Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____	_____
Yajie Liu	Date

Correlation analysis between MRI brain image and gene expression using ADNI data

By

Yajie Liu
Master of Public Health

Department of Biostatistics and Bioinformatics

Steve Qin, PhD
(Thesis Advisor)

Xiangqin Cui, PhD
(Reader)

Correlation analysis between MRI brain image and gene expression using ADNI data

By

Yajie Liu

B.E., Southwest Jiaotong University, 2017

M.Ms, East China University of Science and Technology, 2020

Thesis Committee Chair: Steve Qin, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2022

Abstract

Correlation analysis between MRI brain image and gene expression using ADNI data

By Yajie Liu

Background: With the development of medical imaging studies, numerous studies investigated the correlations with brain-related disease status, age, and the prognostic or diagnostic biomarkers for brain disease diagnosis [1]. There was a kind of regression called the image-on-scalar model applied in this area. called image-on-scalar model aims to delineate the relationship between voxels or areas of interest (ROI) and a set of covariates of interest, such as demographic data, clinical features, and gene expression data, in which images were treated as a functional response variable [2].

Objectives: Examine the ability of elastic net regression in identifying genes that have an impact on brain MRI and examine whether there is any detectable correlation between age and MRI images.

Methods: Elastic net regression model was applied to examine the correlation. We built one model for each gene. In each model, we treated gene expression as dependent variable and image data as the independent variable, gender was tested as an independent variable. MRI image data, gene expression data, and age were from ADNI. And MRI image data were converted into 4232-dimension arrays by the auto-encoder method.

Results: We have selected the top 1000 genes with the highest variances to build elastic net regression models, and the top 25 models with the highest R-squares were analyzed. There were 12 of 25 models identified genes that were brain-related with the R-square of the model greater than 0.4. Models including gender had better performance than models that didn't include gender. The R-square of examining the correlation between age and MRI image data had an R-square of about 0.45 and considered gender as a cofounder couldn't improve the model performance.

Conclusions: In this study, we introduced an application of elastic net regression in identifying genes that have an impact on brain MRI, or in detecting correlation between age and MRI images. We found that the elastic net regression had relatively good in identifying genes that have an impact on brain MRI when considering gender effects. And the elastic net regression didn't have a good performance in detecting the correlation between age and MRI image data.

Correlation analysis between MRI brain image and gene expression using ADNI data

By

Yajie Liu

B.E., Southwest Jiaotong University, 2017

M.Ms, East China University of Science and Technology, 2020

Thesis Committee Chair: Steve Qin, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2022

Acknowledgement

First of all, I would like to thank the Rollins School of Public Health at Emory University for always providing me with all sorts of support. The two-year study at RSPH built my knowledge system in Biostatistics. During the two-year study, I have found my research interests in bioinformatics and set my career direction for the further.

I would like to thank my thesis advisor Steve Qin, who provided me the opportunity to do research in bioinformatics and is always willing to support me when I need help. During doing the thesis, he shared expertise, provided sincere and valuable guidance, and encouraged me when I felt disappointed.

I would like to thank the team member Shaojun Yu who helped a lot in the progress of doing my thesis. He provided me with all the ADNI data and the 4232-dimension array data converted from MRI image data which were the source data for my study. And thank him for him friendly and willing to help. He helped me in figuring out problems when I got stuck.

Finally, I would like to thank my family. Thank them for supporting my study abroad and comforting me when I'm desperate. Although we are thousands of miles apart, their love made me feel not alone.

Table of contents

1. Introduction

With the development of medical imaging studies, numerous studies investigated the correlations with brain-related disease status, age, and the prognostic or diagnostic biomarkers for brain disease diagnosis [1]. There is a kind of regression called the image-on-scalar model applied in this area. Image-on-scalar model aims to delineate the relationship between voxels or areas of interest (ROI) and a set of covariates of interest, such as demographic data, clinical features, and gene expression data, in which images are treated as a functional response variable [2].

Previous studies have developed several methods for fitting the image-on-scalar regression. Univariate analysis such as using the general linear model (GLM) with pre-smoothed imaging data through a kernel convolution is a traditional method. However, GLM usually has low power, low efficiency, low accuracy, and a high false-positive rate [3]. Another kind of model considers all the voxels of the interested variables as a tensor, such as parsimonious tensor response regression [4]. One of the challenges for this model is how to reduce dimensions in an efficient and accurate way. Additional, deep neural networks models, such as the neural network-based image-on-scalar regression model was applied in imaging studies.

In our study, we aimed to examine the ability of elastic net regression in identifying genes that have an impact on brain MRI and examine whether there is any detectable correlation between age and MRI images. A machine learning method elastic net, which can achieve multivariate analysis and automatically select associated covariates, was applied to build the image-on-scalar regression to identify the correlation between AD and age or gene expression.

