Network-Guided Sparse Learning for Predicting Cognitive Outcomes from MRI Measures*

Jingwen Yan[1,2], Heng Huang[3], Shannon L. Risacher[1], Sungeun Kim[1], Mark Inlow[4], Jason H. Moore[5], Andrew J. Saykin[1], and Li Shen[1,2,**]

[1] Radiology and Imaging Sciences, Indiana University School of Medicine, IN, USA
[2] School of Informatics, Indiana University Indianapolis, IN, USA
shenli@iu.edu
[3] Computer Science and Engineering, University of Texas at Arlington, TX, USA
[4] Department of Mathematics, Rose-Hulman Inst. of Tech., IN, USA
[5] The Geisel School of Medicine at Dartmouth College, NH, USA

Abstract. Alzheimer's disease (AD) is characterized by gradual neurodegeneration and loss of brain function, especially for memory during early stages. Regression analysis has been widely applied to AD research to relate clinical and biomarker data such as predicting cognitive outcomes from MRI measures. In particular, sparse models have been proposed to identify the optimal imaging markers with high prediction power. However, the complex relationship among imaging markers are often overlooked or simplified in the existing methods. To address this issue, we present a new sparse learning method by introducing a novel network term to more flexibly model the relationship among imaging markers. The proposed algorithm is applied to the ADNI study for predicting cognitive outcomes using MRI scans. The effectiveness of our method is demonstrated by its improved prediction performance over several state-of-the-art competing methods and accurate identification of cognition-relevant imaging markers that are biologically meaningful.

1 Introduction

Characterized by gradual loss of brain function, especially the memory and cognitive capabilities, Alzheimer's disease (AD) is a neurodegenerative disorder that has attracted tremendous research attention due to its significant public health impact and unknown disease mechanisms. Neuroimaging data, which characterize brain structure and function and its longitudinal changes, have been studied

* For the Alzheimer's Disease Neuroimaging Initiative. 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

** Corresponding author.

L. Shen et al. (Eds.): MBIA 2013, LNCS 8159, pp. 202–210, 2013.

as potential biomarkers for early detection of AD. Regression models have been studied to relate imaging markers to AD phenotypes such as cognitive outcomes.

Early applications focused on traditional regression models such as stepwise regression [6], which predicted cognitive outcomes one at a time. To address the relationships among multiple outcomes, multi-task learning strategies were recently proposed for achieving improved prediction performance. For example, $\ell_{2,1}$-norm [8, 11] was employed to extract features that have impact on all or most clinical scores; and a sparse Bayesian method [7] was proposed to explicitly estimate the covariance structure among multiple outcome measures.

Despite of the above achievements, few regression models take into account the covariance structure among predictors. Since brain structures tend to work together to achieve a certain function, brain imaging measures are often correlated with each other. A recent study proposed a prior knowledge guided regression model, using the group information to enforce the intra-group similarity [10]. However, the relationships among brain structures are much more complicated than a simple partitioning of all the structures into non-overlapping groups. To overcome this limitation, we present a new sparse learning method by introducing a novel network term to more flexibly model the relationship among brain imaging measures. This new model not only preserves the strength of $\ell_{2,1}$-norm to enforce similarity across multiple scores from a cognitive test, but also takes into account the complex network relationship among imaging predictors. We empirically demonstrate its effectiveness by applying it to the ADNI data.

2 Network-Guided Sparse Regression

Throughout this section, we write matrices as boldface uppercase letters and vectors as boldface lowercase letters. Given a matrix $\mathbf{M} = (m_{ij})$, its i-th row and j-th column are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. The Frobenius norm and $\ell_{2,1}$-norm (also called as $\ell_{1,2}$-norm) of a matrix are defined as $||\mathbf{M}||_F = \sqrt{\sum_i ||\mathbf{m}^i||_2^2}$ and $||\mathbf{M}||_{2,1} = \sum_i \sqrt{\sum_j m_{ij}^2} = \sum_i ||\mathbf{m}^i||_2$, respectively.

