PREDICTING TEMPORAL LOBE VOLUME ON MRI FROM GENOTYPES USING L^1-L^2 REGULARIZED REGRESSION

Omid Kohannim[1], Derrek P. Hibar[1], Neda Jahanshad[1], Jason L. Stein[1], Xue Hua[1], Arthur W. Toga[1], Clifford R. Jack, Jr.[2], Michael W. Weiner[3,4], Paul M. Thompson[1], and the Alzheimer's Disease Neuroimaging Initiative

[1]Laboratory of Neuro Imaging, Dept. of Neurology, UCLA School of Medicine, Los Angeles, CA, USA, [2]Mayo Clinic, Rochester, MN, USA, [3]Depts. of Radiology, Medicine and Psychiatry, UC San Francisco, San Francisco, CA, USA, [4]Dept. of Veterans Affairs Medical Center, San Francisco, CA, USA,

ABSTRACT

Penalized or sparse regression methods are gaining increasing attention in imaging genomics, as they can select optimal regressors from a large set of predictors whose individual effects are small or mostly zero. We applied a multivariate approach, based on L^1-L^2-regularized regression (elastic net) to predict a magnetic resonance imaging (MRI) tensor-based morphometry-derived measure of temporal lobe volume from a genome-wide scan in 740 Alzheimer's Disease Neuroimaging Initiative (ADNI) subjects. We tuned the elastic net model's parameters using internal cross-validation and evaluated the model on independent test sets. Compared to 100,000 permutations performed with randomized imaging measures, the predictions were found to be statistically significant ($p \sim 0.001$). The rs9933137 variant in the *RBFOX1* gene was a highly contributory genotype, along with rs10845840 in *GRIN2B* and rs2456930, discovered previously in a univariate genome-wide search.

Index Terms— Neuroimaging, MRI, Prediction, Elastic net, Imaging Genetics

1. INTRODUCTION

Many early studies in imaging genetics explored univariate associations between genotypes and imaging measures, assuming each gene acted independently. One disadvantage of such studies is their limited statistical power to detect gene effects on the brain. Meta-analyses such as the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) project [1] have boosted statistical power, by analyzing MRI and genome-wide genotype data from over 20,000 subjects, gaining power from very large sample sizes. Multivariate approaches, which simultaneously consider entire sets of genotypes, sets of voxels in an image, or both, have also become more popular [2], as they also handle potential problems in high-dimensional data, such as highly correlated predictors, where almost all have no detectable effects.

In [2], we reviewed several recent multivariate, imaging genetics studies that applied principal component regression [3], sparse reduced rank regression [4], or independent components analysis [5] to discover genetic influences on the brain that would have been missed by using only univariate techniques. Regularized, sparse regression methods, in particular, use penalty terms to tackle the problems of high dimensionality (e.g., having more predictors than samples), multiple highly correlated measures, and multiple comparisons across an image, the genome, or both. The "elastic net" combines L^1- and L^2-norm regularization and benefits from the advantages of both methods, to handle high-dimensional, highly correlated data. The algorithm takes advantage of the sparsity properties of L^1 (Least Absolute Shrinkage and Selection Operator, or LASSO), along with the stability of L^2 (ridge) regression [6]. Here, we introduce an elastic net approach to predict an imaging measure from top genotypes. We aim to incorporate top genetic variants (i.e., single nucleotide polymorphisms or SNPs), screened based on univariate genome-wide search (as in a genome-wide association analysis or GWAS), into an elastic net model, to predict temporal lobe volume on MRI. Recently, the elastic net has been applied to genomics [7,8], for jointly considering genetic polymorphisms as well as imaging [9], to integrate large numbers of imaging and clinical predictors. More recently, the algorithm has also been used to detect multi-SNP associations with hippocampal surface morphometry [10], and to integrate imaging and proteomic data in Alzheimer's disease [11].

