



### 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원 저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리와 책임은 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)



보건학 석사학위논문

Polygenic risk  
prediction models  
for Alzheimer's disease

다유전적 위험 구조을 고려한  
알츠하이머 질병 예측

2018년 02월

서울대학교 보건대학원  
보건학과 보건학전공  
서 수 진

Polygenic risk  
prediction models  
for Alzheimer's disease

다유전적 위험 구조을 고려한  
알츠하이머 질병 예측

지도교수 원 성 호

이 논문을 보건학 석사학위논문으로 제출함

2017년 12월

서울대학교 보건대학원

보건학과 보건학전공

서 수 진

서수진의 석사학위논문을 인준함

2017년 12월

위 원 장 김 호 (인)

부 위 원 장 성 주 헌 (인)

위 원 원 성 호 (인)

# ABSTRACT

## Polygenic risk prediction models for Alzheimer's disease

Sujin Seo

Department of Public Health  
The Graduate School of Public Health  
Seoul National University

**Background:** Alzheimer's disease (AD) is known to have polygenic architecture, which indicates a large proportion of the susceptible single-nucleotide polymorphisms (SNPs) collectively account for a significant portion of variation of AD. Furthermore, since the effect of APOE e4 on the risk of AD is substantial, the impact of genetic factors other than the APOE gene may be masked when APOE e4 carriers and non-carriers are analyzed together. Patients with and without APOE e4 have different genetic bases and pathogenic distinctions (Jiang, et al., 2016). Therefore, stratification based on APOE e4 status can allow for the exploration of the underlying mechanism between APOE e4 carriers and non-carriers.

**Objective:** In this study, we assess the performance of penalized methods and non-penalized methods in the prediction of AD to consider the polygenic architecture of AD in the model. In addition, we compare the models stratified by APOE e4 status to the models with combined data.

**Method:** In this paper, penalized regression methods are used alternative to PRS. Unlike PRS, where a large number of underlying susceptibility genes are combined into one variable in the prediction model, the penalized regression methods consider those genes as separate variables. Owing to the penalty term to the coefficients, the problem of much larger number of genetic variants than the sample size. Some penalized methods, such as lasso (Tibshirani, 1996) and elastic–net (Zou and Hastie, 2005), conduct automatic variable selection and give more interpretable results. Furthermore, group lasso (Meier, et al., 2008) is the extension of lasso penalty and it selects variables at the predefined group level. In this paper, we grouped SNPs of APOE of non–carrier group and carrier group and apply group lasso regression. In addition, we explored the mechanisms of the two groups individually by stratify according to the presence of APOE e4 alleles. We applied the various models to National Research Center for Dementia (NRCD) data consisting of all Koreans. The predictive performance is evaluated by AUC.

**Result:** We assessed the odds ratio resulted from GWAS for the combined data, carrier group data and non–carrier group data. When comparing the odds ratios of 100 SNPs of which the p–value is the lowest, for the combined data, the most of large effects SNPs are on chromosome 19 where APOE gene locates and the others sporadically distributed having modest odds ratio, approximately 1.5. Looking at the odds ratios in separate groups, for APOE carrier group, some have high odds ratio exceeding 3 and the others have low value, less than 1. On the other hands, APOE non–carrier group doesn't seem to have high effect SNPs as carrier group, instead, most of SNPs have between 1 and 2 odds ratio.

The best accuracy was found in penalized methods for both the combined case and the separate cases. For the combined case, the largest AUC was 0.6520 with only 100 SNPs and 0.6671 with 10,000 SNPs, in PRS model and in

ridge regression model, respectively. For the separate case, AUC 0.6551 with 100 SNPs and 0.6741 with 1,000 SNPs, in PRS model and in ridge regression model, respectively. When we crossly combined the models where  $y \sim \text{APOE4} + \text{APOE2} + \text{AGE} + \text{SEX}$  model for the carrier group and ridge regression model with 10,000 SNPs for non-carrier group, AUC was 0.6773.