We focus on multi-task learning paradigm, where imaging measures are used to predict one or more cognitive outcomes. Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subseteq \Re^d$ be imaging measures and $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\} \subseteq \Re^c$ cognitive outcomes, where n is the number of samples, d is the number of predictors (feature dimensionality) and c is the number of response variables (tasks). Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$.

Motivated by using the ℓ_1 norm (Lasso, [5]) to impose sparsity on relevant features, the $\ell_{2,1}$ norm [3] was first proposed to taking into account the relationship among responses while still preserving the sparsity advantage of Lasso. The object function is:

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma||\mathbf{W}||_{2,1} \ . \tag{1}$$

This approach couples multiple tasks together, with ℓ_2 norm within tasks and ℓ_1 norm within features. While the ℓ_2 norm enforces the selection of similar features across tasks, the ℓ_1 norm helps achieve the final sparsity. It has been widely applied to capture biomarkers having affects across most or all responses. Yet in

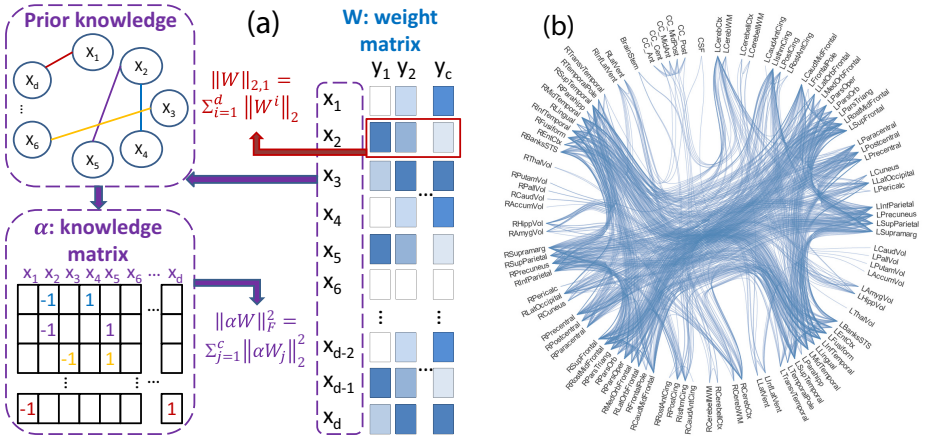

Fig. 1. (a) Illustration of the proposed NG-L21 model: This model enforces $\ell_{2,1}$-norm regularization ($||\mathbf{W}||_{2,1}$) to jointly select prominent predictors for all response variables, and introduces a new regularization term ($||\alpha\mathbf{W}||_F^2$) to flexibly model the relationship among predictors based on prior knowledge. (b) Correlation network among 99 FreeSurfer measures in an example cross-validation trial: Two measures are connected if their Pearson correlation coefficient, calculated from the training data, is ≥ 0.5.

this model the rows of $\mathbf{W}$ are equally treated, which implies that the underlying structures among predictors are ignored. To address this issue, Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) method [9] was proposed to exploit the interrelated structures within and between the predictor and response variables. It assumes 1) possible partition exists among predictors, and 2) predictors within one partition should have similar weights.

However, in practice the relationship among predictors may not be as simple as a straightforward partition. For example, imaging markers can be grouped by different brain circuitries, which may overlap with each other. In addition, instead of partitioning predictors into groups, the relationship among predictors can be represented more generally by a network (e.g., Figure 1(a)). To model these more complicated but more flexible structures among predictors, we propose a new *Network-Guided* $\ell_{2,1}$ *Sparse Learning (NG-L21)* model as follows.

The key idea here is to introduce a new regularization term ($||\alpha\mathbf{W}||_F^2$) to the $\ell_{2,1}$ model (Eq (1)) and formulate the objective function as:

$$\min_{\mathbf{W}} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1||\alpha\mathbf{W}||_F^2 + \gamma_2||\mathbf{W}||_{2,1} \ , \tag{2}$$

where α is a sparse matrix in which each row indicates a neighborhood relationship within a network of connected predictors.

