \hat{s}_i)^2}{\sum_{i=1}^{N}(s_i - \bar{s})^2}$, where $\hat{s}_i$ is the estimated RAVLT for subject i, s_i is the true RAVLT score for subject i, and $\bar{s}$ is mean of the true RAVLT scores. Q^2 is bounded above by 1 but is not bounded from below. Note that Q^2 does not equal R^2, i.e., the correlation squared, but the Q^2 value can never exceed R^2, see the methods supplement of (Moradi et al., 2016).

[9] http://web.stanford.edu/~hastie/glmnet_matlab/.

[10] http://www.miketipping.com/sparsebayes.htm.

Table 2
The generalization performance based on correlation score (R), coefficient of determination (Q^2) and mean absolute error (MAE) for different experiments. *** means that the value was not meaningful, because Q^2 values were below −100 and MAE values were above 100. The values are averages across 100 CV runs. The values in parentheses show the standard deviations across 100 CV runs. RAVLT-Immediate is abbreviated as RAVLT-IR and RAVLT-Percent Forgetting is abbreviated as RAVLT-PF.

Data		RAVLT IR	RAVLT IR	RAVLT IR	RAVLT PF	RAVLT PF	RAVLT PF
		ENLR	KRVR	RVR	ENLR	KRVR	RVR
AD, MCI, NC	R	0.50 (0.007)	0.46(0.01)	0.27 (0.02)	0.43 (0.01)	0.41(0.01)	0.28 (0.02)
	Q2	0.25 (0.007)	0.17 (0.01)	−0.71 (0.06)	0.185 (0.01)	0.14 (0.01)	−0.645 (0.07)
	MAE	7.86 (0.043)	8.21 (0.08)	11.90 (0.23)	25.53 (0.18)	26.65 (0.18)	34.52(0.82)
AD, NC	R	0.61 (0.008)	0.53(0.01)	0.38 (0.03)	0.53 (0.01)	0.50 (0.01)	0.32 (0.03)
	Q2	0.37 (0.01)	0.24 (0.02)	−0.37 (0.07)	0.28 (0.01)	0.23 (0.02)	−0.56 (0.08)
	MAE	8.30 (0.07)	9.11 (0.13)	12.23 (0.35)	25.33(0.16)	25.75 (0.37)	35.58 (1.11)
AD, MCI	R	0.39 (0.01)	0.32(0.01)	0.21 (0.03)	0.29(0.02)	0.255(0.02)	0.15(0.03)
	Q2	0.15 (0.01)	−0.03 (0.02)	−0.78 (0.08)	0.08 (0.01)	−0.05 (0.03)	−0.93 (0.08)
	MAE	6.57 (0.04)	7.26 (0.09)	9.76 (0.24)	23.39(0.14)	24.52(0.38)	32.60 (0.76)
MCI, NC	R	0.43 (0.01)	0.41(0.01)	0.26(0.03)	0.32 (0.02)	0.32 (0.01)	0.19(0.03)
	Q2	0.18 (0.01)	0.10 (0.02)	−0.70 (0.10)	0.09 (0.02)	0.06 (0.01)	−0.88 (0.08)
	MAE	67.88 (0.06)	8.21(0.09)	11.34(0.38)	26.58 (0.21)	26.49(0.19)	36.11 (0.83)
AD	R	0.32 (0.03)	0.28(0.02)	0.08 (0.05)	−0.14 (0.06)	0.06 (0.03)	−0.09 (0.06)
	Q2	0.10 (0.02)	−0.02 (0.03)	−1.08 (0.16)	−0.03 (0.02)	−0.31 (0.05)	−1.48 (0.22)
	MAE	5.75 (0.07)	6.22 (0.11)	8.84 (0.37)	14.08 (0.15)	16.17 (0.35)	22.8 (1.12)
MCI	R	0.15 (0.02)	−0.03(0.03)	0.06 (0.06)	0.16 (0.02)	−0.01 (0.02)	0.05 (0.04)
	Q2	0.02 (0.01)	***	***	0.02 (0.01)	***	−1.11 (0.14)
	MAE	6.92 (0.035)	***	***	26.07 (0.15)	***	33.65 (1.19)

3.1.2. Top predictors for RAVLT scores

Since we standardized the data before applying ENLR, the absolute value of each regression coefficient provides the importance of the corresponding predictor in the predictive model. Therefore, we computed the importance of each brain region based on the maximum value of the average magnitudes of regression coefficients. The magnitude of standardized regression coefficients was averaged across 100 different 10-fold CV iterations. The top predictors (brain regions) for estimation of RAVLT scores in the ENLR model are listed in Table 3 (RAVLT Immediate) and Table 4 (RAVLT Percent Forgetting). We considered only the maximum of the average magnitudes within a region to discount for poor predictors within a region. To compute the 95% confidence intervals (CIs) for the maximum of average magnitudes of regression coefficients, we calculated first the 2.5% and 97.5% percentiles of magnitudes of regression coefficients for each voxel within 100 runs of 10-fold CV, and then took the maximum values of these as the lower and upper bound of the CI. The lower CI limit larger than zero provides strong evidence that the region in the question contributes to the prediction model independent of the training set used. In addition, we computed the selection probability for each voxel across 100 different 10-fold CV runs (see Fig. 2).

3.1.3. Accuracy of estimated RAVLT scores with reduced subject sets

Removing MCI subjects significantly improved the performance of the estimation (see Table 2, the first and second rows, the improvement in R was significant with all three methods and both scores ($p < 0.0001$)). Albeit the predictive performance improved in terms of correlation score and coefficient of determination, the MAE increased in all experiments.

Excluding either the NC or AD group from the dataset notably decreased the prediction performance when comparing to that of using all subjects (see Table 2, first, third and forth rows). The decline in the performance of model was highly significant ($p < 0.0001$) in all experiments. As the results show, removing either AD or NC groups and including subjects from the groups with more similarities such as “AD and MCI” or “NC and MCI” rendered the prediction problem more challenging.

We experimented with using a single group of subjects for learning and evaluating of the model. The results are presented in the last two rows of the Table 2. As it was expected, the estimation of RAVLT scores with a single group of subjects proved to be a difficult problem due to lack of significant differences in the AD related structural changes within subjects of a single group. However, even within MCI and AD groups, the correlation between the estimated and observed RAVLT Immediate score was significant when using ENLR for prediction. With the AD group, the estimation of RAVLT percent forgetting was not successful with any method. However, ENLR could estimate the RAVLT percent forgetting within the MCI group, where the correlation was low but significant.

The scatter plots of the estimated and observed RAVLT scores of the CV run with the median R within 100 computation times, with the proposed approach for different experiments are illustrated in Fig. 3. The scatter plots corresponding to the KRVR and RVR approaches are provided in the supplement.