We hypothesize that this doubly regularized, multivariate regression method would allow us to make significant predictions of MRI-derived temporal lobe volume from

genotypes. This predictive approach, we propose, may have implications for early, personalized risk assessment of brain disorders such as Alzheimer's disease, where the temporal lobes undergo significant atrophy.

2. METHODS

2.1. MRI Measures

ADNI subjects were scanned with a standard MRI protocol optimized for reproducibility and consistency across 58 sites in North America. Temporal lobe volumes were derived from an anatomically defined region-of-interest (ROI) on three-dimensional maps of relative volumes generated with tensor-based morphometry (TBM), a well-established method to map volumetric differences in the brain [12]. Temporal lobe volume is particularly interesting, as this structure is prone to atrophy in Alzheimer's disease (AD). There is interest in discovering genes that may promote or resist the atrophy, or contribute to normal variations in its volume. A total of 740 subjects with both imaging and genotype data were included (173 with AD, 361 with mild cognitive impairment or MCI, and 206 cognitively healthy controls; 438 men and 302 women; mean ± SD age: 75.55 ± 6.79 years).

2.2. Genotypes

Genotyping procedures for ADNI are described in [13]. SNPs with minor allele frequencies less than 0.01 and Hardy-Weinberg equilibrium *p*-values less strict than 5.7 x 10^{-7} were excluded. Genotypes were imputed to infer missing information.

2.3. Elastic net method

The elastic net [6] is a form of penalized regression, where both L^1 and L^2 regularizations are introduced into the standard multiple linear regression model, as formulated below for n subjects and p predictors:

$$\beta^* = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \qquad (1)$$

Here, y represents the vector whose n components are the imaging measure for each subject, after adjusting for sex and age (residuals of regression). X is the n x p matrix of genotypes for top genetic variants across the genome. β^* represents the vector of fitted regression coefficients for each SNP's effect on the imaging measure. λ_1 is a positive weighting parameter on the L^1 penalty, which promotes sparsity in the resulting set of fitted regression coefficients, as many coefficients are likely to be exactly zero. λ_2 is a positive weighting parameter on the L^2 penalty, which promotes stability in the regularization path and precludes a limit on how many variables are selected (in strict LASSO, at most n variables can be selected in an n by p case).

In ten separate experiments (**Figure 1**), we randomly split the data into training sets with $3n/4$ and testing sets with $n/4$ subjects. Standard univariate associations were performed for all ~500,000 genotyped variants with the imaging measure, using the training set only, and top 4,000 SNPs were then fed into the elastic net algorithm. This is a common pre-screening step that has been used in similar contexts [7]. Leave-one-out cross-validation was performed within the training sets to determine the optimal penalty parameters with the mean squared error criterion. Both λ_1 and α are optimized with a grid search, where $a = \lambda_2 / (\lambda_1 + \lambda_2)$, such that the penalty term of (1), P, is restated as below:

$$P = \alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1 \qquad (2)$$

Mean squared error is commonly minimized for parameter tuning using cross-validation, similarly to previous studies in this context [10,11]. To avoid bias, cross-validation for selecting hyperparameters is done separately from evaluation of the model. Models trained to have optimal penalty parameters were tested on the test sets to obtain mean squared errors for predicting the imaging measure from genotypes. For our analyses, we used the 'glmnet' package [14] implemented in R (http://cran.r-project.org). This optimizes model fitting parameters via an efficient, coordinate descent algorithm.

A similar procedure was repeated 100,000 times. To reduce computational time, unlike the actual experiments, only the optimal penalty parameters were used and a fixed set of top 4,000 SNPs from a univariate genome-wide search were incorporated into the models. Imaging measures were randomly assigned to all subjects, after which the data was randomly split into training and testing sets as above. Mean squared errors for prediction of test set temporal lobe volumes were then obtained for each permutation.