2. Methods

2.1 Data

2.1.1 Data source

All data were obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a National Instituted of Health (NIH) funded project launched in 2004. ADNI unites researchers between academia and industry to determine the biomarker of AD in clinical, cognitive, imaging, genetic and biochemical areas. ADNI researchers collected and shared and utilized those data as predictors to predict and prevent AD.

Amyloid Beta (Aβ), phosphorylated tau, synaptic loss, and neurodegeneration are hallmark lesions of AD. However, Clinical/cognitive evaluations are ineffective at detecting pathologic changes in Alzheimer's disease due to a lack of sensitivity and specificity. Biomarkers, on the other hand, are more consistently utilized to identify people who are at risk of cognitive decline and to track illness progression. To determine the relationship of biomarkers to baseline clinical status and cognitive decline, ADNI collected MRI (anatomic, diffusion, perfusion, and resting-state images), amyloid PET (18F-FDG-PET (FDG PET)), CSF (Aβ, total tau, phosphorylated tau, and other proteins), AV-1451 PET, and genetic and autopsy data.

In our study, MRI data, gene expression data from blood samples, and the basic demographic data (gender and age) were used.

2.1.2 Image data

Image data were from ADNI project, which enrolled participants from 59 clinical sites and collected data from 57 imaging centers. Coronal plane slices extracted from ADNI MRI data post-processed using FreeSurfer were the raw data. Although there were follow-ups for each

participant, only the image data collected at baseline were used. To analyze the image data, auto-encoder methods were used to convert image data into a 4232-dimension array data.

After matching with gene expression data, there were total 445 images (samples) to do the analyses, and 336 images (samples) to identify the relationships between participant's age and image signal after matching with age.

2.1.3 Gene data

Gene data were obtained from ADNI. 811 blood samples from the ADNI WGS cohort were collected and performed by Bristol-Myers Squibb (BMS). Microarray (the Affymetrix Human Genome U219 Array, Affymetrix, Santa Clara, CA) was used for expression profiling. The raw gene expression data have then been normalized by RMA (Robust Multi-chip Average) methods[5].

After matching with MRI image data, there were 445 samples that could be used. For each subject, 49386 genes were measured. We selected the top 1000 genes with the greatest variances as our outcome variables. Both the gene expression data and the data after removing gender-effect were used to build the elastic net.

2.2 Statistical analysis

2.2.1 Elastic Net Regression

Considering a linear regression model:

$$\hat{\boldsymbol{y}} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_n}\, x_n$$

For a usual linear regression model, the loss function is residual sum squares (RSS) between the predicted value and true value. However, models fitted by RSS can be sensitive and unstable when coefficients are large, or the number of samples is less than the number of predictors.

To enhance the stability of the regression model, there is penalized linear regression called penalized linear regression. Penalized linear regression models include magnitudes of coefficients as the penalty in the loss function. Two popular penalties are L_1 and L_2 penalties. The L_1 penalty is based on the absolute value of coefficients. Models using L_1 penalty called the Lasso model, which can achieve variate selection simultaneously by minimizing the size of all coefficients to zero[6].

$$L_1\ penalty = \sum_{0}^{n} |\beta_i|\,, \qquad Lasso\ Loss = RSS + L_1\ penalty$$

Another popular penalty is the L_2 penalty, which is based on the squares of coefficients. Models using L_2 penalty called Ridge model, which minimizes the magnitudes of all coefficients, but keep all coefficients in the model[7].

$$L_2\ penalty = \sum_{0}^{n} |\beta_i|^2\,, \qquad Ridge\ Loss = RSS + L_2\ penalty$$

However, both Lasso and Ridge models have limitations in the scenarios when the number of predictors (p) is greater than the sample sizes (n), and when the predictors are correlated. Lasso allows at most n coefficients and selects only one predictor when the predictors are correlated. And Ridge has a poor performance than Lasso in the scenario when $n > p$ and there exist high correlations among predictors[8].

Therefore, we applied elastic net regression, which included both are L_1 and L_2 penalties. The elastic net penalty is a combination of L_1 and L_2 penalties with a hyperparameter α to assign the weight. α is the weight of L_1 ranging from 0 to 1, when $\alpha = 0$, the elastic net regression becomes a lasso regression, and when $\alpha = 1$, the model becomes a Ridges regression. There is another hyperparameter called λ, which assigns the weight of elastic net penalty in the loss function.

$$Elastic\ net\ penalty = \alpha \times L_1\ penalty + (1-\alpha) \times L_2\ penalty,$$

$$Elastic\ net\ Loss = RSS + \lambda \times elastic\ net\ penalty$$

The elastic net regression balanced both Lasso and Ridge, which would have a better performance identifying the relationship between the MRI data of Alzheimer's disease and gene expression or demographic data.

We built one elastic net regression model for each gene, with the gene expression as the dependent variable and the 4232-dimension MRI image data as the independent variable.