3.2. AD conversion prediction based on RAVLT measures

We studied the use of RAVLT Immediate and RAVLT Percent forgetting for predicting conversion to AD in MCI patients. For this, we classified subjects with MCI as pMCI (progressive MCI) if the subject converted to AD within 1, 2 or 3 years follow-up without reversion to MCI or NC at any available follow-up (0–96 months), sMCI (stable MCI) if the diagnosis was MCI at all available time points (0–96 months), but at least for 36 months and uMCI (unlabeled MCI) if the diagnosis was missing at 36 months from the baseline or the diagnosis was not stable at all available time points. The definition of these groups was the same as in our previous work (Moradi et al., 2015). We used only sMCI and pMCI subjects in order to evaluate the effectiveness of RAVLT scores (acquired at baseline) for predicting conversion to AD.

The baseline RAVLT scores differed significantly between the two MCI groups (pMCI and sMCI) in terms of both RAVLT Immediate ($p < 0.0001$) and RAVLT Percent Forgetting ($p < 0.0001$). The average RAVLT Immediate was 35.08 (standard deviation 9.69) in the sMCI group and 26.94 (standard deviation 6.19) in the pMCI group. The average RAVLT Percent Forgetting was 55.35 (standard deviation 30.91) in the sMCI group and 77.48 (standard deviation 27.99) in the pMCI group.

Furthermore, the longitudinal RAVLT measurements showed considerable changes during the 3 years follow-up in pMCI subjects while they were relatively stable in sMCI subjects as shown in Fig. 4, which is provided to confirm the close relationship between the

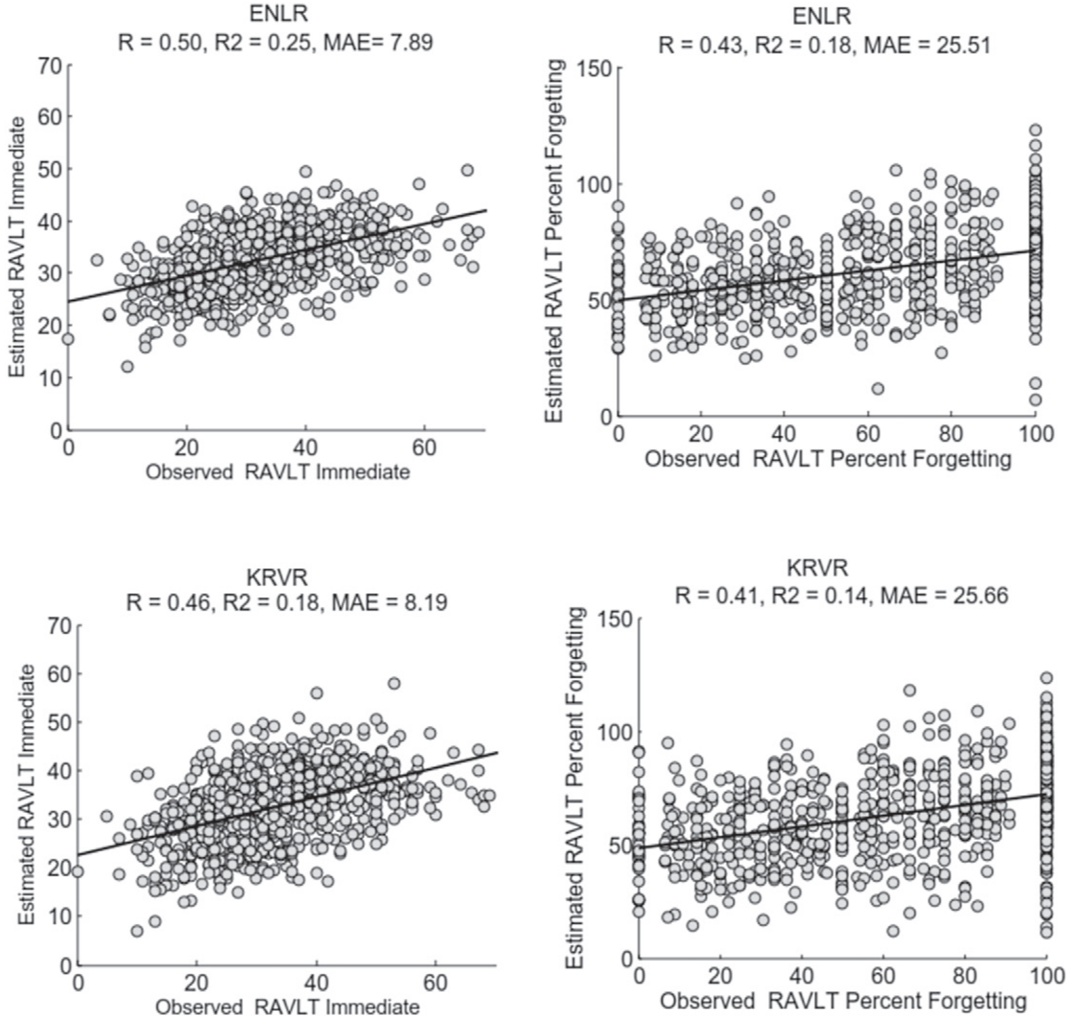

Fig. 1. Scatter plot for estimation of RAVLT Immediate (left) and RAVLT Percent Forgetting (right) using ENLR (top) and KRVR (bottom) with all available subjects, i.e., AD, MCI and NC subjects.

RAVLT scores and the suspected AD pathology. Interestingly, in the pMCI group, RAVLT Immediate displayed a more clear declining trajectory than the RAVLT percent forgetting.

Fig. 5 shows the ROC curves for discrimination of pMCI and sMCI subjects of observed baseline RAVLT scores and the estimated RAVLT scores. The estimated RAVLT scores were learned with all data (AD, MCI and NC subjects). From these estimated scores, we then selected the scores of pMCI and sMCI subjects in order to calculate AUC and plot the ROC curves. The AUC of observed RAVLT Immediate was 0.75 and the AUC of observed RAVLT Percent Forgetting was 0.71, thus indicating that these scores are powerful in predicting conversion to AD in MCI subjects. The AUC of estimated RAVLT Immediate was 0.72 (ENLR), 0.72 (KRVR) and 0.63 (RVR). The AUC of estimated RAVLT Percent Forgetting was 0.71 (ENLR), 0.69 (KRVR) and 0.60 (RVR). The difference between observed and estimated AUCs (based on either ENLR or KRVR) was 0.03 with the 95 % confidence interval (CI) of [−0.05, 0.11] for RAVLT Immediate. For RVR, the difference was 0.12 with the CI of [0.03, 0.21]. In the case of RAVLT Percent Forgetting, the difference between observed and estimated AUCs was 0.01 with the CI of [−0.07, 0.09] (ENLR), 0.02 with the CIs of [−0.07,0.10] (KRVR) and 0.12 with the CI of [0.03,0.20] (RVR). As the results indicate, the AUCs obtained based on estimated RAVLT scores using ENLR and KRVR methods were similar to AUCs obtained the observed RAVLT scores, i.e., estimated scores demonstrated similar power in the detection of AD conversion compared to the observed scores.