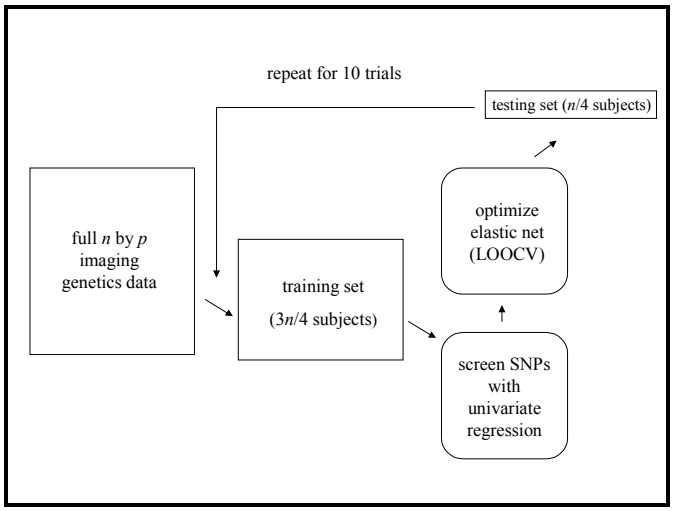

Figure 1: **Validation framework.** Different loops of cross-validation are necessary to prevent over-fitting of a predictive model. We pre-screen the single nucleotide polymorphisms (SNPs) for dimension reduction, and elastic net parameter optimization, is

only performed within the training data. The mean squared errors of predictions in 10 separate trials on independent test sets are averaged. LOOCV = Leave-one-out cross-validation.

Standard multiple regression cannot be used in our scenario, as the multivariate analysis for all top SNPs would fail (i.e., the model fitting equation would be ill-conditioned), as there are many more variants than subjects ($p >> n$ problem).

To perform *post-hoc*, exploratory tests on our top SNPs, we created voxelwise statistical maps to reveal the spatial profile of associations with regional brain volumes. We fitted linear associations at each voxel, adjusted for covariates (sex and age). To correct for multiple spatial comparisons, we used a regional False Discovery Rate (FDR) method, which is now fairly standard in neuroimaging [15].

3. RESULTS

We averaged the mean squared errors of the optimized predictive models on test sets. An average mean squared error of 3,147 was obtained with the elastic net predictor in independent sets of test subjects. The average mean squared error in the 100,000 permutations was 4,257 with a standard deviation of 397. Compared to the distribution of the errors across the permutations (**Figure 2**), the *p*-value is found to be close to 0.001.

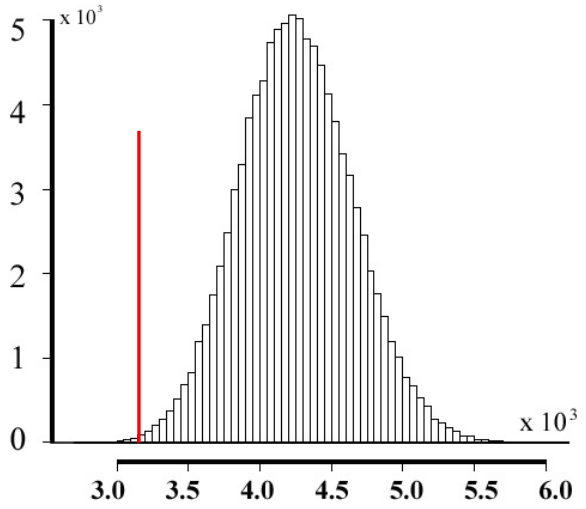

Figure 2: Distribution of mean squared errors for the 10^5 simulations conducted with the optimal elastic net parameters. Errors are approximately normally distributed (mean, 4,257; SD: 397). 131 permutations had errors smaller than our predictive model's error (*red line*), yielding an empirical *p*-value ~ 0.001.