We further investigated whether the stratified strategy helps to improve AD prediction. For each model when the number of SNPs is fixed, stratified model outperformed the combined model when relatively smaller number of SNPs are included but it was opposite when the number of SNPs is large. When we crossly calculated AUC and got the best accuracy 0.6773.

**Conclusion:** This study supports that AD has different polygenic architectures according to APOE types. First, the results of GWAS for combined data and separated data have shown that different kinds of SNPs affect AD with different effects. Second, we show that stratified analysis improves AUC over combined one. For extension of our analysis, we may identify people of high risk of AD without any APOE alleles. That is, the suggested method can provide more variation in estimated risk in the population.

**Keywords:** Alzheimer's disease (AD), Group lasso regression, Lasso regression, Penalized regression, Prediction, Polygenic architecture, Polygenic risk score (PRS), Ridge regression

**Student Number:** 2016-24008

# Contents

I. INTRODUCTION .....	1
II. METHODS .....	5
1. Data description .....	5
2. Model selection .....	6
3. Variable Selection .....	8
4. Non-penalized regression .....	10
Polygenic risk score .....	10
5. Penalized regression .....	10
Ridge regression .....	11
Lasso regression .....	12
Elastic-net regression .....	12
Group lasso regression .....	13
III. RESULTS .....	14
IV. DISCUSSION .....	23
REFERENCE .....	25

## List of Tables

Table 1. Cross table of APOE type and AD .....	6
Table 2. Model .....	7
Table 3. Catalog SNP and corresponding proxies ..	9
Table 4. Predictive accuracy for combined data ..	22
Table 5. Predictive accuracy for separated data ..	22
Table 6. Predictive accuracy of group lasso .....	23

## List of Figures

Figure 1. The workflow of the analysis .....	4
Figure 2. Diagram of design matrix .....	14
Figure 3. Manhattan plot and Q–Q plot .....	18
Figure 4. Odds ratio of the largest 100 SNP .....	19
Figure 5. Negative log(p) of selected top SNPs ..	20

## I. INTRODUCTION

Research has shown that clinical symptoms appear for decades before the onset of Alzheimer's disease (AD) (Bird, 2015). Therefore, for effective prevention, the risk of AD should be predictable 10 years before the onset of symptoms. Predicting the risk of developing those at high risk for AD is becoming increasingly important, and it may help develop methods aimed at preventive interventions such as risk reduction, behavior modification or pharmacologic treatment.

Genome-wide association studies (GWAS) have identified a number of susceptible loci contributing to the risk of AD. However, some of them have small effect size and the known variants explain only a small proportion of the estimated heritability. Apolipoprotein E (APOE), most significant risk factor for developing AD explained only 6% of total phenotypic variance (Ridge, et al., 2013). Lee, et al. estimated substantially increased heritability, 24%, when the weak effect loci are included in the model simultaneously (Lee, et al., 2012). This result implies that a large proportion of the susceptible single-nucleotide polymorphisms (SNPs) lie below the genome-wide significant threshold but collectively account for a significant portion of variation. Desikan, et al. and Escott-Price, et al.

also found that polygenic architecture plays an important role in prediction the risk of AD (Desikan, et al., 2017; Escott-Price, et al., 2015).

To attempt explore the polygenic architecture and predict the risk of AD, Escott-Price, et al. adopts polygenic risk score (PRS) in the prediction models, achieving prediction accuracy  $AUC = 78.2\%$ . PRS is calculated as the weighted sum of risk alleles with the weights are simple linear logistic regression coefficients. This approach is computationally efficient as PRS could be built from the result of GWAS. However, since the model is based only on marginal effects of variants, it leads to biased score and less accuracy when there are joint effects between variants (Won, et al., 2015). For considering the joint effect of a small number of variables, logistic regression can be used, but this approach is not feasible for large number of variables.