It is interesting to study whether pMCI and sMCI subjects can be more effectively separated if using both observed and estimated scores instead of only using observed scores. To test this, we trained a Gaussian plug-in classifier (Duda et al., 2012) using Matlab's classify function. The accuracy of the classifier was measured using 100 runs of 10 fold CV. The average accuracy when using both estimated and observed values for RAVLT Immediate (percent forgetting) was 0.75 (0.71). When using only the observed values the accuracy was 0.70 (RAVLT Immediate) and 0.67 (RAVLT percent forgetting)[11] . The performance improvement was significant in terms of run-wise applied permutation test ($p < 0.0001$). By combining the two observed RAVLT scores, the classification accuracy was 0.71. These results indicated that estimated and observed RAVLT scores contained different information that may be useful for early AD diagnosis.

[11] The difference to the AUCs reported above is because the resubstitution method, not dependent on any classifier, used to compute the values 0.75 and 0.71 above and the cross-validation based estimate (tied to the specific classifier) led to the AUCs of 0.70 and 0.67

Table 3
The top predictors for estimating RAVLT Immediate in all subjects (AD, MCI and NC). For each voxel, the average magnitude of the standardized regression coefficients (normalized with respect to the standard deviation of the response variable) across 100 different 10-fold CV iterations are calculated. The third column shows the number of voxels with the average magnitude greater than or equal to 0.01 in the corresponding region and the fourth and fifth columns show the maximum value of the average magnitude of regression coefficients and its CI within the region. The ranking is based on the maximum value of the average magnitude of regression coefficients in each region. The region definitions correspond to those of the AAL atlas and we abbreviate gyrus as G.

Region definition	Label	Number of voxels	Max weight	95 % CI for max weight
Middle temporal G right	86	3	0.05	[0.0185, 0.0784]
Amygdala right	42	4	0.04	[0.0123, 0.0815]
Insula left	29	2	0.04	[0.0076, 0.0645]
Hippocampus left	37	7	0.03	[0.003, 0.0637]
Sup temporal G left	81	2	0.03	[0.0075, 0.0637]
Calcarine right	44	1	0.03	[0.0007, 0.0641]
Thalamus right	78	1	0.03	[0.0074, 0.0540]
Inf parietal G left	61	1	0.02	[0.00004, 0.0479]
Middle cingulum left	33	2	0.02	[0, 0.0440]
Parahippocampal G left	39	1	0.02	[0, 0.0462]
Anterior cingulate left	31	2	0.02	[0, 0.0483]
Supplementary motor area left	19	1	0.02	[0, 0.0435]
Middle temporal G left	85	2	0.02	[0, 0.0469]
Middle frontal G right	8	1	0.02	[0, 0.0419]
Precuneus left	67	2	0.01	[0, 0.0358]
Lingual G right	48	1	0.01	[0, 0.0397]
Inf occipital G left	53	1	0.01	[0, 0.0360]
Inf frontal G, oper. right	12	1	0.01	[0, 0.0382]
Parahippocampal G right	40	1	0.01	[0, 0.0408]
Fusiform G left	55	1	0.01	[0, 0.0435]

4. Discussion

The purpose of the current study was to analyze the relationships between AD related structural changes within the brain with RAVLT cognitive measures in order to find how accurately RAVLT cognitive measures reflect the structural atrophy caused by AD. To this end, we build a predictive model to estimate RAVLT scores from gray matter density via elastic net penalized linear regression model by considering various datasets of subjects with different AD severity levels in the learning and evaluation procedures. The aim of considering different

Table 4
The top predictors for estimating RAVLT Percent Forgetting in all subjects (AD, MCI and NC). For each voxel, the average magnitude of the standardized regression coefficients (normalized with respect to the standard deviation of the response variable) across 100 different 10-fold CV iterations are calculated. The third column shows the number of voxels with the average magnitude greater than or equal to 0.01 in the corresponding region and the fourth column shows the maximum value of the average magnitude of regression coefficients with the region. The ranking is based on the maximum value of the average magnitude of regression coefficients within each region. The region definitions correspond to those of the AAL atlas and we abbreviate gyrus as G.

Region definition	Label	Number of voxels	Max weight	95 % CI for max weight
Angular G right	66	1	0.07	[0,0433, 0.0879]
Hippocampus right	38	1	0.05	[0.0208, 0.0855]
Hippocampus left	37	6	0.05	[0.0148, 0.0863]
Amygdala left	41	2	0.04	[0.0122, 0.0795]
Amygdala right	42	4	0.04	[0.0042, 0.0814]
Insula left	29	1	0.04	[0.002, 0.0683]
Parahippocampal G right	40	3	0.04	[0.0067, 0.0674]
Middle occipital G left	51	2	0.04	[0.0073, 0.0631]
Calcarine left	43	2	0.03	[0.0012, 0.0682]
Temporal pole, middle temporal G right	88	1	0.03	[0, 0.0702]
Sup temporal G right	82	1	0.03	[0, 0.0647]
Lingual G left	47	2	0.03	[0, 0.0644]
Inf occipital G right	54	2	0.03	[0, 0.0597]
Middle cingulum left	33	1	0.03	[0, 0.0528]
Sup frontal G, orb. left	5	1	0.02	[0, 0.0539]
Middle frontal G left	7	2	0.02	[0, 0.0523]
Temporal pole; sup temporal G left	83	2	0.02	[0, 0.0586]
Cerebellum-6 right	100	1	0.02	[0, 0.0465]
Middle frontal G right	8	2	0.02	[0, 0.0477]
Fusiform G left	55	1	0.02	[0, 0.0506]
Inf temporal G right	90	1	0.02	[0, 0.0450]
Inf frontal G, orb. right	16	1	0.02	[0, 0.0647]
Inf parietal G left	61	3	0.02	[0, 0.0450]
Cerebellum-6 left	99	1	0.02	[0, 0.0562]
Precuneus left	67	1	0.02	[0, 0.0434]
Olfactory G left	21	1	0.02	[0, 0.0535]
Parahippocampal G left	39	2	0.02	[0, 0.0443]
Thalamus right	78	2	0.01	[0, 0.0417]
Sup frontal G right	4	2	0.01	[0, 0.0378]
Sup frontal G left	3	1	0.01	[0, 0.0393]
Middle temporal G right	86	1	0.01	[0, 0.0422]