Fig. 1(a) shows a schematic example of α as well as the entire NG-L21 model. A network is given as prior knowledge, where nodes are predictors. In this study, the network is constructed as follows: An edge (i, j) is inserted to the network if and only if $r(i, j)$ exceeds a given threshold (e.g., 0.5 used in our experiments),

where $r(i,j)$ is the Pearson correlation coefficient between predictors i and j calculated based on the training data. Fig. 1(b) shows an example correlation network. Based on the network, we can define the knowledge matrix α as follows: for each edge i,j in the network, we create a row in α with i-th entry as -1, j-th entry as 1 and all the other entries as zeros. The intuition is that the weight difference between two correlated predictors should be minimized, which is reflected by the new regularization term of $||\alpha\mathbf{W}||_F^2$. We call this model *NG-L21*. Instead of using -1 and 1 in α, we can fill in the actual $-r(i,j)$ and $r(i,j)$ values for each edge (i,j). Thus, the more correlated a feature pair is, the more constraint the pair is imposed by. We call this *weighted* model *NG-L21*$_w$.

Eq. (2) can be solved by taking the derivative w.r.t $\mathbf{W}$ and setting it to 0:

$$\mathbf{X}\mathbf{X}^T\mathbf{W} - \mathbf{X}\mathbf{Y}^T + \gamma_1\mathbf{D_1}\mathbf{W} + \gamma_2\mathbf{D_2}\mathbf{W} = 0, \tag{3}$$

where $\mathbf{D_1} = \alpha^T\alpha$, a matrix in which each row integrates all the neighboring relationships. For i-th row, it is the sum of all the rows in α whose i-th element is not zero. $\mathbf{D_2}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$. Thus, we have

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D_1} + \gamma_2\mathbf{D_2})^{-1}\mathbf{X}\mathbf{Y}^T, \tag{4}$$

where $\mathbf{W}$ can be efficiently obtained by solving the linear equation $(\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}_1 + \gamma_2\mathbf{D}_2)\mathbf{W} = \mathbf{X}\mathbf{Y}^T$. Following [9], an efficient iterative algorithm based on Eq. (4) can be easily developed as follows.

Input: $\mathbf{X}$, $\mathbf{Y}$
Initialize $\mathbf{W}^1 \in \mathbb{R}^{d\times c}$, $t = 1$;
while *not converge* **do**
1. Calculate the diagonal matrices $\mathbf{D}_2^{(t)}$, where the i-th diagonal element of $\mathbf{D}_2^{(t)}$ is $\frac{1}{2\|\mathbf{w}_t^i\|_2}$;
2. $\mathbf{W}^{(t+1)} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}_1 + \gamma_2\mathbf{D}_2^{(t)})^{-1}\mathbf{X}\mathbf{Y}^T$;
3. $t = t + 1$;
end
Output: $\mathbf{W}^{(t)} \in \mathbb{R}^{d\times c}$.

Next, we prove that the above algorithm converges to the global optimum. According to Step 2 in the algorithm, we have

$$W_{t+1} = \min_W Tr(X^TW - Y)^T(X^TW - Y) + \gamma_1 Tr(W^TD_1W) + \gamma_2 Tr(W^TD_{2(t)}W)$$

$$\begin{aligned} &Tr(X^TW_{t+1} - Y)^T(X^TW_{t+1} - Y) + \gamma_1 Tr(\alpha W_{t+1})^T\alpha W_{t+1} + \gamma_2\sum_{i=1}^{d}\left\|w_{t+1}^i\right\|_2 \\ \leq\ &Tr(X^TW^{(t)} - Y)^T(X^TW^{(t)} - Y) + \gamma_1 Tr(\alpha W_t)^T\alpha W_t + \gamma_2\sum_{i=1}^{d}\left\|w_t^i\right\|_2 \end{aligned}$$

Finally we have:

$$\begin{aligned} &\left\|X^T W_{t+1} - Y\right\|_F^2 + \gamma_1 \left\|\alpha W_{t+1}\right\|_F^2 + \gamma_2 \left\|W_{t+1}\right\|_{2,1} \\ \leq\ &\left\|X^T W_t - Y)\right\|_F^2 + \gamma_1 \left\|\alpha W_t\right\|_F^2 + \gamma_2 \left\|W_t\right\|_{2,1} \end{aligned}$$

The last but one step holds, because [8] for any vector w and w_0, we have $\|w\|_2 - \frac{\|w\|_2^2}{2\|w_0\|_2} \leq \|w_0\|_2 - \frac{\|w_0\|_2^2}{2\|w_0\|_2}$. Thus, the algorithm decreases the objective value in each iteration. Since the problem is convex, satisfying the Eq. (2) indicates that W is the global optimum solution. Therefore, this algorithm will converge to the global optimum of the problem.

