Prioritization of Cognitive Assessments in Alzheimer's Disease via Learning to Rank using Brain Morphometric Data

Bo Peng
Computer & Information Science
IUPUI, Indianapolis, IN
peng10@iu.edu

Xiaohui Yao
Biostatistics, Epidemiology & Informatics
University of Pennsylvania, Philadelphia, PA
Xiaohui.Yao@pennmedicine.upenn.edu

Shannon L. Risacher
Radiology & Imaging Sciences
Indiana University, Indianapolis, IN
srisache@iupui.edu

Andrew J. Saykin
Radiology & Imaging Sciences
Indiana University, Indianapolis, IN
asaykin@iupui.edu

Li Shen
Biostatistics, Epidemiology & Informatics
University of Pennsylvania, Philadelphia, PA
Li.Shen@pennmedicine.upenn.edu

Xia Ning
Biomedical Informatics
Ohio State University, Columbus, OH
ning.104@osu.edu

***Abstract*—We propose an innovative machine learning paradigm enabling precision medicine for prioritizing cognitive assessments according to their relevance to Alzheimer's disease at the individual patient level. The paradigm tailors the cognitive biomarker discovery and cognitive assessment selection process to the brain morphometric characteristics of each individual patient. We implement this paradigm using a newly developed learning-to-rank method `PLTR`. Our empirical study on the ADNI data yields promising results to identify and prioritize individual-specific cognitive biomarkers as well as cognitive assessment tasks based on the individual's structural MRI data. The resulting top ranked cognitive biomarkers and assessment tasks have the potential to aid personalized diagnosis and disease subtyping.**

1. Introduction

Identifying structural brain changes related to cognitive impairments is an important research topic in the study of Alzheimer's Disease (AD). Regression models have been widely investigated to predict cognitive outcomes using morphometric measures extracted from structural magnetic resonance imaging (MRI) scans (e.g., [1]). Such studies can improve the understanding of the neuroanatomical basis of cognitive impairments, but are not designed to directly impact clinical practice. To bridge this gap, here we propose a new learning paradigm which ranks cognitive assessments according to their relevance to AD using brain MRI data.

Cognitive assessments provide the most common clinical routine for the diagnosis of AD. Given a large number of cognitive assessment tools and a time-limited office visit, determining a proper set of cognitive tests is a widely studied topic. Most existing studies aim to create selection guidelines for a targeted population [2]. In this work, we propose a novel learning paradigm that embraces the concept of precision medicine and tailors the cognitive test selection process to the individual characteristics of a given patient. Specifically, we perform an innovative application of a newly developed learning-to-rank method, denoted as `PLTR` [3], to the structural MRI and cognitive assessment data of the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort [4]. Using structural MRI measures as the individual characteristics, we aim to not only identify individual-specific cognitive biomarkers but also prioritize them and their corresponding assessment tasks according to AD-specific abnormality.

The uniqueness of our study is twofold. First, traditional regression-based studies for prediction of cognitive performances from MRI data focuses on identifying relevant imaging biomarkers at the population level. The proposed new model aims to identify AD-relevant cognitive biomarkers tailored to each individual patient. Second, the identified cognitive biomarkers and assessments are prioritized based on the individual's brain characteristics, which can be used to guide the determination of cognitive assessments in a personalized fashion in clinical practice. It has the potential to enable personalized diagnosis and disease subtyping.

2. Materials

The study sample from the ADNI cohort [4] consists of 819 ADNI-1 subjects, including 229 healthy control (HC), 397 mild cognitive impairment (MCI, a prodromal stage of AD) and 193 AD participants. Combining MCI and AD subjects as patients, we have 590 cases and 229 controls.

