Hippocampus as a Predictor of Cognitive Performance: Comparative Evaluation of Analytical Methods and Morphometric Measures

Taiyong Li$^{1,2,*}$, Jing Wan$^{1,3,*}$, Zhilin Zhang$^4$, Jingwen Yan$^{1,5}$, Sungeun Kim$^1$, Shannon L. Risacher$^1$, Shiaofen Fang$^2$, M. Faisal Beg$^6$, Lei Wang$^7$, Andrew J. Saykin$^1$, and Li Shen$^{1,3,5,*}$, for the ADNI$^{**}$

$^1$ Radiology and Imaging Sciences, Indiana University School of Medicine, IN, USA
$^2$ School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, China
$^3$ Computer and Information Science, Purdue University Indianapolis, IN, USA
$^4$ Electrical and Computer Engineering, Univ. of California, San Diego, CA, USA
$^5$ School of Informatics, Indiana University Indianapolis, IN, USA
$^6$ School of Engineering Science, Simon Fraser University, BC, Canada
$^7$ Psychiatry and Behavioral Sciences, Northwestern University, IL, USA

Abstract. Predicting cognitive status using hippocampal imaging measures is an important research topic in the study of Alzheimer’s disease (AD). Multivariate regression with sparsity constraint has been shown as an effective approach for identifying AD biomarkers related to cognitive performance. In particular, the multiple measurement vector (MMV) model with common sparsity assumption holds great promise. In this work, we compare four typical MMV algorithms with two traditional regression models in the task of predicting cognitive outcomes from four types of hippocampal measurements: volumes, subfield volumes, surface deformations and voxel-based gray matter measures. The temporal sparse Bayesian learning algorithm (T-MSBL) demonstrates the best performance, suggesting that the MMV-model based sparse Bayesian learning exploiting the correlation structure is a valuable framework in discovering biomarkers related to cognitive performance. Among four different hippocampal measurements, the subfield measures yield the most powerful and stable prediction rates across all the algorithms. Replication of these results in independent cohorts warrants further investigation.

* Equal contribution by Taiyong Li (litaiy@iupui.edu) and Jing Wan (wanjing@iupui.edu). Correspondence to Li Shen (shenli@iupui.edu).

** Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.
1 Introduction

Alzheimer’s disease (AD) is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions. Hippocampus is known to play important roles in consolidating information from short-term memory to long-term memory and is one of the first regions of the brain to suffer damage in the progression of AD. Hippocampal measures extracted from magnetic resonance imaging (MRI) scans have been widely studied to detect the status of AD or mild cognitive impairment (MCI, thought to be a prodromal stage of AD) [9, 15, 17, 19, 26, 28], predict MCI conversion to AD [3], or infer cognitive status [2, 12]. These hippocampal measures include (1) volumes [9, 17, 19], (2) subfield volumes [15, 26], (3) surface deformations [3, 12, 28], and (4) gray matter (GM) measures using voxel-based morphometry (VBM) [19].

Predicting cognitive performance from MRI-based measures is an important research topic in the study of AD. Substantial attention has recently been given to identifying neuroimaging predictors for cognitive decline in AD. Regression models have been investigated to predict clinical scores from individual MRI and/or positron emission tomography (PET) scans [3, 18, 23, 27]. For example, morphometric features of the entire brain were jointly analyzed to predict each selected clinical score using relevance vector regression in [23]. Stepwise regression was performed in a univariate, pairwise fashion to relate each imaging measure to each cognitive score in [18]. A sparse-based multi-task learning approach was employed to select features that could predict cognitive scores in [27].

Given a variety of hippocampal measures and multiple prediction algorithms, in this paper, we conduct a comparative study using the MRI and cognitive data available in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. We apply six state-of-the-art algorithms to the above four types of hippocampal data to evaluate the performance of the algorithms and the predictive power of each data type. Our empirical results demonstrate that the temporal sparse Bayesian learning algorithm (T-MSBL) [31] can achieve the best performance among the algorithms and hippocampal subfield volumes have the strongest predictive power on cognitive scores among all the measures.