2.2.2 gender effect

Gender differences have been reported in brain function, and Xi Zhang's research indicated that gender differences are encoded in brain structure which can be revealed by MRI[9]. This finding is consistent with our results that most of top genes with great performance in elastic net regression (with largest R-squares) are gender-related. Since gender became a confounder in our regression model, we then removed the gender effect on our covariates of interest.

As shown below, for each outcome of interest (expression data of genes or age), we build a linear regression model with gender as the covariate, the residuals of the model then became the outcome of interests.

$$Y_i = \beta_0 + \beta_1 Gender + \varepsilon_i$$

3. Results

3.1 Correlation between gene expression and MRI image data

Our interest was to examine the ability of elastic net regression in identifying genes that have an impact on brain MRI. The top 1000 genes with the greatest variances were selected to test the relationships with MRI data using the elastic net model, we built one model for each gene. R-

squares were used to select good models, top 25 models with the highest R-squares were shown below:

Table 3.1.1 Top 25 elastic net models in predicting gene expression

Index	Gene Symbol	α	λ	Number of coefficients	R-square	RMSE	MAE
1	ALOX15	0.55	0.01	388	0.45	1.04	0.84
2	ALOX15	0.1	0.05	611	0.43	1.01	0.83
3	LOC100653057 || CES1	0.1	0.02	662	0.42	0.91	0.71
4	TXLNG2P	1	0.02	299	0.42	1.31	1.00
5	CLCN4	0.1	0.04	617	0.41	0.82	0.65
6	KIAA1324	0.1	0.01	683	0.41	1.06	0.83
7	KIAA1324	0.1	0.09	522	0.40	0.80	0.62
8	SERPING1	0.1	0.01	682	0.40	1.01	0.79
9	TMEM176B	0.1	0.03	707	0.40	1.78	1.55
10	KLRC3	0.1	0.04	659	0.40	1.03	0.84
11	IFIT3	0.1	0.10	515	0.39	0.79	0.61
12	ADAMTS5	0.1	0.01	673	0.39	0.86	0.70
13	VNN1	0.55	0.01	376	0.39	0.90	0.63
14	ETV7	0.1	0.05	659	0.39	1.14	0.89
15	PZP	0.1	0.03	664	0.39	0.79	0.63
16	S100B	0.1	0.04	624	0.39	1.14	0.91
17	FAM3C	0.1	0.09	542	0.39	0.89	0.72
18	PIGR	0.1	0.02	667	0.38	1.52	1.04
19	APOBEC3A_B || APOBEC3B	1	0.03	319	0.38	2.63	1.97
20	ORM1	0.1	0.01	702	0.38	1.11	0.89
21	HIP1	0.55	0.01	396	0.38	0.78	0.64
22	PDK4	0.1	0.01	739	0.38	0.81	0.62
23	FAM3C	0.1	0.03	670	0.38	0.98	0.78
24	APOBEC3B	1	0.02	315	0.38	1.74	1.34
25	COMMD3-BMI1 || BMI1	0.1	0.01	692	0.38	0.78	0.63

Related diseases or biological functions of selected genes were summarized in table 3.1.2.

Table 3.1.2 Related diseases for the genes

Index	Gene Symbol	Related Diseases or biological functions *	Tissues with high-level of gene expression **
1	ALOX15	Asthma, Nasal Polyps, And Aspirin Intolerance and Periventricular Leukomalacia	Adipose
2	ALOX15	Asthma, Nasal Polyps, And Aspirin Intolerance and Periventricular Leukomalacia	Adipose
3	CES1	Drug Metabolism, Altered, Ces1-Related and Egasyn.	Liver
4	TXLNG	Enable syntaxin binding activity, in chromosome Y	Adipose, Brain, Ovary
5	**CLCN4**	Raynaud-Claes Syndrome and Non-Syndromic X-Linked Intellectual Disability, in chromosome X	Brain, Muscle
6	KIAA1324	Baastrup's Syndrome and Endometrial Serous Adenocarcinoma	Minor Salivary Gland
7	KIAA1324	Baastrup's Syndrome and Endometrial Serous Adenocarcinoma	Minor Salivary Gland
8	SERPING1	Angioedema	Adipose, Artery, Breast, Liver, Lung
9	TMEM176B	Spinocerebellar Ataxia	Liver
10	KLRC3	Nasal Type Extranodal Nk/T-Cell Lymphoma and Monkeypox	Brain, Lung, Spleen, Whole Blood
11	IFIT3	Lupus Erythematosus and Systemic Lupus Erythematosus	Brain, EBV-transformed lymphocytes, Nerve
12	ADAMTS5	Osteoarthritis and Bone Deterioration Disease	Adipose, Breast, Ovary, Uterus
13	VNN1	Obstructive Nephropathy and Methylmalonic Aciduria, Cblb Type	Whole Blood, Liver
14	ETV7	Ewing Sarcoma and Fetishism	EBV-transformed lymphocytes
15	**PZP**	Cumulated in serum in AD patients[10]	Liver, Brain
16	**S100B**	Malignant Peripheral Nerve Sheath Tumor and Neurofibroma	Brain
17	FAM3C	Pancreatic Cancer and Breast Cancer	All Tissues
18	PIGR	Iga Glomerulonephritis and Protein-Energy Malnutrition	Colon, Minor Salivary Gland, Small Intestine
19	APOBEC3A_B || APOBEC3B	Immunity	EBV-transformed lymphocytes, Cultured fibroblasts
20	ORM1	Appendicitis and Dry Eye Syndrome	Liver
21	**HIP1**	Chronic Myelomonocytic Leukemia and Huntington Disease	Adipose, Artery, Brain, Lung
22	PDK4	Type 2 Diabetes Mellitus and Rhabdomyosarcoma	Breast, Muscle - Skeletal
23	**FAM3C**	Pancreatic Cancer and Breast Cancer	Brain, Thyroid
24	APOBEC3B	Hepatitis B and Bone Leiomyosarcoma	EBV-transformed lymphocytes, Cultured fibroblasts
25	COMMD3-BMI1 || BMI1	Mantle Cell Lymphoma and Erythroplakia	All Tissues