3 Experimental Results

3.1 Data and Experimental Setting

The magnetic resonance imaging (MRI) and cognitive data were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. One goal of ADNI has been to test whether serial 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 AD. For up-to-date information, see www.adni-info.org.

This study included 179 AD and 205 healthy control (HC) participants (Table 1). For each baseline MRI scan, FreeSurfer V4 was employed for brain segmentation and cortical parcellation, and extracted 73 thickness measures and 26 volume measures. These 99 imaging measures were used to predict three sets of cognitive scores [1] separately: Mini-Mental State Exam (MMSE), Rey Auditory Verbal Learning Test (RAVLT, including 5 scores shown in Table 2 as joint response variables), and Wechsler Memory Scale III logical memory (LogMem). Using the regression weights derived from the HC participants, all the imaging measures were pre-adjusted for the baseline age, gender, education, handedness, and intracranial volume, and all the cognitive measures were pre-adjusted for the baseline age, gender, education and handedness.

Regression was performed separately on each cognitive task (MMSE, RAVLT, or LogMem) using the MRI measures as predictors, where the proposed NG-L21 and NG-L21$_w$ methods and three competing regression methods (Linear, Ridge and L21) were evaluated. Pearson correlation coefficients r between the actual

Table 1. Participant characteristics

Category	HC	AD
Number	205	179
Gender(M/F)	112/93	98/81
Handness(R/L)	191/14	167/12
Age(mean±std)	76.07±4.98	75.58±7.51
Education	16.17±2.74	14.85±2.10

Table 2. RAVLT scores

Score ID	Description
TOTAL	Total score of the first 5 learning trials
TOT6	Trial 6 total number of words recalled
TOTB	List B total number of words recalled
T30	30 minute delay number of words recalled
RECOG	30 minute delay recognition score

Table 3. Mean prediction performance over five cross-validation trials is reported for each experiment, where the performance is measured by correlation coefficients between the actual and predicted cognitive scores in each trial. The p values, calculated from the paired sample t test between two sets of cross-validation correlation coefficients, are shown for comparing two proposed methods with L21.

		TOTAL	T30	RECOG	TOT6	TOTB	MMSE	LogMem
Correlation Coefficients	NG-L21$_w$	0.6511	0.5926	0.5636	0.6137	0.4630	0.7574	0.7076
	NG-L21	0.6505	0.5925	0.5634	0.6130	0.4606	0.7575	0.7068
	L21	0.6306	0.5792	0.5469	0.5967	0.4441	0.7488	0.6977
	Ridge	0.6215	0.5415	0.5368	0.5814	0.4406	0.7478	0.6870
	Linear	0.5396	0.4299	0.4533	0.4741	0.3525	0.6708	0.6071
p values	L21 vs NG-L21$_w$	0.0029	0.0488	0.0476	0.0105	0.0021	0.0104	0.0119
	L21 vs NG-L21	0.0037	0.0469	0.0577	0.0129	0.0024	0.0088	0.0098

and predicted cognitive scores were computed to measure the prediction performance. Five-fold cross validation was employed to obtain an unbiased estimate of regression performance. Paired t-test was applied to the cross-validation results to evaluate whether performance significantly differ between two methods.

3.2 Network Construction

Each MRI measure was treated as a network node, and the connectivity network among 99 MRI measures was constructed based on their pairwise Pearson correlation coefficients. Rather than including all pairwise links, threshold 0.5 was applied to connect only highly correlated nodes. For nodes that were not very correlated, constraints should not be imposed to make their regression weights similar to each other. A network was created using only the training data. Thus, our 5 cross-validation trials yielded 5 networks that were almost identical. One example was shown in Fig. 1(b), where totally 85 structures out of 99 had qualified links with correlation coefficient higher than 0.5. To incorporate this connectivity information into the proposed models, we examined the weighted network in NG-L21$_w$ and non-weighted one in NG-L21. While in the weighted network each link between structures was assigned the value of their correlation coefficient, non-weighted network treated all the links equally.