2 Materials and Methods

2.1 Materials

Data used in this study were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). ADNI is a landmark investigation sponsored by the NIH and industrial partners designed to collect longitudinal neuroimaging, biological and clinical information from over 800 participants that will track the neural correlates of memory loss from an early stage. One goal of ADNI is to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Updated information can be found at www.adni-info.org. Following a previous imaging genetic
Table 1. Participant characteristics

<table>
<thead>
<tr>
<th>Category</th>
<th>HC</th>
<th>MCI</th>
<th>AD</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (M/F)</td>
<td>88/73</td>
<td>175/97</td>
<td>58/56</td>
<td>0.023</td>
</tr>
<tr>
<td>Baseline Age (years; Mean±STD)</td>
<td>76.22±4.95</td>
<td>75.12±7.23</td>
<td>75.57±7.60</td>
<td>0.262</td>
</tr>
<tr>
<td>Education (years; Mean±STD)</td>
<td>16.14±2.61</td>
<td>15.68±2.96</td>
<td>15.03±3.07</td>
<td>0.007</td>
</tr>
<tr>
<td>Handedness (R/L)</td>
<td>150/11</td>
<td>247/25</td>
<td>109/5</td>
<td>0.244</td>
</tr>
</tbody>
</table>

Fig. 1. Hippocampus measures extracted from MRI scans: (a) Volumes or voxel-based GM measures, (b) subfields volumes, and (c) shape measures.

Image Measurements: This work examined four types of hippocampal measurements, including (1) bilateral volumes, (2) subfield volumes, (3) surface deformations from a template hippocampal pair, and (4) GM measures in the hippocampal region generated by VBM. To obtain left and right hippocampal volumes and subfield volumes, the hippocampus from the baseline 1.5T MRI scans (shown in Fig. 1(a)) was segmented by applying the FreeSurfer pipeline (Martinos Center for Biomedical Imaging, Boston, MA). Hippocampal subfields (shown in Fig. 1(b)) were then segmented automatically using a Bayesian modeling approach [22]. There are seven subfields in each of the left and right hippocampi (total 14 subfields): CA1, CA2-3, CA4-DG, fimbria, hippocampal fissure, presubiculum, and subiculum. Total volumes and subfield volumes for left and right hippocampi were then extracted and adjusted for the baseline age, gender, education, handedness and total intracranial volume (ICV) using the regression weights derived from the HC participants.

For generating hippocampal surface measures (shown in Fig. 1(c)), FreeSurfer and Large Deformation Diffeomorphic Metric Mapping (FS+LDDMM) was applied [11]. This fully-automated segmentation pipeline first used FreeSurfer subcortical labeling to provide information for initialization, and then employed LDDMM to generate a diffeomorphic transformation so that anatomical structures can be mapped consistently and smoothly [1]. To remove size effect, each hippocampus was scaled so that its ICV was adjusted to a constant (i.e., mean study [24], 547 non-Hispanic Caucasian participants (161 Healthy Control (HC), 272 MCI, 114 AD participants) were included in this study. Table 1 summarizes their characteristics.
Table 2. Description of cognitive measures.

<table>
<thead>
<tr>
<th>Score Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSE</td>
<td>Mini-Mental State Exam score</td>
</tr>
<tr>
<td>ADAS</td>
<td>Alzheimer’s Disease Assessment Scale</td>
</tr>
<tr>
<td>FLUENCY</td>
<td>FLUENCY score</td>
</tr>
<tr>
<td>RAVLT TOTAL</td>
<td>Total score of the first 5 trials</td>
</tr>
<tr>
<td>T30</td>
<td>30 minute delay total number of words recalled</td>
</tr>
<tr>
<td>RECOG</td>
<td>30 minute delay recognition score</td>
</tr>
<tr>
<td>TRAILS A</td>
<td>Trail making A score</td>
</tr>
<tr>
<td>TRAILS B</td>
<td>Trail making B score</td>
</tr>
<tr>
<td>TR(B-A)</td>
<td>TRAILS B-TRAILS A</td>
</tr>
</tbody>
</table>

ICV of all HCs). Rigid body transformation was then applied to register each hippocampus to a template (defined as the mean of all HCs) in a least square fashion. Surface signals were extracted as the deformation along the surface normal direction of the template, and adjusted for baseline age, gender, education, and handedness using the regression weights derived from the HC participants.