Baseline 1.5T MRI scans and cognitive assessment data were downloaded from the ADNI website (adni.loni.usc.edu). MRI scans were processed using Freesurfer version 5.1

Acknowledgements: This work was supported in part by NIH R01 EB022574, R01 LM011360, U19 AG024904, R01 AG019771, and P30 AG010133; NSF IIS 1837964 and 1855501. The complete ADNI Acknowledgement is available at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

as in [5], where volumetric and cortical thickness measures of 101 regions relevant to AD were extracted to characterize brain morphometry.

In this study, our analysis focuses on 151 measures assessed in 15 neuropsychological tests. For convenience, below we call these measures as *cognitive features* and these tests as *cognitive tasks*. The 15 studied tasks include Alzheimer's Disease Assessment Scale (ADAS), Clinical Dementia Rating Scale (CDR), Functional Assessment Questionnaire (FAQ), Geriatric Depression Scale (GDS), Mini-Mental State Exam (MMSE), Modified Hachinski Scale (MODHACH), Neuropsychiatric Inventory Questionnaire (NPIQ), Boston Naming Test (BNT), Clock Drawing Test (CDT), Digit Span Test (DSPAN), Digit Symbol Test (DSYM), Category Fluency Test (FLUENCY), Weschler's Logical Memory Scale (LOGMEM), Rey Auditory Verbal Learning Test (RAVLT) and Trail Making Test (TRAIL).

3. Methods

We use the joint push and learning-to-rank method as developed in He *et al.* [3], denoted as PLTR, for personalized cognitive feature prioritization. Our goal is to prioritize cognitive features for each patient that are most relevant to his/her disease diagnosis using patients' brain morphometric measures extracted from their MRI scans. In specific, the cognitive features are in the form of scores or answers from cognitive tasks that the patients take. The prioritization result can potentially be used in clinical practice to recommend the most relevant cognitive features or tasks that can most effectively help diagnosis of an individual.

In the context of MCI/AD cognitive feature prioritization, PLTR learns and uses latent vectors of patients and their imaging features to score each cognitive feature for each patient, and ranks the cognitive features based on their scores; patients with similar imaging feature profiles will have similar latent vectors. During the learning process, PLTR explicitly pushes the most relevant cognitive features on top of the less relevant ones for each patient, and therefrom optimizes the latent patient and cognitive feature vectors so they will reproduce the pushed ranking structures. In PLTR, such latent vectors are learned by solving the following optimization problem:

$$\min_{U,V} \mathcal{L}_s = (1-\alpha)P_s^{\uparrow} + \alpha O_s^{+} + \frac{\beta}{2}R_{uv} + \frac{\gamma}{2}R_{\text{csim}}, \quad (1)$$

where $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m]$ and $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ are the latent vector matrices for patients and features, respectively; $\mathcal{L}_s$ is the overall loss function; and $P_s^{\uparrow}$ measures the average number of relevant cognitive features that are ranked below an irrelevant cognitive feature, defined as follows,

$$P_s^{\uparrow} = \sum_{p=1}^{m} \frac{1}{n_p^+ n_p^-} \sum_{f_i^- \in \mathcal{P}_p} \sum_{f_j^+ \in \mathcal{P}_p^+} \mathbb{I}(s_p(f_j^+) \le s_p(f_i^-)), \quad (2)$$

where m is the number of patients, f_j^+ and f_i^- are the relevant and irrelevant features of patient $\mathcal{P}_p$, n_p^+ and n_p^- are their respective numbers, and $\mathbb{I}(x)$ is the indicator function ($\mathbb{I}(x) = 1$ if x is true, otherwise 0). In Problem (2), $s_p(f_i)$ is a scoring function defined as follows,

$$s_p(f_i) = \mathbf{u}_p^{\mathsf{T}} \mathbf{v}_i, \quad (3)$$

that is, it calculates the score of feature f_i on patient $\mathcal{P}_p$ using their respective latent vectors $\mathbf{u}_p$ and $\mathbf{v}_i$. In Problem (1), O_s^+ measures the ratio of mis-ordered feature pairs over the relevant features among all the patients, defined as follows,