* Related Diseases or biological functions for each gene was gotten from GeneCards websit

** Distribution of genes were based on the expression level reported by GTEx websit, high-level expression was defined as the medium TPM greater than 10.

From table 3.1.2, there were only 6 genes either had high-level expression in brain (genes with bolden gene symbol) or had brain-related functions (genes with bolden index). Since we were using the MRI data to do the gene expression prediction, the results were not satisfying. Most of the genes we selected were inflammatory- / immuno- related or gender-related.

Interestingly, we found that there was cluster distribution for some genes caused by gender (Figure 3.1.1), which consisted of the fact that most of the genes we found were at chromosome X or Y, or distributed in tissues like virginal, Uterus, or Testis.

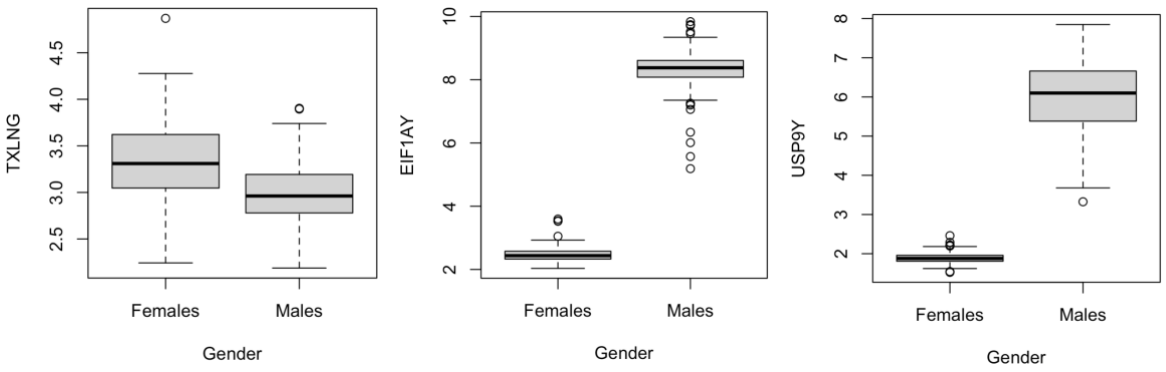

Figure 3.1.1 Distribution of gene expression data by gender

Therefore, for the next step, we built models using gene expression data considering gender as a confounder.

3.2 Correlation between gene expression and MRI image data considering gender

From section 3.1, we found that the distributions of gene expression data were extremely different among gender, especially for gender-related genes in chromosome Y which typically with greatest variances among the samples. Therefore, we then removed gender effects for gene expression data, and then re-selected the top 1000 genes with the greatest variances to build the elastic net model. Models with better performances (high R-square) revealed a good ability to detect the correlation between gene expression and MRI image data. The top 25

models with the highest R-squares were shown in table 3.2.1. These model performances were better than models in section 3.1, which proved that gender had affected gene expression prediction ability.