Voxel-based morphometry (VBM) was performed using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/) to create a smoothed modulated normalized gray matter (GM) map (1 × 1 × 1 mm voxel size, 10 mm FWHM Gaussian kernel for smoothing) for the baseline MRI scan of each participant using a similar procedure described in [16]. The Automated Anatomical Labeling (AAL) atlas was registered with these maps to extract left and right hippocampi [21]. These hippocampal VBM measures were adjusted for baseline age, gender, education, and handedness using the regression weights derived from the HC participants.

**Cognitive Scores:** We examined five sets of baseline cognitive scores: Mini-Mental State Exam (MMSE), Alzheimer’s Disease Assessment Scale (ADAS), FLUENCY, Rey Auditory Verbal Learning Test (RAVLT) and Trail Making (TRAILS) (see Table 2). Details about these cognitive assessments are available in the ADNI procedure manuals (www.adni-info.org).

### 2.2 Mathematical Model and Algorithms

We use a multivariate regression model to select biomarkers from hippocampal measures to predict cognitive scores. Since we believe that the number of effective biomarkers is small (i.e., sparse), we adopt the multiple measurement vector (MMV) model [4], which has been shown to be an effective model in selecting sparse biomarkers [25, 27].

The mathematical expression of the MMV model [4] is given by

\[ \mathbf{Y} = \Phi \mathbf{X} + \mathbf{V}, \]  

where each column of \( \Phi \in \mathbb{R}^{N \times M} \) is a hippocampal measure, \( \mathbf{Y} \in \mathbb{R}^{N \times L} \) is \( L \) cognitive scores of \( N \) subjects after have performed a cognition task (each subject has \( L \) scores from \( L \) scoring systems), \( \mathbf{V} \) is an unknown noise matrix (or
called model error matrix), and $X \in \mathbb{R}^{M \times L}$ is an unknown coefficient matrix. We hypothesize that only a small number of our input imaging measures contribute to a certain cognitive task and so expect $X$ to have only a few nonzero rows. We also assume that these biomarkers more or less affect all the cognitive scores under the task, but still allow a nonzero row to contain some zero entries.

There are many algorithms for this problem. Most of them calculate the solution by solving the following unconstrained problem (or its equivalent constrained problem)

$$X = \arg \min_X \|Y - \Phi X\|_F^2 + \lambda g_1(X)$$

with the mixed $\ell_{q,1}$ penalty (typically, $q = 2$ or $q = \infty$) \(^8\)

$$g_1(X) \triangleq \sum_{i=1}^M \|X_i\|_q,$$

where an $\ell_q$ norm is applied on each row of $X$, and an $\ell_1$ norm is applied on the $M$ calculated $\ell_q$ norms. $\lambda$ is a regularizer, which is generally tuned by cross-validation. Algorithms using this penalty include group Lasso (the variant used for the MMV model) [29], the Mixed $\ell_2/\ell_1$ Program [5], and many domain-specific algorithms for feature extraction [14].

Note that the penalty $g_1(X)$ is a convex penalty. In some scenarios non-convex penalty based algorithms can yield better performance. A typical non-convex penalty is:

$$g_2(X) \triangleq \sum_{i=1}^M (\|X_i\|_2)^p, \quad 0 < p < 1$$

The MMV based FOCal Underdetermined System Solver (M-FOCUSS) method [4] is a representative in this group.

In addition to the penalties $g_1(X)$ and $g_2(X)$, a more powerful penalty is proposed [30, 25]:

$$g_C(X) = \sum_{i=1}^M w_i^{(k)} \sqrt{X_i \cdot (B^{(k)})^{-1} X_i^T}.$$  

Here $k$ is the iteration count. $w_i^{(k)}$ is a weighting at the $k$-th iteration, which is a function of the estimate of $X$ in the previous iteration. $B^{(k)}$ is an estimated positive definite matrix in the $k$-th iteration capturing the correlation structure within each row of $X$. Compared to the penalties $g_1(X)$ and $g_2(X)$, we can see the penalty $g_C(X)$ exploits the correlation within each row $X_i$; by replacing the $\ell_q$ norm $\|X_i\|_q$ with the measure $\sqrt{X_i \cdot (B^{(k)})^{-1} X_i^T}$. This penalty has a number of

\(^8\) Throughout the paper $X_i$ and $X_j$ denotes the $i$-th row and the $j$-th column of $X$, respectively.
advantages over the penalties \( g_1(\mathbf{X}) \) and \( g_2(\mathbf{X}) \), which has been discussed in [25].