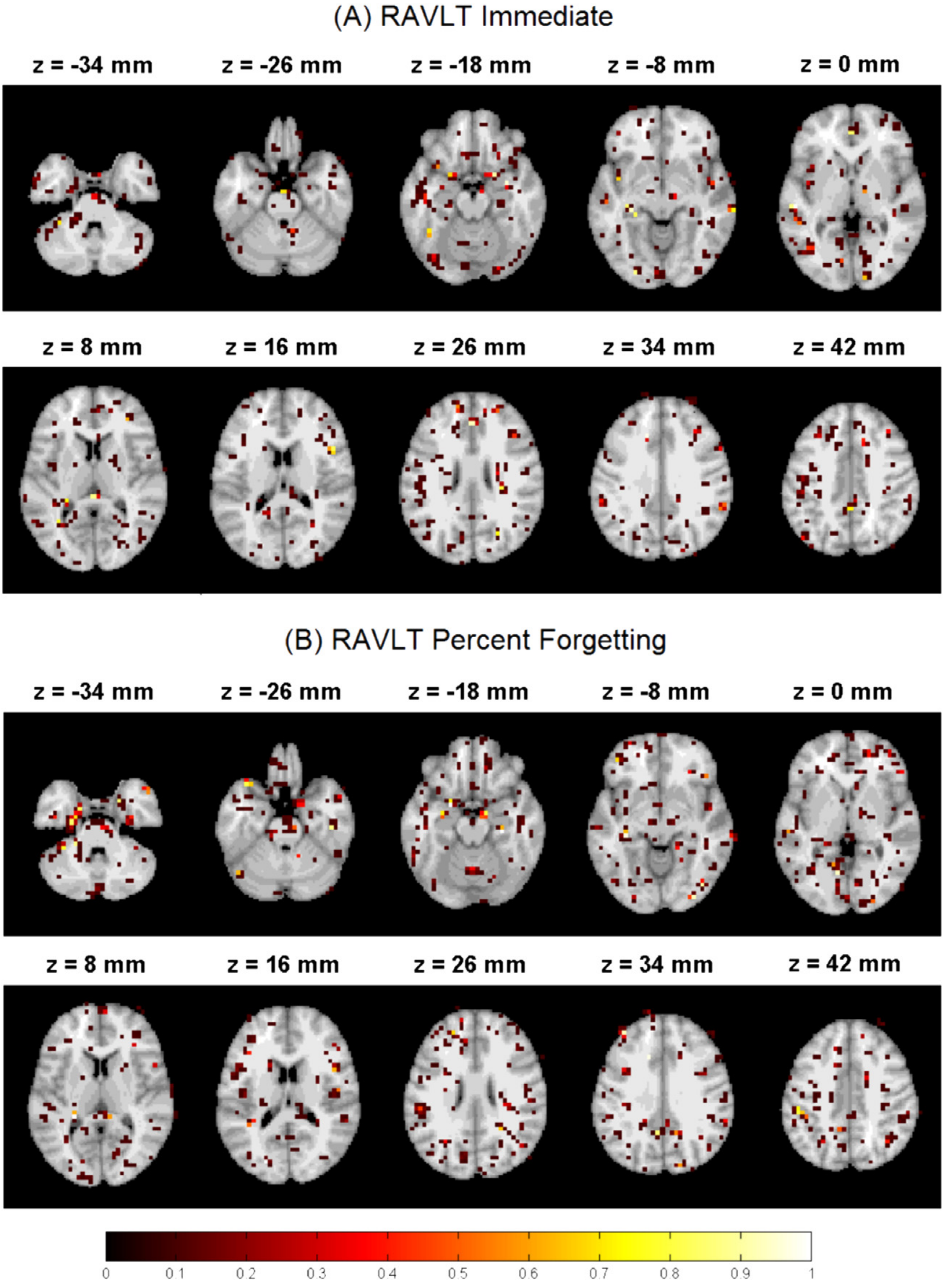

Fig. 2. The selection probability of voxels in the estimation RAVLT Immediate (A) and RAVLT Percent Forgetting (B) across 100 different 10-fold CV iterations. The images are displayed according to the neurological convention.

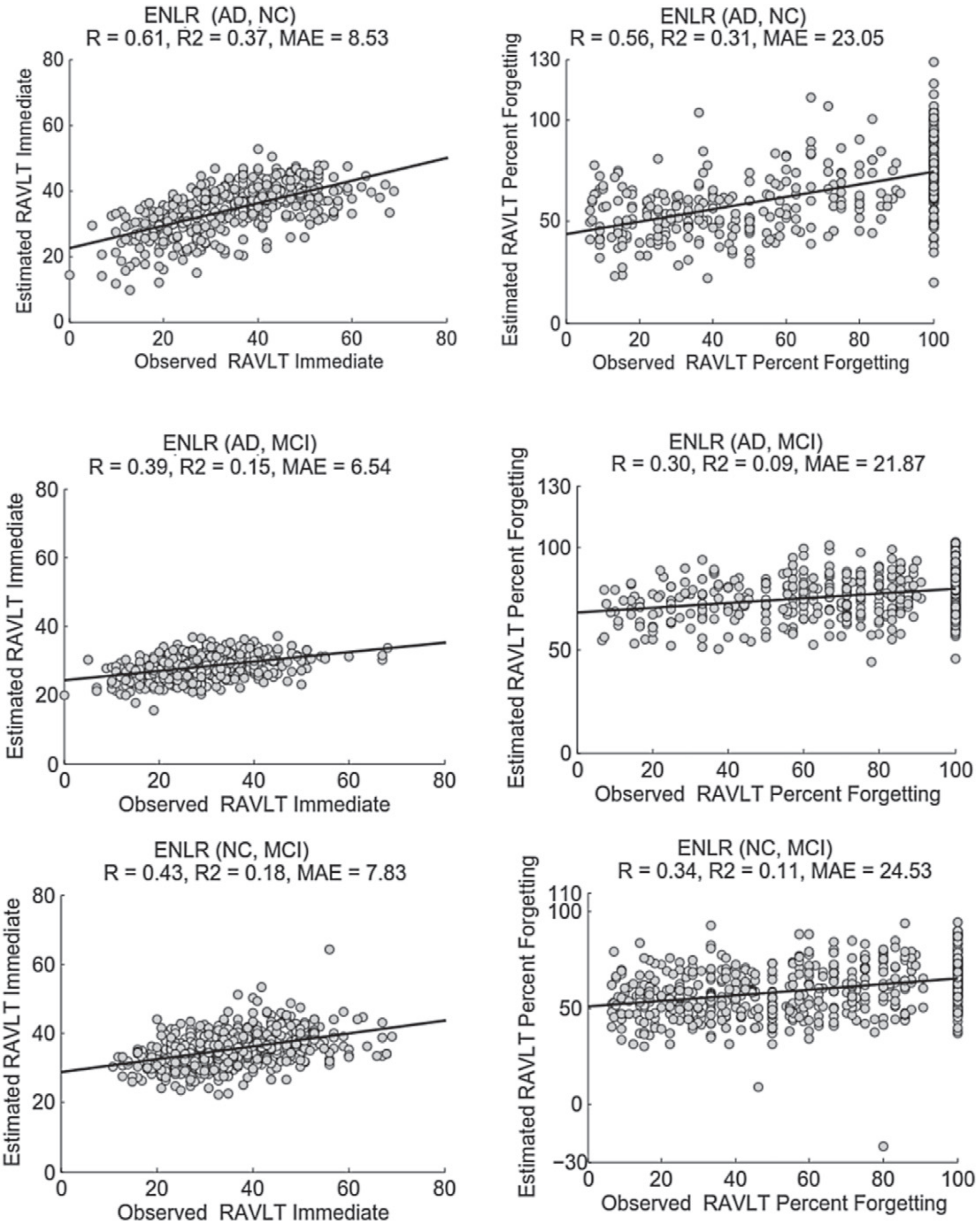

Fig. 3. Scatter plot for estimation of RAVLT Immediate (left) and RAVLT Percent Forgetting (right) based on ENLR using AD and NC subjects (top), AD and MCI subjects (middle) and NC and MCI subjects (bottom).