Penalized regression methods can be used alternatively. Unlike PRS, where a large number of underlying susceptibility genes are combined into one variable in the prediction model, the penalized regression methods consider those genes as separate variables. Thus, we can utilize joint effects of multiple variants. Owing to the penalty term to the coefficients, the problem of much larger number of genetic variants than the sample size, so called

“ $p \gg N$  problem”, is resolved. Some penalized methods, such as lasso (Tibshirani, 1996) and elastic-net (Zou and Hastie, 2005), conduct automatic variable selection and give more interpretable results. Furthermore, group lasso (Meier, et al., 2008) is the extension of lasso penalty and it selects variables at the predefined group level. It is useful for when the covariates have a group structure, and it is desirable to have all coefficients within a group become nonzero or zero simultaneously. In this paper, we grouped SNPs of APOE e4 of non-carrier group and carrier group and apply group lasso regression.

APOE is the strongest susceptible gene of late-onset AD. APOE exists as three polymorphic alleles, e2, e3 and e4. Individuals with copy of e4 have a higher risk developing AD, especially e4/e4 homozygotes have a 14.9 increased odds compared with e3/e3 reference haplotype (Corder, et al., 1994). Since the effect of APOE e4 for the risk fof AD is substantial, the effects of other genetic factors could be masked when APOE e4 carriers and non-carriers are analysis together. When APOE e4 carrier and non-carriers are combined, it is not possible to identify a high risk person who does not have APOE e4. In addition, the basic assumption of a combined analysis that does not into account the interaction of APOE is that the risk factors and there effects are the same for the APOE e4 carriers and

non-carriers. However, patients with and without APOE e4 have many different clinic-pathologic features. Jiang, et al. reveals that the two groups have different genetic bases and pathogenic distinctions (Jiang, et al., 2016). Therefore, stratification based on APOE e4 status can allow for the exploration of the underlying mechanism between APOE e4 carriers and non-carriers.

In this report, we assess the performance of penalized methods and non-penalized methods in the prediction of AD. In addition, we compare the models stratified by APOE e4 status to the models with combined data. We applied the various models to National Research Center for Dementia (NRCD) data consisting of all Koreans. The predictive performance is evaluated by AUC. The whole workflow of the analysis is provided in Figure 1. Our results indicate that the penalized regression methods with stratified analysis captures the different underlying architecture of APOE carriers and non-carriers and give higher prediction accuracy.

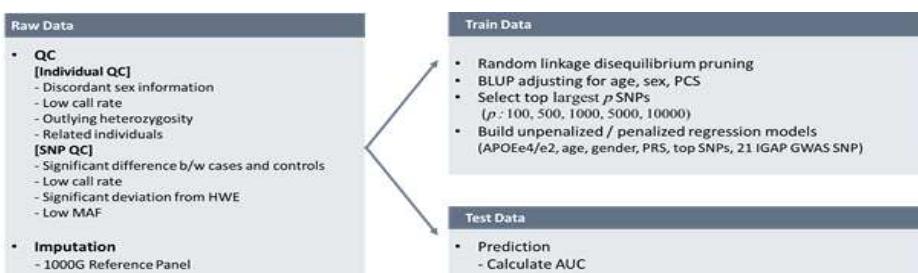


Figure 1. The workflow of the analysis

## II. METHODS

### 1. Data description

Dataset is based on Korean population consisting 4,391 individuals from NRCD. Subjects were gathered from 8 centers, Chosun University Hospital, Chungnam National University Hospital, Donga University Hospital, Kyungpook National Universiy Hospital, NRCD, Seoul National University Hospital, Seoul National University Bundang Hospital and Pusan National University Hospital. They were genotyped for 833,535 probe sets with Axiom\_KORV1\_0 chip. At first, we excluded individuals with low sex inconsistency, low call rate (call rate < 97%), outlying heterozygosity (heterozygosity rate  $>$  mean  $\pm$  3s.d.) and related individuals (identical by state  $>$  0.9). SNPs which are significantly different between cases and controls ( $p$ -value  $<$  10E-05), have low call rate (call rate  $<$  0.03), significantly deviate from Hardy-Weinberg equilibrium ( $p$ -value  $<$  10E-05) and have low minor allele frequency (MAF  $<$  0.05) were eliminated. The quality controlled data were imputed with Shape-IT (Delaneau, et al., 2008) and Impute2 (Howie and Marchini, 2011) using 1000 Genomes data as a reference penal. For targeting our prediction model to late-onset AD, we further filtered out who are mild cognitive impairment (MCI) and under age 60. In total, 2,372 individuals, 1,371 cases and 1,001 controls,