The Temporal MMV Sparse Bayesian Learning (T-MSBL) [31] and the temporal M-FOCUSS algorithm [30] are representative algorithms. In our experiments, we choose a fast variant of T-MSBL [25], which is derived using MacKay’s fixed point method [13].

In our experiments, we also choose three algorithms which do not use the above penalties. One is the Simultaneous Orthogonal Matching Pursuit (S-OMP) [20], which uses a greedy method to seek the solution to (1). Another is the Multi-Task Compressive Sensing (MT-CS) [10]. MT-CS is a Sparse Bayesian Learning (SBL) algorithm, which treats the MMV model (1) as \( L \) dependent single measurement vector (SMV) models, i.e., \( \mathbf{Y}_i = \mathbf{\Phi X}_i + \mathbf{V}_i \) \((i = 1, \cdots, L)\), where every \( \mathbf{X}_i \) shares a common prior. Note that this model is an alternative one to the MMV model in multi-task learning. The last one is the traditional Ridge Regression [8]. It is originally proposed for an SMV model. To use it in our problem, we apply it to each \( \mathbf{Y}_i = \mathbf{\Phi X}_i + \mathbf{V}_i \) \((i = 1, \cdots, L)\) independently.

3 Experimental Results

For comparative evaluation of analytical methods and hippocampal measures for predicting cognitive outcomes, we have performed extensive experiments using six regression methods on four types of hippocampal predictors and five types of cognitive outcomes. The six regression methods included the T-MSBL [25], the mixed \( \ell_2/\ell_1 \) Program [5], M-FOCUSS [4], Simultaneous Orthogonal Matching Pursuit (S-OMP) [20], Multi-Task Compressive Sensing (MT-CS) [10], and Ridge Regression [7]. The four types of hippocampal predictors included baseline hippocampal volume, subfield, surface, and VBM measures. The five types of cognitive scores included baseline MMSE, ADAS, FLUENCY, RAVLT, and TRAILS (see Table 2). Our experiments have also been conducted on the total sample (ALL) as well as three sub-samples including HC & AD, HC & MCI,
Fig. 3. Example five-fold cross-validation performance comparison among different hippocampal measures, where the data set of HC & AD was used in the analysis. The correlation coefficients between the actual values and the predicted values are plotted against cognitive outcomes. Each panel shows the result using one of the six algorithms.

and MCI & AD. Five-fold cross validation was applied in the experiments to estimate prediction performance in an unbiased fashion. The prediction performance was measured by the correlation coefficients between actual cognition scores and predicted scores of all the test samples.

Fig. 2 shows an example performance comparison among six methods, where the data set of HC & AD was used in the analysis. The significance map of each comparison is shown in Fig. 4(a), where the p value was calculated from the paired sample t test between two sets of cross-validation correlation coefficients. Note that the results of $\ell_2/\ell_1$ and M-FOCUSS were extremely similar such that the line of $\ell_2/\ell_1$ was covered by that of M-FOCUSS. Among all the methods, T-MSBL achieved the highest correlation coefficients in most cases.

We also compared the predictive power of each type of hippocampal measures. Fig. 3 shows the mean cross-validation results and Fig. 4(b) shows the significance map of each comparison. The subfield measures achieved the highest performance than any other data. Although detailed surface and VBM measures can potentially provide more discriminative power, this hypothesis is not validated by our results. Simple hippocampal volume or more advanced hippocampal subfield volumes have already contained adequate information to outperform more detailed surface or voxel-based measures.