$$O_s^{+} = \sum_{p=1}^{m} \frac{1}{|\{f_i^+ \succ_{\mathcal{P}_p} f_j^+\}|} \sum_{f_i^+ \succ_{\mathcal{P}_p} f_j^+} \mathbb{I}(s_p(f_i^+) < s_p(f_j^+)), \quad (4)$$

where $f_i \succ_R f_j$ represents that f_i is ranked higher than f_j under the relation R. In Problem (1), R_{uv} is a regularizer on U and V to prevent overfitting, defined as $R_{uv} = \frac{1}{m}\|U\|_F^2 + \frac{1}{n}\|V\|_F^2$, where $\|X\|_F$ is the Frobenius norm of matrix X. R_{csim} is a regularizer on patients to constrain patient latent vectors, defined as $R_{\text{csim}} = \frac{1}{m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} w_{pq}\|\mathbf{u}_p - \mathbf{u}_q\|_2^2$, where w_{pq} is the similarity between $\mathcal{P}_p$ and $\mathcal{P}_q$ that is calculated using the imaging features of the patients.

4. Data Processing

4.1. Data Normalization

We selected all the MCI/AD patients from the dataset and did the following data normalization for the patients. We first conducted a t-test on each of the cognitive features between patients and controls, and selected cognitive features if there is a significant difference between patients and controls on these features. Then we converted each of the selected features into $[0, 1]$ by shifting and scaling the feature values. We also converted all the normalized feature values based on the Cohen's d of the features between patients and controls, so that smaller values always indicate more AD possibility. After that, we filtered out features whose values are 0, 1 or 0.5 for more than 95% patients, in order to remove features that are either extremely dominated by patients or controls, or not discriminative. We conducted the same process as above on the imaging features.

4.2. Patient Similarities from Imaging Features

After the above normalization and filtering steps, we have 86 normalized imaging features remaining in the study. We represent each patient as a vector of these features, denoted as $\mathbf{r}_p = [r_{p1}, r_{p2}, \cdots, r_{p86}]$, where r_{pi} $(i = 1, \cdots, 86)$ is an imaging feature for patient p. We calculate the patient similarity from imaging features using the radial basis function (RBF) kernel, that is, $w_{pq} = \exp(-\frac{\|\mathbf{r}_p - \mathbf{r}_q\|^2}{2\sigma^2})$, where w_{pq} is the patient similarity used in R_{csim}.

5. Experimental Protocol

5.1. Training-Testing Data Splits

We test our methods in two settings: cross validation and leave-out validation. In the cross validation (CV), we split the cognitive tasks for each patient into 5 folds. That is, all the features in a cognitive task will be either split

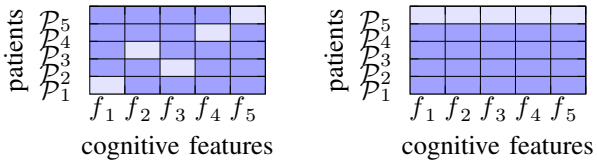

Figure 1: Data split for cross validation (left) and leave-out validation (right)

into training or testing set. We use 4 folds for training and the rest fold for testing, and do such experiments 5 times, each with one of the 5 folds as the testing set. The overall performance of the methods is averaged out over the 5 testing sets. This setting corresponds to the goal to prioritize additional cognitive tasks that a patient should complete. In the leave-out validation (LOV), we split patients into training and testing sets, such that a certain patient and all his/her cognitive features will be either in the training set or in the testing set. This corresponds to the use scenario to identify the most relevant cognitive tasks that a new patient needs to take, based on existing imaging information of the patient, when the patient has not completed any cognitive tasks. Figure 1 demonstrates the CV and LOV data split process.