3.3 Prediction Performance and Biomarker Identification

Shown in Table. 3 is the performance comparison among all five methods. NG-L21 and NG-L21$_w$ both demonstrated an improved performance over the other three methods, while L21 performed the best among the three competing methods. The difference between NG-L21 and NG-L21$_w$ was minor, and the weighted method only led to slight improvements than non-weighted one for TOTAL, TOT6 and LogMem. This could be partially due to the small range of the edge

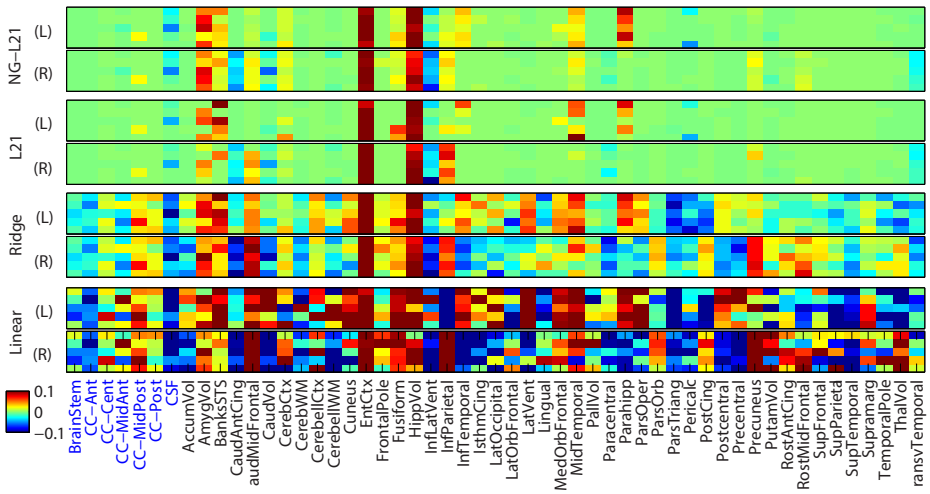

Fig. 2. Heat maps of regression weights for predicting MMSE scores using MRI measures. Five-fold cross-validation regression weights are plotted for NG-L21, L21, Ridge and Linear regression models respectively. Each panel corresponds to the measures from the left (L) or right (R) hemisphere. The measures shown in the first seven column (highlighted in blue) are unilateral, and the remaining ones are bilateral.

weights (0.5-1.0). To further make sure the improvements of the proposed methods were not by chance, we calculated p-values from the paired sample t test between two sets of cross-validation correlation coefficients from two different methods. According to the last two rows in Table 3, both NG-L21 and NG-L21$_w$ outperformed L21 significantly for predicting all the tested cognitive outcomes.

Finally, we examined the biomarkers identified by different methods. Shown in Fig. 2 was an example comparison of resulting regression coefficients among four methods (NG-L21$_w$ was extremely similar to NG-L21 and thus not shown), where 99 MRI measures were used to predict MMSE score. Each methods occupied two panels, representing the left and right hemispheres respectively. Apparently NG-L21 and L21 both showed sparse patterns while Linear and Ridge methods yielded non-sparse patterns that were hard to interpret. In addition, NG-L21 tended to select slightly more features than L21 as correlated measures were forced to be selected together in NG-L21, which yielded not only more stable patterns across cross-validation trials but also more biologically meaningful and more interpretable results. The MRI markers identified by NG-L21 yielded promising patterns that matched prior knowledge on neuroimaging and cognition. MMSE measured overall cognitive impairment; and thus its result (Fig. 2) included important AD-relevant imaging markers such as hippocampus, amygdala, inferior lateral ventricle, entorhinal cortex, and middle temporal gyri. Both LogMem and RAVLT were memory tests; and thus their results (Fig. 2) included regions relevant to memory, such as hippocampus, amygdala, entorhinal cortex, middle temporal gyri and parahippocampal gyri.