To investigate which genetic variants contributed most to the predictions, we examined the average absolute values of coefficients for each fitted predictor. Out of the 4,000 variants incorporated into the elastic net models in each of the ten trials, 105 were screened for all trials. We investigated the coefficients obtained by these SNPs. The top ten are shown in **Table 1**. To ensure that the findings were robust, we also counted the number of times the variants received nonzero coefficients across the ten runs (**Table 1**). With permutations, each SNP obtained a nonzero coefficient only about 2.0 ± 0.5 SD times, on average.

SNP	Gene	Chr	$\|\beta\|_{average}$	$\|\beta\|$>0 count
rs2456930	-	15	2.32	10
rs10518480	-	4	1.96	10
rs17476752	-	5	1.78	9
rs9933137	*RBFOX1*	16	1.75	8
rs10845840	*GRIN2B*	12	1.64	9
rs997972	-	20	1.50	9
rs1929933	*GLDC*	9	1.44	9
rs1564348	*SLC22A1*	6	1.41	9
rs309800	-	4	1.37	10
rs11204135	-	8	1.33	10

Table 1: List of single nucleotide polymorphisms (SNPs) with the highest contribution to the elastic net models predicting temporal lobe volume on MRI. These ten SNPs had the largest elastic net coefficients (absolute values), and their selection was robust, as they obtained nonzero coefficients at least 8 out of the 10 total trials. Corresponding gene names and chromosome numbers are displayed for the variants.

We noted that rs10845840 in the *GRIN2B* gene and the intergenic rs2456930, which were the top findings with a univariate genome-wide search [16], also appeared in our top list, which is a re-assuring validation. Interestingly, rs9933137 in the *RBFOX1* gene also obtained a very high mean $|\beta|$ and outperformed the top univariate SNP in *GRIN2B*. To explore the profile of effects of the *RBFOX1* SNP on temporal lobes in more detail, we performed an exploratory, *post-hoc* voxelwise test, shown in **Figure 3**.

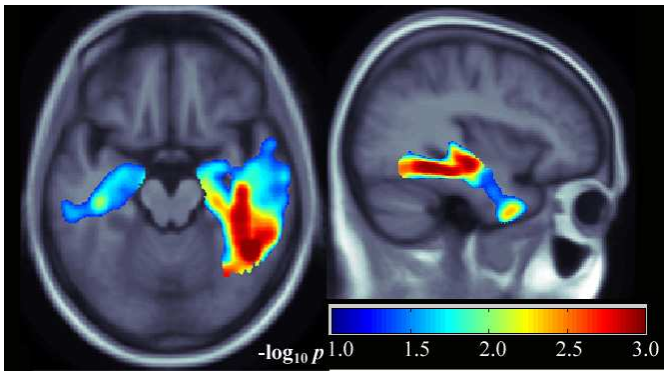

Figure 3: The *post-hoc* voxelwise effects of the *RBFOX1* rs9933137 polymorphism are shown on TBM-derived maps of the temporal lobes, using linear regression. Volumetric change at each voxel is linearly regressed against the genetic variant, along with covariates such as sex and age. *P*-values for the associations are corrected for multiple spatial comparisons using regional false discovery rate (FDR). Warmer colors represent more significant

effects. Images are in radiological convention. Results survived multiple comparisons correction across both lobes, but the left temporal lobe showed stronger effects (also seen in the left sagittal slice). Although this does not add new information to the multivariate, prediction study, it confirms that the highly predictive polymorphism's diffuse effects on the temporal lobes at a voxel-by-voxel basis.