datasets with different levels of memory problems was to determine the dependency between the RAVLT performance and the dementia related atrophy. The results of the current study revealed strong association between information detected by RAVLT scores and AD related structural atrophy. As the results show (see Table 2), including subjects from similar groups such as "AD and MCI" or "NC and MCI" produced lower predictive performance compared to using groups of subjects with significant structural differences within the brain, such as "AD and NC".

Several studies have investigated the role of RAVLT cognitive measures in the evaluation of AD as well as the relationship between AD related atrophy and RAVLT measures (Estévez-González et al., 2003; Balthazar et al., 2010; Stonnington et al., 2010; Wang et al., 2011). A recent study by Stonnington et al. (2010) investigated the association between AD related structural changes and a RAVLT measure (percent retention) by applying relevance vector regression for the estimation of RAVLT based on MR structural images. However, they did not find a significant correlation between estimated and observed values (R = 0.13, normalized RMSE = 1) in an ADNI dataset of 39 AD, 92 MCI and 32 NC subjects. For comparison purposes, we also calculated normalized RMSE (by normalizing the observed scores to have zero mean and unit variance) for the estimation of RAVLT immediate (RMSE = 0.87, R = 0.50) and RAVLT Percent Forgetting (RMSE = 0.90, R = 0.43). In contrast to Stonnington et al. (2010), our study indicated a significant

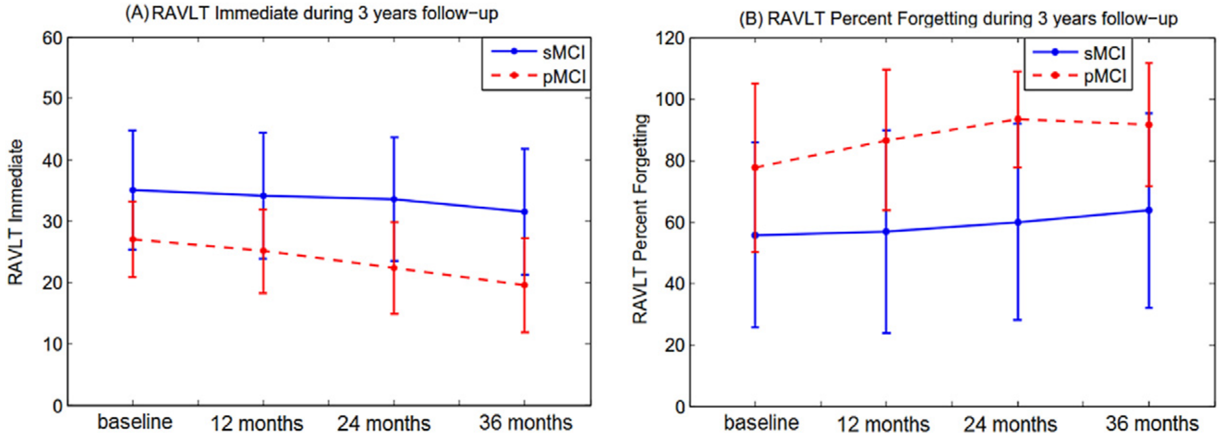

Fig. 4. Mean RAVLT scores (A–B) during 3years follow-up assessment in pMCI and sMCI subjects with error bars representing the standard deviation.

relationship between RAVLT measures and structural atrophy caused by AD. The improved prediction performance of our model stems both from the larger number of subjects used to train the model and from a better approach for learning the model (ENLR in contrast to KRVR used by Stonnington et al., 2010). Relative to the machine learning approach used, Stonnington et al. (2010) speculated that the estimation of RAVLT, which focuses on the specific aspects of cognitive ability, might be challenging based on the whole brain MRI. However, our results demonstrate that this challenge can be in part overcome by using sparsity inducing learning methods, such as ENLR. In addition to RAVLT Immediate and RAVLT Percent Forgetting, we also estimated the delayed recall score from gray matter density using proposed approach in a full dataset (AD, MCI and NC; Results of this experiment are available in the Supplement). As expected, the predictive accuracy evaluated by cross-validation ($R = 0.44, Q^2 = 0.19, MAE = 2.83$) was almost equivalent to that of RAVLT Percent Forgetting, which is a measure of delayed recall taking into account the relationship of immediately and delayed recalled words.

The knowledge of top predictors is crucial to understand which brain regions are most influential in estimation of RAVLT scores as well as how strongly these measures are related to brain atrophy caused by AD. One proposed use of the elastic net penalized linear regression for constructing predictive model was to obtain an interpretable model. As stated in Section 2.4, the ENLR performs variable selection simultaneously with model estimation, thus providing a subset of relevant voxels for the learning procedure. Note that while also KRVR provided relatively high predictive performance for the estimation of both RAVLT scores (although the predictive performance of KRVR was consistently lower than the predictive performance of ENLR in all experiments, see Table 2), the interpretation of the KRVR model is hard due to kernelization. The top ranked predictors for estimating RAVLT Immediate (learning) are listed in Table 3 and for estimating RAVLT Percent Forgetting are listed in Table 4. Our finding of top predictors of medial temporal lobe structures and amygdala for estimation of RAVLT Immediate and angular gyrus, hippocampus and amygdala for estimation of RAVLT Percent Forgetting are consistent with previous knowledge. The essential role of medial temporal lobe structures, especially hippocampus, for episodic memory has been known for long (Squire and Wixted, 2011; Jeong et al., 2015). Specifically, these structures are thought to be involved for the formation and the maintenance of memories after learning before storing to other cortical areas (Squire and Wixted, 2011). In addition, atrophy in bilateral temporal white matter close to the structures involved in memory formation including the hippocampus, entorhinal cortex, and amygdala has been consistently combined with AD pathology (Li et al., 2012).