Table 3.2.1 Top 25 elastic net models in predicting gene expression

Index	Gene Symbol	α	λ	Number of coefficients	R-square	RMSE	MAE
1	MPO	0.1	0.01	620	0.50	0.84	0.63
2	IGHV4-31	0.1	0.12	531	0.47	1.14	0.89
3	BCL11A	0.1	0.01	670	0.45	0.75	0.59
4	C19ORF33	0.55	0.01	394	0.44	0.99	0.63
5	CEACAM6	0.1	0.05	646	0.43	1.07	0.87
6	GSTM1	0.1	0.05	607	0.42	1.05	0.92
7	FMOD	0.1	0.02	610	0.42	0.70	0.31
8	FCRL5	0.1	0.01	654	0.42	1.08	0.81
9	ITGAV	1	0.01	331	0.42	0.94	0.69
10	CRISP3	0.55	0.02	384	0.42	1.18	0.96
11	TUBB2A	0.1	0.09	504	0.41	0.94	0.63
12	IFI44L	0.1	0.02	681	0.41	1.25	0.96
13	IGKV3-11	0.1	0.04	613	0.41	0.88	0.66
14	CHURC1	0.1	0.01	682	0.41	0.79	0.70
15	REEP3	0.1	0.03	679	0.40	0.80	0.61
16	EFNA5	0.1	0.01	720	0.40	0.74	0.56
17	SH3BGRL2	1	0.01	307	0.40	0.86	0.64
18	PZP	0.1	0.04	637	0.40	0.81	0.64
19	OAS3	0.55	0.01	393	0.40	1.64	0.74
20	PDK4	0.1	0.01	721	0.40	0.80	0.62
21	HERC5	0.1	0.03	646	0.40	0.89	0.69
22	HPGD	0.1	0.01	684	0.40	0.76	0.60
23	ABLIM3	0.1	0.12	466	0.40	1.89	0.73
24	MFAP3L	0.1	0.10	531	0.39	0.82	0.65
25	GBP1	0.55	0.01	391	0.39	0.81	0.61

Next, we summarized the functions and related diseases of the selected genes (table 3.2.2) to examine the correlation detection ability of the elastic net.

Table 3.2.2 Related diseases for the biomarker genes

Index	Gene Symbol	Related Diseases or biological functions *	Tissues with high-level of gene expression **
1	MPO	Myeloperoxidase Deficiency and Alzheimer Disease	Whole Blood
2	IGHV4-31	Activation of immune response	EBV-transformed lymphocytes
3	**BCL11A**	Intellectual Developmental Disorder With Persistence Of Fetal Hemoglobin and Intellectual Disability - Hypoplastic Corpus Callosum	Brain, Blood
4	C19ORF33	Pre-eclampsia, which could trigger Alzheimer's disease	Muscularis, Vagina, skin
5	CEACAM6	Crohn's Disease and Cystic Fibrosis	Mucosa, Lung
6	GSTM1	Asbestosis and Oral Leukoplakia	Bladder, Vagina
7	FMOD	Pseudo achondroplasia and Myopia	Artery
8	FCRL5	Hairy Cell Leukemia and Lymphoma	EBV-transformed lymphocytes, Spleen
9	ITGAV	West Nile Virus and Herpes Simplex	Artery, Heart
10	CRISP3	Ectopic Pregnancy and Prostate Cancer.	Minor Salivary Gland
11	**TUBB2A**	Cortical Dysplasia, Complex, With Other Brain Malformations and Tubulin, Beta	Brain
12	IFI44L	Multisystem Inflammatory Syndrome In Children and Orofaciodigital Syndrome V	EBV-transformed lymphocytes
13	IGKV3-11	Immune response	EBV-transformed lymphocytes
14	**CHURC1**	Hypoglycemia, Leucine-Induced and Polycystic Kidney Disease 1 With or Without Polycystic Liver Disease	Brain, Artery
15	**REEP3**	Played roles in GPCR signals, which been studied as therapeutic target for AD[11]	Brain, Artery
16	EFNA5	Cortical Senile Cataract and Septal Myocardial Infarction	Skin, Minor Salivary Gland, Brain
17	**SH3BGRL2**	Leber Congenital Amaurosis	Adipose, Brain, Esophagus
18	**PZP**	Cumulated in serum in AD patients[10]	Liver, Brain
19	OAS3	Chikungunya and Tick-Borne Encephalitis	EBV-transformed lymphocytes
20	PDK4	Type 2 Diabetes Mellitus and Rhabdomyosarcoma	Breast, Muscle - Skeletal
21	HERC5	Influenza	EBV-transformed lymphocytes, Testis
22	HPGD	Digital Clubbing, Isolated Congenital and Hypertrophic Osteoarthropathy, Primary, Autosomal Recessive	Bladder, Vagina
23	**ABLIM3**	Aromatic L-Amino Acid Decarboxylase Deficiency and Fetal Erythroblastosis	Adipose, Brain, Breast
24	**MFAP3L**	Played roles in EGFR and MAPK1/ERK2 signals, which been identified as target for treating Amyloid-β–induced memory loss[12]	Brain, Testis
25	GBP1	Chronic Active Epstein-Barr Virus Infection and Aneurysmal Bone Cysts	Whole body

* Related Diseases or biological functions for each gene was gotten from GeneCards websit

** Distribution of genes were based on the expression level reported by GTEx websit, high-level expression was defined as the medium TPM greater than 10.