and 344,101 SNPs were analyzed. Among the QC data, the distribution of APOE and case/control is in Table 1.

	Control	Case	Total
e2/e2	4	2	6
e2/e3	142	68	210
e2/e4	17	19	36
e3/e3	996	513	1,509
e3/e4	211	329	540
e4/e4	1	70	71
Total	1,371	1,001	2,372

Table 1. Cross table of APOE type and AD

## 2. Model selection

To find the best prediction models for AD, we designed variety scenarios (Table 2). First, since we want to assess the different effects of markers between APOE e4 carriers and non-carriers, we compare the model by separating and combining data by APOE types. People who have at least one allele are assigned to the carrier group and the others to the non-carrier group. In the combined data, we use all the samples to do GWAS of BLUP and select susceptible SNPs, but in separating data, we do that in each group. Second, to compare the penalized method and non-penalized methods for AD prediction, we build the models both of them.

Model		
Non-penalized Regression	$Y \sim APOE4 + APOE2$ $Y \sim APOE4 + APOE2 + AGE + SEX$ $Y \sim APOE4 + APOE2 + AGE + SEX + PRS$	
Penalized Regression	$Y \sim APOE4 + APOE2 + AGE + SEX + \text{top SNP}$	Ridge Elastic-net Lasso Group lasso

Table 2. Model

The variables used in the model is coded as following. The dependent variable is coded as binary, people with AD is 1 and the normal people is 0. APOE4 and APOE2 are the count of e4 alleles and e2 alleles, respectively and they are regarded as a continuous variable to take account of the exponential effect of APOE e4 alleles. AGE is the age of individuals itself, without any transformation and SEX is coded as 1 to male and 2 to female. The methods selecting top SNPs and calculating the PRS is described in detail in the following sections.

Accuracies of the disease prediction models were assessed via 10-fold cross validated AUC. To prevent overfitting problem, the data is randomly divided into 10 different subgroups, and one of them is used to test data and the others train is used to train data. Train data is used to select SNPs and calculate genetic effects and test data is evaluated the models. Since we will build the penalized

regression models which leads to choosing the tuning parameters, we again do 10-fold cross validation within the train data. We repeat it 10 times changing the test data, so all the subgroups are used for test data.

### 3. Variable Selection

Prior to construct models considering polygenic effects of variants, which variables would be included in the model should be determined. Under an extreme polygenic architecture, for example tens of thousands of common SNP have small effects, including all the SNPs in a model may not be computationally feasible or may induce noise that are not truly associated with the disease. The simplest and most common popular approach is to select SNPs based on the GWAS (Chatterjee, et al., 2016). The best linear unbiased prediction (BLUP) also gains popularity as improve prediction accuracy (Zhang, et al., 2014) by taking into account the similarities between pairs of individuals. In this paper, we select variables based on the result of the BLUP with GCTA (Yang, et al., 2011). The SNPs with the largest absolute values of effects are selected for construction of further model. Since how total heritability distributed over the genome is unknown and the number of susceptible markers may be related to the prediction accuracy, we

assessed a variety of thresholds by the number of SNPs, 100, 500, 1,000, 5,000 and 10,000.