4. CONCLUSION

We proposed a multivariate model to predict an imaging measure from genotypes, using L^1-L^2 regularized regression, also known as the elastic net. We split 740 ADNI subjects into training and test sets in ten separate trials. We optimized elastic net parameters in the training set using leave-one-out cross-validation, and predictions were made on the independent test sets. This is a rigorous predictive framework, as it avoids the overfitting that can arise if training data are used for testing. We also compared the performance of our predictor with that of 10^5 permutations, where MRI measures were randomly assigned to the subjects. Our predictions were significantly better than those made by random models. Although the main goal of our study was prediction rather than discovery, we also looked for the variants that most strongly contributed to the predictions. Using average elastic net coefficients as a metric, we found a single nucleotide polymorphism in the *RBFOX1* gene to be most contributory to the predictive models, which also showed significant 3D effects on the temporal lobes. This gene, also known as *A2BP1*, has been previously characterized as an autism risk gene [17], and regulates neuronal excitation in the brain [18]. Interestingly, it has also been discovered in another sparse regression imaging genetics study as a highly significant gene [19]. Future studies are needed to compare the performance of this predictor with other multivariate techniques. Pre-screening of genetic variants, which was done as a way of reducing dimensionality similarly to previous studies [7], may be a limitation, as it might lead to missing potential effects from contributory genes. Furthermore, applying multi-voxel methods [4,5,19] and incorporating biological pathway information may yield more statistically powerful predictions.

5. ACKNOWLEDGMENTS

ADNI data collection was supported by federal and private funds including NIH grants U01 AG024904, P30 AG010129, K01 AG030514, and the Dana Foundation. The ADNI Genetics Core, led by Andrew Saykin, performed the ADNI genotyping. OK was partially supported by the UCLA Medical Scientist Training Program. Algorithm development was supported by AG016570, EB01651, RR019771 (to PT).

6. REFERENCES

[1] The ENIGMA Consortium (2011). "Genome-Wide Association Meta-Analysis of Hippocampal Volume: Results from the ENIGMA Consortium", *Organization for Human Brain Mapping meeting*, Quebec City, Canada, June 2011, http://enigma.loni.ucla.edu/

[2] D.P. Hibar, et al., "Multilocus Genetic Analysis of Brain Images," *Front. Genet.* 2(73), pii:00011, 2011.

[3] D.P. Hibar et al., "Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects," *NeuroImage* 56:1875-1891, 2011.

[4] M. Vounou et al., "Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach," *NeuroImage* 53:1174-1159, 2010.

[5] J. Liu et al., "Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Hum. Brain Mapp.* 30:241-255, 2009.

[6] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J Royal Stat. Soc. B* 67:301-320, 2005.

[7] S. Cho et al., "Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis," *Ann. Hum. Genet.* 74:416-428, 2010.

[8] S. Cho et al., "Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis," *BMC Proc.* 3(Suppl 7):S25, 2009.

[9] F. Bunea et al., "Penalized least squares regression methods and applications to neuroimaging," *NeuroImage* 55:1519-1527, 2011.

[10] J. Wan et al., "Hippocampal surface mapping of genetic risk factors in AD via sparse learning models," *MICCAI* 14:376-383, 2011.

[11] L. Shen et al., "Identifying Neuroimaging and Proteomic Biomarkers for MCI and AD via the Elastic Net," *Lect. Notes Comput. Sci.* 7012:27-34, 2011.

[12] X. Hua et al., "3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry," *NeuroImage* 41:19-34, 2008.

[13] A.J. Saykin et al., "Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims progress and plans," *Alz. Dement.* 6(3):265-273, 2010.

[14] J. Friedman et al., "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Software* 33:1-22, 2010.

[15] D.R. Langers, "Enhanced signal detection in neuroimaging by means of regional control of the global false discovery rate," *NeuroImage* 38:43-56, 2007.

[16] J.L. Stein et al., "Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease," *NeuroImage* 51:542-554, 2010.

[17] C.L. Martin et al., "Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism," *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 144:869-876, 2007.

[18] L.T. Gehman et al., "The splicing regulator RBFOX1 (A2BP1) controls neuronal excitation in the mammalian brain," *Nat. Genet.* 43:706-711, 2011.

[19] M. Vounou et al., "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease," *NeuroImage* 60(1):700-716, 2011.