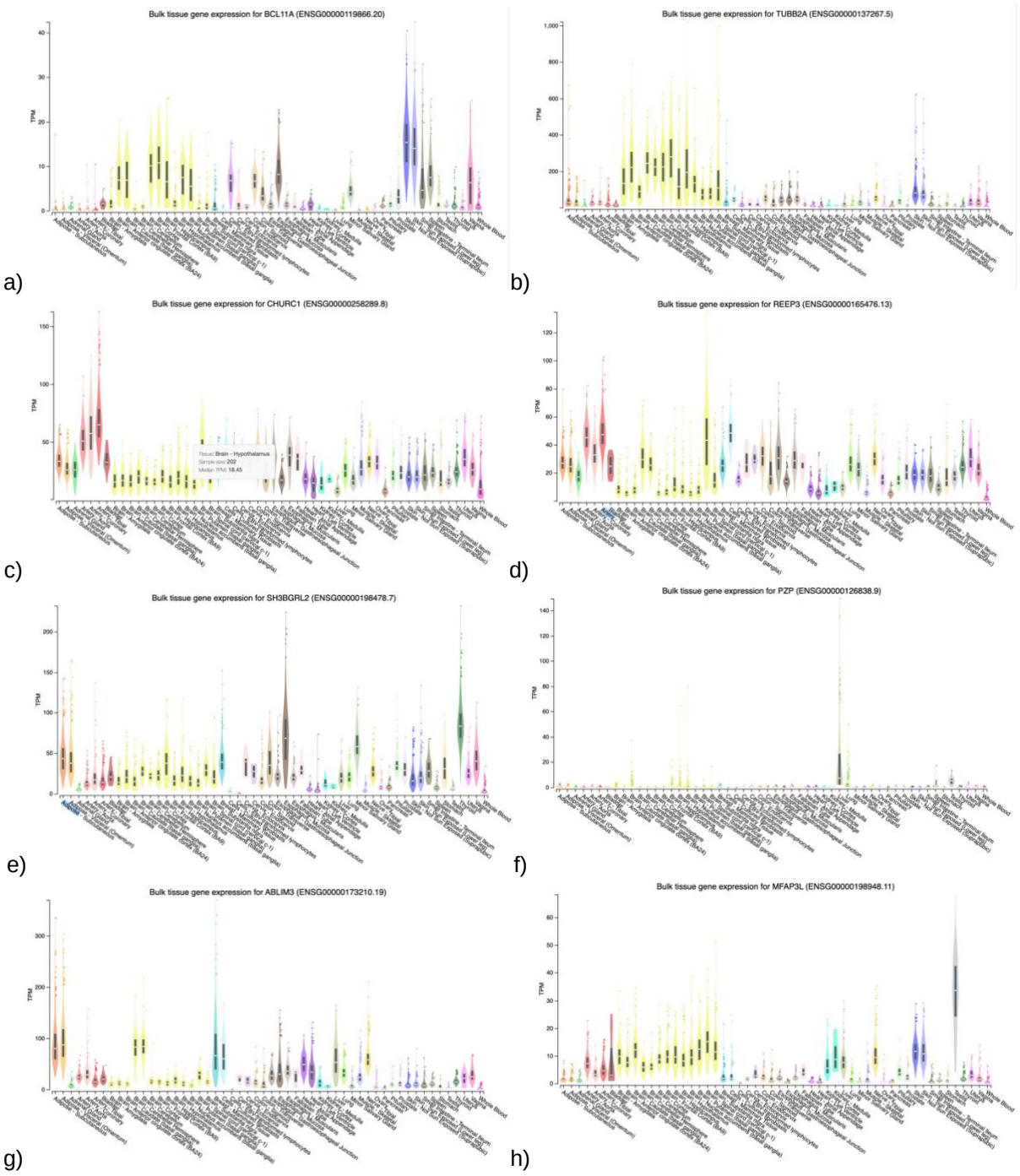

Figure 3.2.1 Distributions of the brain-related genes. a) BCL11A, b) TUBB2A, c) CHURC1, d)REEP3, e)SH3BGRL2, f) PZP, g) ABLIM3, h) MFAP3L

Table 3.2.2 showed that among the top 25 models with the highest R-squares, 10 of them identified genes (genes with bolden index) had reported playing a role in AD. And 8 genes (with bolden gene symbols in table 3.2.2) had high-level expression in brain based on figure 3.2.1.

Together, there were about half (12 of 25) genes we identified were either mainly distributed in brain or had AD-related functions. These correlation detection results were way better than that using raw gene expression data, suggesting that taking gender effects into consideration could greatly improve the ability to identify the biomarkers of AD for our model.

It's interesting that among other non-brain-related genes, almost half of them were either inflammatory- / immuno- related or gender-related. Immune-related genes were more likely to be detected as feature genes because expression varied widely among individuals[13]. Also, the gender-related genes we detected were mainly distributed in the vagina or testis, indicating that there were still gender effects in our models.

3.3 Correlation between age and MRI image data

Many machine learning approaches are applied in building brain age prediction models. Convolutional methods of artificial neural network, support vector machine method, and deep learning methods such as convolutional neural network which doesn't rely on manual designs in feature extraction had been studied in this area [14].