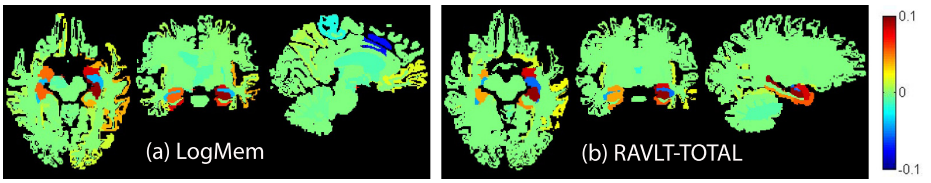

Fig. 3. NG-L21 weight maps on brain for (a) RAVLT-TOTAL and (b) LogMem scores

4 Conclusions

We presented a new network-guided sparse learning model NG-L21 and demonstrated its effectiveness by applying it to the ADNI data for predicting cognitive outcomes from MRI scans. While spatial correlation had been considered in several voxel-based feature selection and learning models [2, 4], the existing studies on predicting cognitive outcomes from ROI-based MRI measures often ignored [7, 8] or simplified [10] the relationships among these ROI predictors. The proposed NG-L21 model aimed to bridge this gap and introduced a novel network term to flexibly model the relationship among imaging markers. An efficient algorithm was developed to implement this model and was shown to be able to achieve global optimum. Its application to the ADNI data exhibited the following strengths of the NG-L21 model: (1) It could flexibly take into account the complex relationship among imaging markers in a network format rather than a simple grouping scheme used in [10]. (2) As a multi-task sparse learning framework, it could identify a compact set of imaging markers related to multiple cognitive outcomes. (3) By considering the correlation among predictors, it yielded not only improved prediction performance but also more stable cross-validation feature selection patterns. Different from traditional Lasso and L21 methods that tended to select only one relevant feature from a group of highly correlated ones, the NG-L21 model could jointly identify these correlated features, making the results more stable and easier to interpret.

Acknowledgement. This research was supported by NIH R01 LM011360, U01 AG024904, RC2 AG036535, R01 AG19771, P30 AG10133, and NSF IIS-1117335 at IU, by NSF CCF-0830780, CCF-0917274, DMS-0915228, and IIS-1117965 at UTA, and by NIH R01 LM011360, R01 LM009012, and R01 LM010098 at Dartmouth.

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; Alzheimer's Association; Alzheimer's 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; Innogenetics, N.V.; 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, K01 AG030514, and the Dana Foundation.

References

1. Aisen, P.S., Petersen, R.C., et al.: Clinical core of the Alzheimer's Disease Neuroimaging Initiative: progress and plans. Alzheimers Dement 6(3), 239–246 (2010)
2. Liu, M., Zhang, D., Yap, P.-T., Shen, D.: Tree-guided sparse coding for brain disease classification. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 239–247. Springer, Heidelberg (2012)
3. Obozinski, G., Taskar, B., Jordan, M.: Multi-task feature selection. Technical Report, Statistics Department, UC Berkeley (2006)
4. Sabuncu, M.R., Van Leemput, K.: The relevance voxel machine (RVoxM): A bayesian method for image-based prediction. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 99–106. Springer, Heidelberg (2011)
5. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288 (1996)
6. Walhovd, K., Fjell, A., et al.: Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobiol Aging 31(7), 1107–1121 (2010)
7. Wan, J., et al.: Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease. In: CVPR 2012, pp. 940–947 (2012)
8. Wang, H., Nie, F., et al.: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: ICCV 2011, pp. 557–562 (2011)
9. Wang, H., Nie, F., et al.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics 28(2), 229–237 (2012)
10. Yan, J., et al.: Multimodal neuroimaging predictors for cognitive performance using structured sparse learning. In: Yap, P.-T., Liu, T., Shen, D., Westin, C.-F., Shen, L. (eds.) MBIA 2012. LNCS, vol. 7509, pp. 1–17. Springer, Heidelberg (2012)
11. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 59(2), 895–907 (2012)