Chr	IGAP			NRCD		
	SNP	Position	Closest gene	SNP	Position	Closest gene
1	rs6656401	207692049	<i>CR1</i>	rs11117949	207682256	CR1
2	rs6733839	127892810	<i>BIN1</i>			
6	rs10948363	47487762	<i>CD2AP</i>	rs10948368	47591856	CD2AP
7	rs11771145	143110762	<i>EPHA1</i>	rs3885667	143103155	EPHA1
8	rs9331896	27467686	<i>CLU</i>	rs9331896	27467686	CLU
11	rs983392	59923508	<i>MS4A6A</i>	rs7232	59940599	MS4A6A
11	rs10792832	85867875	<i>PICALM</i>	rs11234569	85886994	
19	rs4147929	1063443	<i>ABCA7</i>	rs3752242	1053677	ABCA7
19	rs3865444	51727962	<i>CD33</i>	rs3826656	51726613	CD33
6	rs9271192	32578530	<i>HLA-DRB5-HLA-DRB1</i>	rs9271198	32578726	HLA-DRB1
8	rs28834970	27195121	<i>PTK2B</i>	rs13266887	27223342	PTK2B
11	rs11218343	121435587	<i>SORL1</i>	rs1792124	121441520	SORL1
14	rs10498633	92926952	<i>SLC24A4-RIN3</i>			
18	rs8093731	29088958	<i>DSG2</i>			
2	rs35349669	234068476	<i>INPP5D</i>	rs80106733	234080309	INPP5D
5	rs190982	88223420	<i>MEF2C</i>			
7	rs2718058	37841534	<i>NME8</i>			
7	rs1476679	100004446	<i>ZCWPW1</i>	rs1476679	100004446	<i>ZCWPW1</i>
11	rs10838725	47557871	<i>CELF1</i>	rs59409728	47548318	CELF1
14	rs17125944	53400629	<i>FERMT2</i>	rs17125944	53400629	<i>FERMT2</i>
20	rs7274581	55018260	<i>CASS4</i>	rs6024879	55017560	<i>CASS4</i>

**Table 3. Catalog SNPs and corresponding proxies in NRCD data**

Inclusion of known genome-wide significant SNPs had improved the model (Escott-Price, et al., 2015). Thus, we included 21 genome-wide significant SNPs (Lambert, et al., 2013) listed in the selected SNPs in BLUP. We refer them as catalog SNPs. The exactly same SNPs are not necessarily exist in our data, thus we choose the best substitutes.

Proxies with 16 SNPs in the NRCD data were selected and they are closest to and highest linkage disequilibrium with the catalog SNPs within the LD blocks. Table 3 shows the catalog SNPs and their proxies in our data.

## 4. Non-penalized regression

### Polygenic risk score

PRS is the marginal effects of susceptibility SNPs and calculated as the weighted sum of risk alleles with the weights are coefficients of simple logistic regression. To construct PRS, two procedure are required, variable selection and coefficients of selected SNPs. In this paper, we use the variables previously selected by BLUP and the estimated OR from logistic regression with the selected SNPs were used to calculate the total genetic effects of each individuals. PRS is practically useful for disease prediction as it is computationally efficient. However, it has limit under the situation where the joint effects are substantial or variants are highly correlated.

## 5. Penalized regression

Penalized regression methods allow to resolve  $p \gg N$  problem by posing penalties to the coefficients. Depending on the formation of penalties, we can explore different penalized

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n \left\{ -y_i x_i^t \beta + \log(1 + \exp(x_i^t \beta)) \right\} + \sum_{j=1}^p F_\lambda(|\beta_j|)$$

regression methods. General form of solution for the dimensional coefficient vector  $\beta = (\beta_1, \dots, \beta_p)^t$  is estimated by minimizing the penalized negative log-likelihood, i.e., where  $F_\lambda$  is a penalty function and  $\lambda$  is the tuning parameter for the amount of shrinkage which chosen to have the best performance based on the grid search. When  $\lambda$  is too small, we tend to overfit the data and have high variance. On the other hands, when  $\lambda$  is too large, we may underfit the data and the models will be potentially biased.