We tested the age prediction efficiencies in MRI image data using the elastic net regression model. There were total of 336 samples after matching the age and MRI image data, and when $\alpha = 0.55$ and $\lambda = 1.34$, the model could achieve the detection of the correlation between age and MRI image data with the result of $R\ square = 0.45$, $RMSE = 5.23$ and $MAE = 3.30$. In this model, 233 dimensions of 4232 dimensions of the image data were selected.

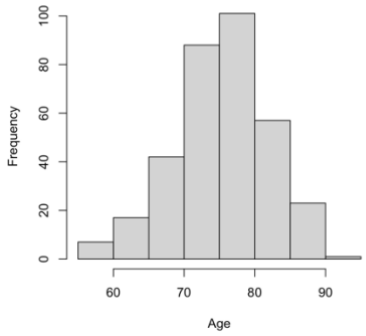

Figure 3.3.1 Distribution of age in ADNI data.

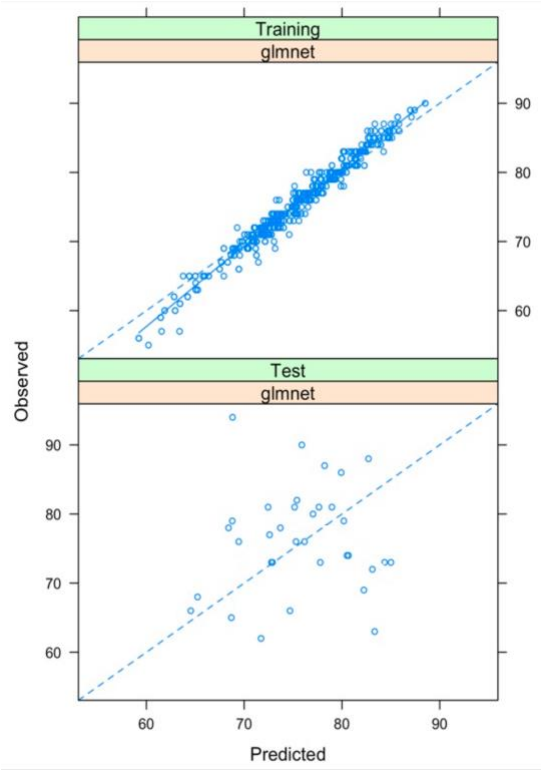

Figure 3.3.2 Results of elastic net on correlation between age and MRI image data.

Based on the gender differences discovered in brain structure and our previous findings of gender effects in gene expression prediction, we then take gender into consideration in building models. The distribution of age among gender was shown below:

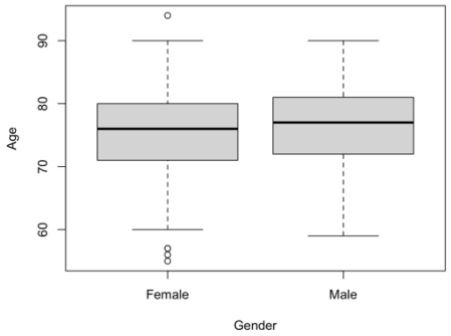

Figure 3.3.3 Distribution of age in ADNI data among gender.

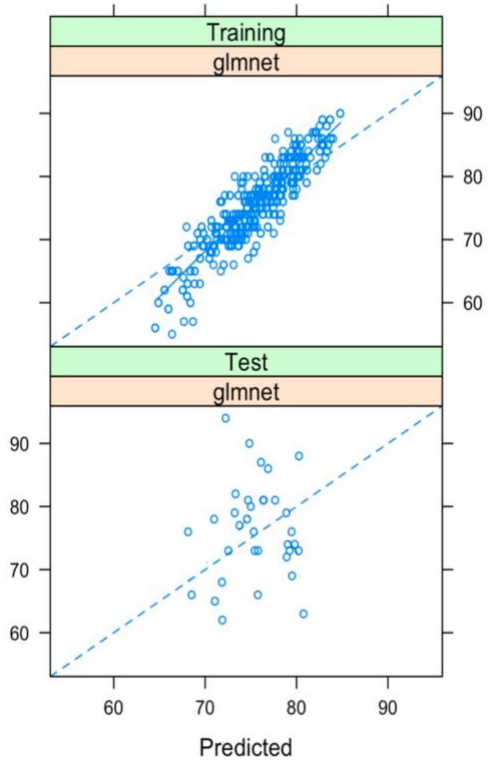

Figure 3.3.4 Results of elastic net on predicting age considering gender.

Age was evenly distributed among gender, and the elastic net model achieve the detection of the correlation between age and MRI image data with the result of $R\ square\ =\ 0.41$, $RMSE\ =\ 5.04$ and $MAE\ =\ 3.78$ when $\alpha = 0.55$ and $\lambda = 0.42$. In this model, 169 dimensions of 4232 dimensions of the image data were selected. The results were similar to the model that didn't consider gender, suggesting that gender was not a confounder in predicting age by MRI data. This result consisted of the age distribution by gender.