For demonstrating how to identify potential neuroimaging biomarkers associated with cognitive outcome, we provide a visualization in Fig. 5 to color map the regression weights of each hippocampal subfield on the ADAS score for two methods: the T-MSBL and the mixed $\ell_2/\ell_1$ Program. We can see that the
biodmarker pattern identified by the T-MSBL was much sparser than that of the mixed $\ell_2/\ell_1$ Program. Specifically, all of the weights of the biomarkers identified by $\ell_2/\ell_1$ are non-zero while only 20 out of 56 ones identified by T-MSBL are non-zero, which demonstrates the latter is sparse. The main four subfields (Left CA1, Right CA1, Left Subiculum and Right Subiculum) identified by T-MSBL have been shown to be associated with AD [6].

Finally, we focused on the T-MSBL and the mixed $\ell_2/\ell_1$ Program, and compared the results on different data sets in Fig. 6. While all the results showed a similar trend, we achieved the best prediction accuracy using the data set of HC & AD. Given the big cognitive difference between HC and AD participants, it is expected that using such a data set tends to yield a good prediction performance. Compared with the mixed $\ell_2/\ell_1$ Program, the T-MSBL performed better in all the cases, in particular when the detailed measures were used. This suggests that the T-MSBL is a promising approach for feature selection from a high dimensional data set.

4 Discussion

In this work, we compared four typical MMV algorithms with two traditional SMV algorithms on predicting a variety of cognitive scores using four types of hippocampal morphometric measures. We observed that, in the most cases, T-MSBL had the best performances among all the methods. This is due to the following advantages of T-MSBL.

1. In our problem a given imaging biomarker can affect multiple cognitive scores, so the coefficients in the same row of $X$ are largely correlated. Recently, it has been found [31] that when such correlation is present, most
existing methods have seriously degraded performance due to the ignorance of the correlation. In contrast, T-MSBL can adaptively estimate and exploit the correlation structure in $\mathbf{X}_i \cdot (\forall i)$ to improve performance.

2. It has been shown [25] that T-MSBL is equivalent to an iterative reweighted $\ell_2/\ell_1$ algorithm, and thus it can obtain more sparse features than existing $\ell_2/\ell_1$ algorithms. This is a very desired property for our complicated datasets, where sparse biomarkers generally lead to stable performance and good generalization.

3. It has been shown [31] that in noiseless situation the global minimum of T-MSBL’s cost function corresponds to the sparsest solution, as long as any $N$ columns of $\Phi$ are linearly independent. However, under the same situation, we cannot guarantee that the global minimum of the algorithms with the $\ell_2/\ell_1$ penalty corresponds to the sparsest solution, unless more strict conditions on $\Phi$ and $\mathbf{X}$ are satisfied.

4. In our problem the columns of $\Phi$ could be highly correlated. In some datasets used in our experiments the maximum correlation reaches 0.99. The correlation in $\Phi$ can result in poor performance for most algorithms. In contrast, experiments have shown that T-MSBL still maintains its superior performance in this situation.

For the evaluation of the predict power of different hippocampal measures, we observed that hippocampal subfield volumes have performed the best among all four sets of measures. The detailed measures of hippocampal surface or volume (i.e. hippocampal surface signals or hippocampal VBM data) could not provide additional information for improving the predicting cognitive scores. T-MSBL seemed to be the most stable method, and achieved similar performance while using different sets of hippocampal measures. Using independent cohorts to

---

Fig. 5. Heat maps of regression weights in hippocampal subfield analyses using L2/L1 or T-MSBL. The same experiment was tested on each of the following four sample sets: HC & AD, HC & MCI, MCI & AD, and ALL.
Fig. 6. Example five-fold cross-validation performance comparison among different data sets using the T-MSBL and the mixed $\ell_2/\ell_1$ Program. The correlation coefficients between the actual values and the predicted values are plotted against cognitive outcomes. Each panel shows the result using one type of machine learning methods.

replicate these biomedical findings and to confirm the promise of the T-MSBL method warrants further investigation. Another interesting future topic could be to examine whether dimensionality reduction or feature extraction methods such as principal component analysis can help improve the prediction power of detailed hippocampal surface or gray matter measures.

Acknowledgement

This research was supported by NSF IIS-1117335, NIH UL1 RR025761, U01 AG024904, NIA RC2 AG036535, R01 AG19771, and P30 AG10133-18S1.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimers Association; Alzheimers Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Inno- genetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda
Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References