Recent studies have suggested the involvement of widely distributed cortical network and the importance of its interactive roles

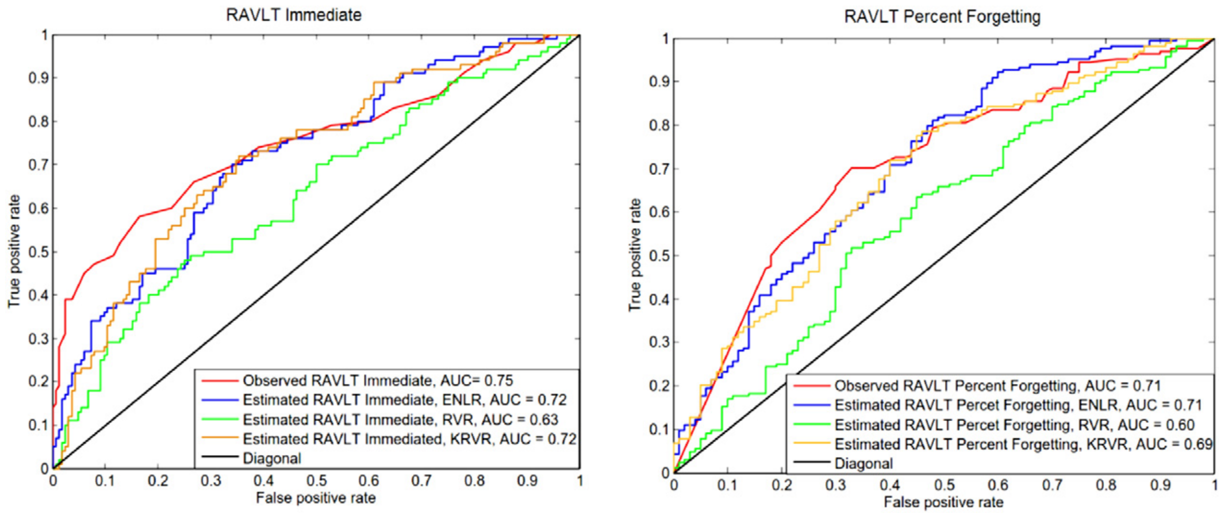

Fig. 5. ROC curves of MCI subjects classification to sMCI or pMCI using observed RAVLT and estimated RAVLT based on different methods (ENLR, RVR, KRVR). The learning was done using all subjects (AD, MCI and NC) and the evaluation was done on pMCI and sMCI subjects (median within 100 runs). Left: RAVLT Immediate, Right: RAVLT Percent Forgetting.

in the memory process (Jeong et al., 2015). In addition to temporal lobe, prefrontal and parietal cortical areas have been associated with episodic memory (Squire and Wixted, 2011; Brem et al., 2013; Jeong et al., 2015).

The involvement of angular gyrus, located in inferior parietal cortex, in retrieval has been confirmed by functional neuroimaging studies (Kwok et al., 2012; Sestieri et al., 2011) and is also reported in a review study by Jeong et al. (2015). The insular cortex has been related with taste memory processes but may have a role in interaction with amygdala in non-taste recognition memory as well (Bermudez-Rattoni, 2014). Insula and angular gyrus are also parts of the default network (including also anteromedial prefrontal cortex, the precuneus, and the medial temporal lobe) which has been discovered to be disrupted in AD (Jeong et al., 2015). Our findings of the brain regions best predicting learning and retrieval in RAVLT are in line with previous research based on neuroimaging data of neurobiological changes associated with disorders causing dementia and normal memory processes. Specifically, our results indicate that in addition to well-known hippocampus and amygdala, also middle temporal gyrus, angular gyrus and insula are also associated with verbal episodic memory tasks.

Furthermore, our results suggest that a wide network of brain regions is involved in memory processes. While making interpretations about importance of brain regions for prediction is certainly possible with sparse linear regularization based models such as ENLR, this does not mean that ranking the importance of different brain regions in the machine learning analysis of whole brain imaging data would be straight-forward. Even within the same machine learning algorithm, different complementary measures of variable importance can be derived. For example, we have provided two separate and complementary indicators of voxel/region importance in Fig. 2 and Tables 3 and 4. Also, it is important to bear in mind that the weights in machine learning models have a different meaning than the parameter estimates in the forward models produced by a standard mass-univariate analysis (Haufe et al., 2014).

The accuracy of estimated RAVLT measures improved little by adding age-correction procedure in the learning process(although the improvement was statistically significant by run-wise applied permutation test). Studies of normal memory processes have indicated that subject demographics, and especially age, have considerable effect on the RAVLT cognitive test in the cognitively normal individuals (Magalhães and Hamdan, 2010; Malloy-Diniz et al., 2007) and at the same time, aging changes the brain structure Good et al. (2001). However, in our experiments removing the normal aging effect resulted only in slight improvement in the estimated RAVLT scores. We hypothesize that this was due to a large effect of AD pathology on both MRI and RAVLT that completely overshadows the effects of normal aging.

In the current work, we explored the utility of estimated and observed RAVLT measures for predicting conversion to AD in MCI subjects. The AD conversion prediction in MCI patients has attracted increasing interest recently, due to an opportunity for an early-stage AD diagnosis (Eskildsen et al., 2013; Wee et al., 2013; Gaser et al., 2013). Previous studies have assessed the predictive value of different neuroimaging techniques in AD conversion prediction. In our previous work (Moradi et al., 2015), we developed a MRI based biomarker by using MRI data and age information which resulted in cross-validated AUC of 77% for discriminating pMCI and sMCI patients, we further obtained an AUC of 90% by integrating MRI biomarker with neuropsychological test results. In another recent study by Eskildsen et al. (2015), an AUC of 76% was reported for predicting AD in MCI patients based on structural MRI and age information using machine learning algorithms. Moreover, the prediction of AD in MCI patients using different biomarkers was recently studied by Dukart et al. (2015). Within different single biomarkers including sMRI, positron emission tomography (FDG-PET) and apolipoprotein (APOE), the highest performance was achieved by FDG-PET (AUC = 82%). They also showed that integrating several biomarkers significantly improved the AD conversion prediction in MCI patients (AUC = 84%). In overall, the reported accuracies based on single neuroimaging modalities in recent studies varies between 70–80% (Moradi et al., 2014; Eskildsen et al., 2015; Salvatore et al., 2016), however, studies based on combination of several data sources such as neuroimaging, genetics information and cognitive test results, have been reported higher performance for predicting AD in MCI patients (accuracy between 80–90%) (Moradi et al., 2015; Dukart et al., 2015; Ritter et al., 2015). Although the current work did not focus on the AD conversion prediction, the achieved performance for predicting conversion to AD in MCI patients based on both RAVLT Immediate (AUC = 0.75) and RAVLT Percent Forgetting (AUC = 0.71) were comparable to the predictive performance of neuroimaging biomarkers (Teipel et al., 2015; Salvatore et al., 2016).Moreover, the analysis of longitudinal 3 years follow-up assessments of RAVLT measures in MCI subjects showed a notable decline in the RAVLT Immediate score and an increase in RAVLT percent Forgetting in pMCI subjects while remaining relatively stable for both scores in sMCI subjects. These findings reconfirm the diagnostic power of RAVLT for early diagnosis of Alzheimer's disease as reported elsewhere Estévez-González et al. (2003). Interestingly, the estimated RAVLT scores were almost as good as the observed ones in predicting conversion to AD indicating that structural brain imaging representations of episodic memory displayed most of the essential information in RAVLT for detecting AD pathology. However, the conversion predictions improved when observed and estimated scores were combined suggesting that the differential information contained in these two types of scores might be useful for early AD diagnosis.