5.2. Evaluation Metrics

5.2.1. Metrics on Cognitive Feature Level. We define average feature hit at k (QH@k) to evaluate the ranking performance as follows,

$$\text{QH@}k(\tau^q, \tilde{\tau}^q) = \sum_{i=1}^{k} \mathbb{I}(\tilde{\tau}_i^q \in \tau^q(1:k)), \quad (5)$$

where τ^q is the ground-truth ranking list of all the features in all the tasks, $\tau^q(1:k)$ is the top k features in the list, $\tilde{\tau}^q$ is the predicted ranking list of all the features, and $\tilde{\tau}_i^q$ is the i-th ranked features in $\tilde{\tau}^q$. That is, QH@k calculates the number of features among top k in the predicted feature lists that are also in the ground truth (i.e., hits). Higher QH@k values indicate better prioritization performance.

We define a second evaluation metric weighted average feature hit at k (WQH@k) as follows:

$$\text{WQH@}k(\tau^q, \tilde{\tau}^q) = \sum_{j=1}^{k} QH@j(\tau^q, \tilde{\tau}^q)/k, \quad (6)$$

that is, WQH@k is a weighted version of QH@k that calculates the average of QH@j ($j = 1, \cdots, k$) over top k. Higher WQH@k indicates more feature hits and those hits are ranked on top in the ranking list.

5.2.2. Metrics on Cognitive Task Level. We use the mean of the top-g normalized ground-truth scores/predicted scores on the features of each cognitive task for a patient as the score of that task for that patient. For each patient, we rank the tasks using their ground-truth scores and use the ranking as the ground-truth ranking of these tasks. Thus, these scores measure how much relevant to AD the task indicates for the patients. We use the predicted scores to rank cognitive tasks into the predicted ranking of the tasks. We define a third evaluation metric task hit at k (NH_g@k) as follows to evaluate the ranking performance in terms of tasks,

$$\text{NH}_g\text{@}k(\tau_g^n, \tilde{\tau}_g^n) = \sum_{i=1}^{k} \mathbb{I}(\tilde{\tau}_{gi}^n \in \tau_g^n(1:k)), \quad (7)$$

where τ_g^n/$\tilde{\tau}_g^n$ is the ground-truth/predicted ranking list of all the tasks using top-g question scores.

6. Experimental Results

6.1. Overall Performance

TABLE 1: Overall Performance of PLTR in CV

d	QH@5	WQH@5	NH_1@1	NH_2@1	NH_3@1	NH_5@1	NH_{all}@1
10	**2.665**	3.136	0.605	0.701	0.713	0.725	0.683
10	2.647	**3.191**	0.599	0.677	0.707	0.725	0.677
10	2.569	2.957	**0.635**	0.707	0.689	0.719	0.653
10	2.623	3.073	0.623	**0.713**	0.707	0.719	0.671
50	2.467	2.992	0.605	0.695	**0.725**	0.725	0.653
30	2.491	3.080	0.563	0.689	0.713	**0.749**	**0.689**

The column "d" corresponds to the latent dimension. The best performance under each evaluation metric is in **fold**.

Table 1 presents the performance of PLTR in the CV setting. In terms of cognitive features from all tasks, PLTR is able to identify on average 2.665 out of the top-5 most relevant ground-truth cognitive features among its top-5 predictions. Corresponding to the real scenario to prioritize cognitive tasks that each patient should take, PLTR is able to identify the top-1 most relevant task for 74.9% of all the patients (i.e., NH_5@1). This indicates the strong power of PLTR in prioritizing cognitive features and in recommending relevant cognition tasks for real clinical applications. Note that in CV, each patient has only a few cognitive tasks in the testing set. Therefore, we only consider the evaluation at the top task in the predicted task rankings (i.e., only NH_g@1 in Table 1). In addition, as g in NH_g@1 increases in Table 1 (i.e., more top feature scores are used to score tasks), the performance of PLTR in terms of NH_g@1 first improves and then declines. This may indicate that different questions in a task may test different aspects related to AD, and PLTR is able to identify the most relevant features from each task.