### Ridge regression

Ridge regression (Hoerl and Kennard, 1970) uses sum of squared coefficient for the penalty function :

$$F_\lambda(t) = \lambda t^2$$

This penalty term helps to give unique solution even in the case where  $p$  is much larger than  $n$ . Also the ridge estimates tend to have smaller variance than the least square estimates, which can alleviate the problem of multicollinearity of least square method. However, ridge does not make the model more interpretable because it does not perform variable selections.

## Lasso regression

In contrast to ridge regression, lasso regression (Tibshirani, 1996) selects variables and estimates the coefficients simultaneously. This property is due to the  $l_1$  penalty function which is the sum of the absolute coefficients:

$$F_\lambda(t) = \lambda t$$

One common method for variable selection is forward or backward selection. This approach is unstable and highly variable, in the sense that an infinitesimally small change in data can result in completely different estimates, especially in high dimension. Lasso is superior to have stable performance and higher prediction accuracy(Fan and Li, 2001). Although lasso has many advantages, it has several limitations. The at most number of selected variables is  $n$  when  $p \gg n$ . In addition, if there are correlated variables, then lasso tends to select only one variable regardless of any importance.

## Elastic–net regression

Elastic net (Zou and Hastie, 2005) complements the ridge and lasso by penalized with both  $l_1$  and  $l_2$  norm. This has the effect of effectively shrinking the coefficients and setting some coefficients to zero. Elastic net can handle highly correlated variables better than lasso. The penalty

function for elastic net is following.

$$F_\lambda(t) = \lambda \{ \alpha t + (1 - \alpha)t^2 \}$$

where  $\alpha$  is the additional tuning parameter to balance the ridge and lasso. When  $\alpha=1$ , the above penalty function becomes lasso and when  $\alpha=0$ , it becomes ridge penalty.

### Group lasso regression

The group lasso is an extension of the lasso to do variable selection on groups of variables. It means, predefined groups of predictors to be included or excluded to the model together. The predictors are grouped to  $G$  groups and we can rewrite  $x_i = (x_{i,1}^T, \dots, x_{i,G}^T)^T$  with the group of variables  $x_{i,g} \in R^{df_g}, g = 1, \dots, G$ . We denote  $df_g$  as the degrees of freedom of the  $g$ th group. The estimator for group

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n \{-y_i \eta_\beta(x_i) + \log(1 + \exp(\eta_\beta(x_i)))\} + \lambda \sum_{g=1}^G S(df_g) \|\beta_g\|_2$$

lasso is given by

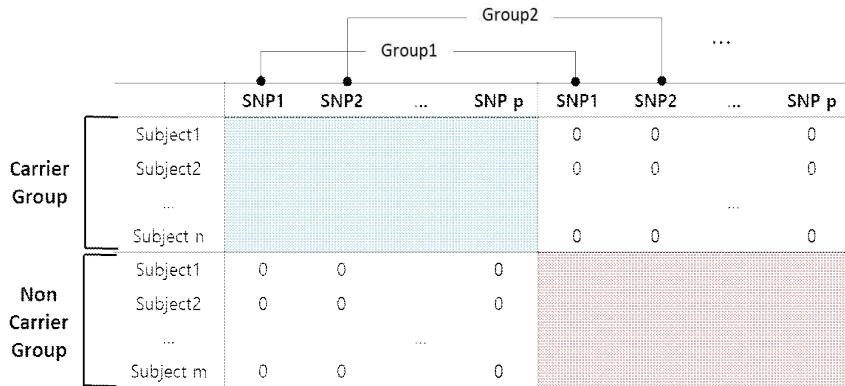
where  $\eta_\beta(x_i) = \beta_0 + \sum_{g=1}^G x_{i,g}^T \beta_g$  and  $S(df_g)$  rescales the penalty with respect to the number of predictors of the group and typically is used.