In summary, we designed a predictive model for analyzing the association between RAVLT measures (learning and retrieval) and AD related structural atrophy using MRI scans in a large ADNI dataset. our experimental results indicated a strong relationship between RAVLT Immediate and Percent Forgetting scores and the brain atrophy caused by AD. Moreover, both RAVLT Immediate and RAVLT Percent Forgetting were found to be reliable for AD diagnosis and reflect well the underlying AD pathology. However, we found that RAVLT Immediate is more correlated with AD related brain atrophy as well as it has a higher predictive accuracy for the AD conversion prediction in MCI patients.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

This project has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258), el Ministerio de Educación, Cultura y Deporte (CEI-15-17) and Banco Santander.

Appendix A. Penalized linear regression

Linear regression models the response variable y as a linear combination of the predictor variables $\mathbf{x}$. The predictor variables $\mathbf{x} \in \mathbb{R}^{N \times D}$ are MRI based gray matter densities, where N is the number of subjects and D is the number of voxels, i.e., the dimensionality of MRI data, and the response variable y is the RAVLT score. The linear model is formalized as

$$y_i = \mathbf{w}^T\mathbf{x}_i + w_0 + \epsilon_i = \sum_{j=1}^{D} w_j x_{i,j} + w_0 + \epsilon_i, \tag{A.1}$$

where the index i refers to a subject, $\mathbf{w}$ and w_0 are the model parameters and ϵ_i is the error term. The ordinary least squares (OLS) estimation determines the model parameters by minimizing the residual sum of squares (RSS):

$$RSS(w) = \sum_{i=1}^{N} (y_i - w_0 - w_1 x_{i1} - \ldots - w_D x_{iD})^2, \tag{A.2}$$

However, when the number of predictors is larger than the number of subjects ($D \gg N$), the OLS does not provide a unique solution. Moreover, a high number of predictors may cause the curse of dimensionality, i.e., the lack of generality caused by over-fitting. For avoiding the curse of dimensionality, many variable/feature selection methods have been proposed in neuroimaging data (Tohka et al., 2016; Mwangi et al., 2014). Among them, the regularization methods have gained considerable attention (Miller, 2002). Similarly to OLS-based parameter estimation, penalized linear regression estimates the model parameters by minimizing RSS, but it also shrinks some of the regression parameters towards zero. In this way, it performs simultaneously parameter estimation and variable selection. Here, as the dimensionality of MRI data is high ($D = 29852$), we used penalized least squares approach with the elastic net penalty (Zou and Hastie, 2005). The elastic net penalty is a weighted average of the LASSO penalty $\sum_{j=1}^{D} |\mathbf{w}_j|$ (Tibshirani, 1996) and the ridge penalty $\sum_{j=1}^{D} \mathbf{w}_j^2$. The LASSO penalty acts as a variable selector by forcing many parameters to have zero values leading to a sparse solution. In neuroimaging applications in which many relevant variables are correlated with each other, LASSO tends to select only one of them while ignoring other correlated variables albeit they would be relevant (Carroll et al., 2009). This is obviously not desired. In contrast, ridge regression penalty shrinks the coefficients of the correlated variables towards each other and assigns similar coefficients values to them. However, ridge regression does not result in a sparse solution, with many zero parameters. However, a combination of these two penalties leads to a sparse model combined with the grouping effect, providing a good solution in neuroimaging applications (Zou and Hastie, 2005; Carroll et al., 2009). In ENLR, the model is solved by minimizing the elastic net cost function:

$$\frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - \mathbf{x}_i^T \mathbf{w})^2 + \lambda[(1-\alpha)\, \|\mathbf{w}\|_2^2 /2 + \alpha \|\mathbf{w}\|_1], \tag{A.3}$$

where the regularization parameter λ is found by cross-validation and $\alpha \in [0, 1]$ defines the compromise between ridge and lasso penalties. In our experiments, we selected $\alpha = 0.5$ to give equal weights for the ridge and lasso penalties. A limitation of the elastic net penalty is that it does not consider spatial relationships of the voxels and neighboring voxels are not required to receive similar weights. While there are regularizers that take into account the spatial relationships among the voxels, such as GraphNet Grosenick et al. (2013), these come with more parameters to select, longer computation times and have found to produce more variable estimate of the generalization error in the case of dementia related classification tasks Tohka et al. (2016).

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org10.1016/j.nicl.2016.12.011.

References

Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., Draganski, B., Initiative, A.D.N., 2013. How early can we predict Alzheimer's disease using computational anatomy? Neurobiol. Aging 34, 2815–2826.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. 99, 6562–6566.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Anderson, M.J., Robinson, J., 2001. Permutation tests for linear models. Aust. N. Z. J. Stat. 43, 75–88.

Arlot, S., Celisse, A., et al. 2010. A survey of cross-validation procedures for model selection. Stat. surv. 4, 40–79.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Balthazar, M.L., Yasuda, C.L., Cendes, F., Damasceno, B.P., 2010. Learning, retrieval, and recognition are compromised in aMCI and mild AD: are distinct episodic memory processes mediated by the same anatomical structures? J. Int. Neuropsychol. Soc. 16, 205–209.

Beheshti, I., Demirel, H., Initiative, A.D.N., et al. 2016. Feature-ranking-based Alzheimer's disease classification from structural MRI. Magn. Reson. Imaging 34, 252–263.

Bermudez-Rattoni, F., 2014. The forgotten insular cortex: its role on recognition memory formation. Neurobiol. Learn. Mem. 109, 207–216.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brem, A.-K., Ran, K., Pascual-Leone, A., 2013. Learning and memory. Handb. Clin. Neurol. 116, 693.

Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., et al. 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage 111, 562–579.

Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., Cohen, R., 2011. Penalized least squares regression methods and applications to neuroimaging. NeuroImage 55, 1519–1527.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44, 112–122.

Coupé, P., Fonov, V.S., Bernard, C., Zandifar, A., Eskildsen, S.F., Helmer, C., Manjón, J.V., Amieva, H., Dartigues, J.-F., Allard, M., et al. 2015. Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: toward an early individual prognosis. Hum. Brain Mapp. 36, 4758–4770.

Cuadra, M.B., Cammoun, L., Butz, T., Cuisenaire, O., Thiran, J.-P., 2005. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Trans. Med. Imaging 24, 1548–1565.

Dubois, B., Feldman, H.H., Jacova, C., Cummings, J.L., DeKosky, S.T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N.C., Galasko, D., et al. 2010. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 9, 1118–1127.

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons.

Dukart, J., Sambataro, F., Bertolino, A., 2015. Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers. J. Alzheimers Dis. 49, 1143–1159.

Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Collins, D.L., Initiative, A.D.N., et al. 2015. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. Neurobiol. Aging 36, S23–S31.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., Initiative, A.D.N., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning. Neuroimage 65, 511–521.

Estévez-González, A., Kulisevsky, J., Boltes, A., Otermín, P., García-Sánchez, C., 2003. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging. Int. J. Geriatr. Psychiatry 18, 1021–1028.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., Initiative, A.D.N., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50, 883–892.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1.

Gaser, C., 2009. Partial volume segmentation with adaptive maximum a posteriori (map) approach. NeuroImage 47, S121.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., Initiative, A.D.N., 2013. Brainage in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. PloS one 8, e67346.

Gomar, J.J., Conejero-Goldberg, C., Davies, P., Goldberg, T.E., Initiative, A.D.N., 2014. Extension and refinement of the predictive value of different classes of markers in ADNI: four-year follow-up data. Alzheimers Dement. 10, 704–712.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage 14, 21–36.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with graphnet. NeuroImage 72, 304–321.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Hastie, T., Tibshirani, R.J., Friedman, J.H., 2011. The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.

Huttunen, H., Manninen, T., Tohka, J., 2012. MEG mind reading: strategies for feature selection. Proc. Fed. Comput. Sci. Event 2012, 42–49.

Jack, C.R., Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 7, 257–262.

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9, 119–128.

Jeong, W., Chung, C.K., Kim, J.S., 2015. Episodic memory in aspects of large-scale brain networks. Front. Hum. Neurosci. 9,

Khundrakpam, B.S., Tohka, J., Evans, A.C., Group, B.D.C., et al. 2015. Prediction of brain maturity based on cortical thickness at different spatial resolutions. Neuroimage 111, 350–359.

Kwok, S.C., Shallice, T., Macaluso, E., 2012. Functional anatomy of temporal organisation and domain-specificity of episodic memory retrieval. Neuropsychologia 50, 2943–2955.

Li, J., Pan, P., Huang, R., Shang, H., 2012. A meta-analysis of voxel-based morphometry studies of white matter volume alterations in Alzheimer's disease. Neurosci. Biobehav. Rev. 36, 757–763.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.

Magalhães, S.S., Hamdan, A.C., 2010. The Rey auditory verbal learning test: normative data for the Brazilian population and analysis of the influence of demographic variables. Psychol. Neurosci. 3, 85.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83.

Malloy-Diniz, L.F., Lasmar, V.A.P., Gazinelli, L.d.S.R., Fuentes, D., Salgado, J.V., 2007. The Rey auditory-verbal learning test: applicability for the Brazilian elderly population. Rev. Bras. Psiquiatr. 29, 324–329.

Masdeu, J.C., Kreisl, W.C., Berman, K.F., 2012. The neurobiology of Alzheimer disease defined by neuroimaging. Curr. Opin. Neurol. 25 (4), 410–420.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., et al. 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 7, 263–269.

Miller, A., 2002. Subset Selection in Regression. CRC Press.

Moradi, E., Khundrakpam, B., Lewis, J.D., Evans, A.C., Tohka, J., 2016. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. NeuroImage In press.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.

Moradi, E., Tohka, J., Gaser, C., 2014. Semi-supervised learning in MCI-to-ad conversion prediction—when is unlabeled data useful? Pattern Recognition in Neuroimaging, 2014 International Workshop on. pp. 1–4.

Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244.

Petersen, R., Aisen, P., Beckett, L., Donohue, M., Gamst, A., Harvey, D., Jack, C., Jagust, W., Shaw, L., Toga, A., et al. 2010. Alzheimer's disease neuroimaging initiative (ADNI) clinical characterization. Neurology 74, 201–209.

Rajapakse, J.C., Giedd, J.N., Rapoport, J.L., 1997. Statistical approach to segmentation of single-channel cerebral MR images. IEEE Trans. Med. Imaging 16, 176–186.

Rey, A., 1964. L'examen clinique en psychologie [the clinical psychological examination]. Paris: Presses Universitaires de France

Ricci, M., Graef, S., Blundo, C., Miller, L.A., 2012. Using the Rey auditory verbal learning test (RAVLT) to differentiate Alzheimer's dementia and behavioural variant fronto-temporal dementia. Clin. Neuropsychol. 26, 926–941.

Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., Haynes, J.-D., Initiative, A.D.N., et al. 2015. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimers Dement.: Diagn., Assessment Dis. Monit. 1, 206–215.

Salvatore, C., Battista, P., Castiglioni, I., 2016. Frontiers for the early diagnosis of AD by means of MRI brain imaging and support vector machines. Curr. Alzheimer Res. 13, 509–533.

Salvatore, C., Cerasa, A., Battista, P., Gilardi, M.C., Quattrone, A., Castiglioni, I., Initiative, A.D.N., 2015. Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. Front. Neurosci. 9,

Schoenberg, M.R., Dawson, K.A., Duff, K., Patton, D., Scott, J.G., Adams, R.L., 2006. Test performance and classification statistics for the Rey auditory verbal learning test in selected clinical samples. Arch. Clin. Neuropsychol. 21, 693–703.

Sestieri, C., Corbetta, M., Romani, G.L., Shulman, G.L., 2011. Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses. J. Neurosci. 31, 4407–4420.

Squire, L.R., Wixted, J.T., 2011. The cognitive neuroscience of human memory since HM. Annu. Rev. Neurosci. 34, 259.

Stonnington, C.M., Chu, C., Klöppel, S., Jack, C.R., Ashburner, J., Frackowiak, R.S., Initiative, A.D.N., 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51, 1405–1413.

Teipel, S., Drzezga, A., Grothe, M.J., Barthel, H., Chételat, G., Schuff, N., Skudlarski, P., Cavedo, E., Frisoni, G.B., Hoffmann, W., et al. 2015. Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. Lancet Neurol. 14, 1037–1053.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.

Tierney, M., Szalai, J., Snow, W., Fisher, R., Nores, A., Nadon, G., Dunn, E., George-Hyslop, P.S., 1996. Prediction of probable Alzheimer's disease in memory-impaired patients a prospective longitudinal study. Neurology 46, 661–665.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Tipping, M.E., Faul, A., et al. 2003. Fast marginal likelihood maximisation for sparse Bayesian models. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics.

Tohka, J., Moradi, E., Huttunen, H., 2016. Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics 14, 279–296.

Tohka, J., Zijdenbos, A., Evans, A., 2004. Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage 23, 84–97.

Vergara, I.A., Norambuena, T., Ferrada, E., Slater, A.W., Melo, F., 2008. Star: a simple tool for the statistical comparison of ROC curves. BMC Bioinf. 9, 1.

Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L., 2011. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 557–562.

Wee, C.-Y., Yap, P.-T., Shen, D., 2013. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum. Brain Mapp. 34, 3411–3425.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat Methodol. 67, 301–320.