Table 2 presents the performance of PLTR in the LOV setting. We first hold out 26 and 52 AD patients as testing patients, respectively. We determine these AD patients as the ones that have more than 10 similar AD patients in the training set with corresponding patient similarities higher than 0.67 and 0.62, respectively. When 26 patients are hold out for testing, PLTR is able to identify the top most relevant questionnaire for 84.6% of the testing patients (i.e., 22 patients) under NH_1@1. When 52 patients are hold out for testing, PLTR is able to identify for 80.8% of the testing patients (i.e., 42 patients) under NH_1@1. Note that the hold-out testing patients in LOV do not have any cognitive features. Therefore, the performance of PLTR as above demonstrates the strong capability of PLTR in identifying most AD related cognitive features based on imaging features only. Also note in Table 2, as the number of feature scores used to score cognitive tasks (i.e., g in NH_g@k) increases, the performance of PLTR in NH_g@1 first declines and then increases, and in NH_g@5 first increases. This indicates that PLTR can still prioritize the most relevant cognitive features among top in the predicted rankings.

TABLE 2: Overall Performance of PLTR in LOV

n	α	β	γ	d	QH@5	WQH@5	NH_1@1	NH_1@5	NH_2@1	NH_2@5	NH_3@1	NH_3@5	NH_5@1	NH_5@5	NH_{all}@1	NH_{all}@5
	0.5	1.5	1.0	30	**1.615**	**1.906**	0.846	3.231	0.577	3.385	0.231	3.654	0.308	3.346	0.808	3.692
	0.1	0.5	0.5	30	1.500	1.778	**0.846**	**3.269**	**0.577**	**3.538**	0.269	3.654	0.269	3.269	0.808	3.577
26	0.3	1.0	1.0	10	1.538	1.856	0.846	3.192	0.577	3.423	**0.308**	**3.731**	0.346	3.346	0.808	3.615
	0.3	1.5	1.0	10	1.577	1.851	0.846	3.192	0.577	3.462	0.308	3.654	**0.346**	**3.462**	0.808	3.654
	0.5	1.5	1.0	30	1.615	1.906	0.846	3.231	0.577	3.385	0.231	3.654	0.308	3.346	**0.808**	**3.692**
	0.5	0.5	1.0	50	**1.385**	**1.668**	0.788	3.212	0.423	3.654	0.115	3.750	0.288	3.423	0.788	3.423
	0.5	0.5	1.0	10	1.327	1.616	**0.808**	**3.269**	0.423	3.654	0.115	3.731	0.173	3.423	0.788	3.404
	0.5	0.5	1.0	30	1.327	1.652	0.788	3.212	**0.423**	**3.712**	0.115	3.750	0.269	3.423	0.788	3.404
52	0.5	0.5	1.0	30	1.327	1.652	0.788	3.212	0.423	3.712	**0.115**	**3.750**	0.269	3.423	0.788	3.404
	0.5	1.5	1.0	30	1.308	1.616	0.788	3.154	0.423	3.654	0.115	3.712	**0.288**	3.481	0.788	3.615
	0.5	1.5	1.0	10	1.288	1.581	0.808	3.173	0.423	3.596	0.115	3.750	0.192	**3.519**	0.788	3.635
	0.3	1.5	1.0	50	1.269	1.616	0.808	3.115	0.423	3.635	0.115	3.731	0.250	3.481	**0.788**	**3.635**

The column "n" corresponds to the number of hold-out testing patients. Best performance under each evaluation metric is in **bold**.