4. Discussion

In our study, we tried to use elastic net regression to identify the correlation between gene expression and MRI image data. From our results, we can see that brain-related genes had better performances in elastic net (with larger R-square values), which indicated that our model had the ability to identify AD biomarkers.

Comparing the gene expression prediction models in sections 3.1 and 3.2, models built on gene expression data that removed gender effects had better performances. These results indicated that gender was a confounder when analyzing the MRI data of AD patients. However, based on the results in section 3.2, we still identified genes, which were gender-related or immune-related but not brain-related, even after removing gender effect. To get more accurate results, more methods in gene selection were needed, especially to remove the gender-related or immune-related genes. Also, the clustered distributions discovered in some (15 out of 1000) of the genes reflecting that there were other confounders in our models. In the further study, the stage of AD might also be considered in predicting gene expression progress.

For most of the models the best αs were 0.11, this suggested the Lasso model had more weight in the model. Less number of coefficients were preferred. This was understandable since most of the dimensions of the MRI data after being converted by auto-encoder methods were similar, Lasso regression could remove useless dimensions in model selection.

In section 3.3, when predicting patients' age by MRI data, it could be found that the number of coefficients was reduced from 233 to 169. This phenomenon can be explained after removing gender effects, brain structure differences caused by gender would not be considered in the models, therefore, the number of coefficients would be reduced. Additionally, even though the R-squares in models with or without removing gender effects were similar, the number of coefficients in the model removed gender effects, which would make the calculation easier.

There were limitations in our study. The R-squares for models were relatively low even after removing gender effects, most of the R-squares were 0.4 - 0.5, which might suggest a poor prediction ability. However, this can be explained that the gene expression data were from blood samples while the MRI data captured brain features, the relationships between those two datasets were not expected to be strong. Since our models could select brain-related genes as biomarkers, we revealed the potential of the elastic net in dealing with MRI data to identify biomarkers.

References

[1]. Noor, M., Zenia, N. Z., Kaiser, M. S., Mamun, S. A., & Mahmud, M. (2020). Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia. Brain informatics, 7(1), 11. https://doi.org/10.1186/s40708-020-00112-2

[2]. Zhang, D., Li, L., Sripada, C., & Kang, J. (2020). Image-on-Scalar Regression via Deep Neural Networks. *arXiv preprint arXiv:2006.09911*.

[3]. Yue, Y., Loh, J. M., and Lindquist, M. A. (2010). Adaptive spatial smoothing of fmri images. Statistics and its Interface 3, 3–13.

[4]. Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. Journal of the American Statistical Association 112, 1131–1146.

[5]. Vawter, M. P., Evans, S., Choudary, P., Tomita, H., Meador-Woodruff, J., Molnar, M., Li, J., Lopez, J. F., Myers, R., Cox, D., Watson, S. J., Akil, H., Jones, E. G., & Bunney, W. E. (2004). Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, *29*(2), 373–384. https://doi.org/10.1038/sj.npp.1300337

[6]. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[7]. Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, *4*(2), 105-123.

[8]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320.

[9]. Zhang, X., Liang, M., Qin, W., Wan, B., Yu, C., & Ming, D. (2020). Gender Differences Are Encoded Differently in the Structure and Function of the Human Brain Revealed by Multimodal MRI. *Frontiers in human neuroscience*, *14*, 244. https://doi.org/10.3389/fnhum.2020.00244

[10]. Nijholt, D. A., Ijsselstijn, L., van der Weiden, M. M., Zheng, P. P., Sillevis Smitt, P. A., Koudstaal, P. J., Luider, T. M., & Kros, J. M. (2015). Pregnancy Zone Protein is Increased in the Alzheimer's Disease Brain and Associates with Senile Plaques. *Journal of Alzheimer's disease : JAD*, *46*(1), 227–238. https://doi.org/10.3233/JAD-131628

[11]. Zhao, J., Deng, Y., Jiang, Z., & Qing, H. (2016). G Protein-Coupled Receptors (GPCRs) in Alzheimer's Disease: A Focus on BACE1 Related GPCRs. *Frontiers in aging neuroscience*, *8*, 58. https://doi.org/10.3389/fnagi.2016.00058

[12]. Wang, L., Chiang, H. C., Wu, W., Liang, B., Xie, Z., Yao, X., Ma, W., Du, S., & Zhong, Y. (2012). Epidermal growth factor receptor is a preferred target for treating amyloid-β-induced memory loss. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(41), 16743–16748. https://doi.org/10.1073/pnas.1208011109

[13]. Li, Y., Ge, X., Peng, F. *et al.* Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biol* **23,** 79 (2022). https://doi.org/10.1186/s13059-022-02648-4

[14]. Jiang, H., Lu, N., Chen, K., Yao, L., Li, K., Zhang, J., & Guo, X. (2020). Predicting Brain Age of Healthy Adults Based on Structural MRI Parcellation Using Convolutional Neural Networks. *Frontiers in neurology*, *10*, 1346. https://doi.org/10.3389/fneur.2019.01346