6.2. Case Study

When NH_1@1 achieves its optimal performance 0.846 for the 26 testing patients in LOV (i.e., the first row block in Table 2), the corresponding most common task that is prioritized for the testing patients is Rey Auditory Verbal Learning Test (RAVLT), including the following cognitive features: 1) trial 1 total number of words recalled; 2) trial 2 total number of words recalled; 3) trial 3 total number of words recalled; 4) trial 4 total number of words recalled; 5) trial 5 total number of words recalled; 6) total Score; 7) trial 6 total number of words recalled; 8) list B total number of words recalled; 9) 30 minute delay total; and 10) 30 minute delay recognition score. This task is also the most relevant task in the ground truth if tasks are scored correspondingly. RAVLT is a well-known cognitive test that assesses learning and memory, and has shown promising performance in early detection of AD [6]. A number of studies have reported high correlations between various RAVLT scores with different brain regions [7]. For example, RAVLT recall is associated with medial prefrontal cortex and hippocampus; RAVLT recognition is highly correlated with thalamic and caudate nuclei. Genetic analysis of $APOE\ \varepsilon 4$ allele, the most common variant of AD, reported its association with RAVLT score in an early-MCI (EMCI) study [5]. The fact that RAVLT is prioritized demonstrates the strong power of PLTR in prioritizing cognitive features to assist AD diagnosis.

Similarly, we find the top-5 most frequent cognitive tasks corresponding to the performance at NH_3@5=3.731 for the 26 hold-out testing patients. They are: Functional Assessment Questionnaire (FAQ), Clock Drawing Test (CDT), Weschler's Logical Memory Scale (LOGMEM), Rey Auditory Verbal Learning Test (RAVLT), and Neuropsychiatric Inventory Questionnaire (NPIQ). In addition to RAVLT discussed above, other top prioritized cognitive tasks have also been reported to be associated with AD or its progression. In an MCI to AD conversion study, FAQ, NPIQ and RAVLT showed significant difference between MCI-converter and MCI-stable groups [8]. These evidences further demonstrate the diagnostic power of our method.

7. Conclusions

We have proposed an innovative machine learning paradigm for prioritizing cognitive assessments according to their relevance to AD at the individual patient level. The paradigm tailors the cognitive biomarker discovery and cognitive assessment selection process to the brain morphometric characteristics of each individual patient. It has been implemented using a newly developed learning-to-rank method PLTR. Our empirical study on the ADNI data has yielded promising results to identify and prioritize individual-specific cognitive biomarkers as well as cognitive assessment tasks based on the individual's structural MRI data. The resulting top ranked cognitive biomarkers and assessment tasks have the potential to aid personalized diagnosis and disease subtyping, and to make progress towards enabling precision medicine in AD.

References

[1] J. Yan, T. Li *et al.*, "Cortical surface biomarkers for predicting cognitive outcomes using group l2,1 norm," *Neurobiol Aging*, vol. 36 Suppl 1, pp. S185–93, 2015.

[2] J. Scott and A. M. Mayo, "Instruments for detection and screening of cognitive impairment for older adults in primary care settings: A review," *Geriatr Nurs*, vol. 39, no. 3, pp. 323–329, 2018.

[3] Y. He, J. Liu, and X. Ning, "Drug selection via joint push and learning to rank," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. Epub ahead of print, 2018.

[4] M. W. Weiner, D. P. Veitch *et al.*, "The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement," *Alzheimers Dement*, vol. 13, no. 5, pp. 561–571, 2017.

[5] S. Risacher, S. Kim *et al.*, "The role of apolipoprotein e (apoe) genotype in early mild cognitive impairment (e-mci)," *Frontiers in Aging Neuroscience*, vol. 5, p. 11, 2013.

[6] E. Moradi, I. Hallikainen *et al.*, "Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease," *NeuroImage: Clinical*, vol. 13, pp. 415–427, 2017.

[7] M. L. F. Balthazar, C. L. Yasuda *et al.*, "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.*, vol. 16, no. 1, p. 205209OA, 2010.

[8] S. L. Risacher, A. J. Saykin *et al.*, "Baseline MRI Predictors of Conversion from MCI to Probable AD in the ADNI Cohort," *Current Alzheimer Research*, vol. 6, no. 4, pp. 347–361, 2009.