In this study, we modify a design matrix as

$$X_g = \begin{pmatrix} X_{carrier} & 0 \\ 0 & X_{non-carrier} \end{pmatrix}$$

and Figure 2 is the schematization of the modified design

matrix. We separate the data by carrier group and non-carrier group and assigned those data to the first and fourth block elements of the matrix. The remained elements are filled with 0. The same SNPs in the carrier group and the non-carrier are grouped into one group. We expect that if a SNP has no predictive power in both groups, then both of the coefficients should be zero. On the other hand, when a SNP is useful for prediction, then both of the coefficients are non-zero.



**Figure 2. Diagram of design matrix for group lasso**

### III. RESULTS

In this study we investigated whether the approach of stratification by APOE types was enriched the prediction performance compared to the combined model. Prior to compare the predictability, we assessed the odds ratio resulted from GWAS for the combined data, carrier group

data and non-carrier group data. Figure 3 show the Q-Q plots and manhattan plots of combined and separated data. Figure 4 shows the odds ratios of 100 SNPs which have the lowest p-value for each case and Figure 5 gives the distribution of negative log p-value of selected top SNPs. For these results, in general, for the combined data (grey), the most of large effects SNPs are on chromosome 19 where APOE gene locates and the others sporadically distributed having modest odds ratio, approximately 1.5. Because of the extreme effect of APOE gene, susceptible SNPs seen in separated groups cannot be found in the combined data. Looking at the odds ratios in separate groups, for APOE e4 carrier group (red), some have high odds ratio exceeding 3 and the others have low value, less than 1. On the other hands, APOE e4 non-carrier group (blue) doesn't seem to have high effect SNPs as carrier group, instead, most of SNPs have between 1 and 2 odds ratio. We inferred that the limited number of SNPs strongly affect AD in carrier group, but combination of a large number of small effect SNPs is associated with AD in non-carrier group. If we analyze with combined data, those tendency could not be reflected in the model.

The APOE is the strongest known genetic risk factor for AD. In the presence of APOE alone, the AUC was 0.6183 and it had improved to 0.6782 when age and sex are

adjusted. However, in contrast to the expectation and the other studies (Desikan, et al., 2017; Escott–Price, et al., 2015; Lee, et al., 2012), PRS could not help to enhance the prediction accuracy, even up to 10,000 SNPs were included. We followed the approach previously described by GERAD (Genetic and Environmental Risk for Alzheimer' s disease) and IGAP (International Genomics of Alzheimer' s disease) (Escott–Price, et al., 2015) and got AUC 0.6518 for our data. We consider this value as a reference to compare our results. To see the differences of non–penalized method penalized methods (Table 4, Table5), we compared the AUCs of those models on the test data changing the number of SNPs to be included in the model. The best accuracy was found in penalized methods for both the combined case and the separate cases (combined AUC and Meta AUC column in Table 5). For the combined case, the largest AUC was 0.6520 with only 100 SNPs and 0.6671 with 10,000 SNPs, in PRS model and in ridge regression model, respectively. For the separate case, AUC 0.6551 with 100 SNPs and 0.6741 with 1,000 SNPs, in PRS model and in ridge regression model, respectively. When we crossly combined the models where  $y \sim \text{APOE4} + \text{APOE2} + \text{AGE} + \text{SEX}$  model for the carrier group and ridge regression model with 10,000 SNPs for non–carrier group, AUC was 0.6773.

We further investigated whether the stratified strategy helps to improve AD prediction. For each model when the number of SNPs is fixed, stratified model outperformed the combined model when relatively smaller number of SNPs are included but it was opposite when the number of SNPs is large. However, there's no need to use same model for the carrier group and the non-carrier group, so we crossly calculated AUC and got the best accuracy 0.6773, as described in the previous paragraph.

To explore the different coefficients by the carrier group and non-carrier group for the same SNP, we did group lasso analysis. Unlike the separated data by APOE type, but like combined data, it used the same top  $p$  SNPs for the carrier group and the non-carrier group. The resulted AUC was the lowest among the listed models in Table 6. In addition, due to the large dimensionality, group lasso could not fit the model of 10,000 SNPs with grpreg function of {grpreg} package in R.

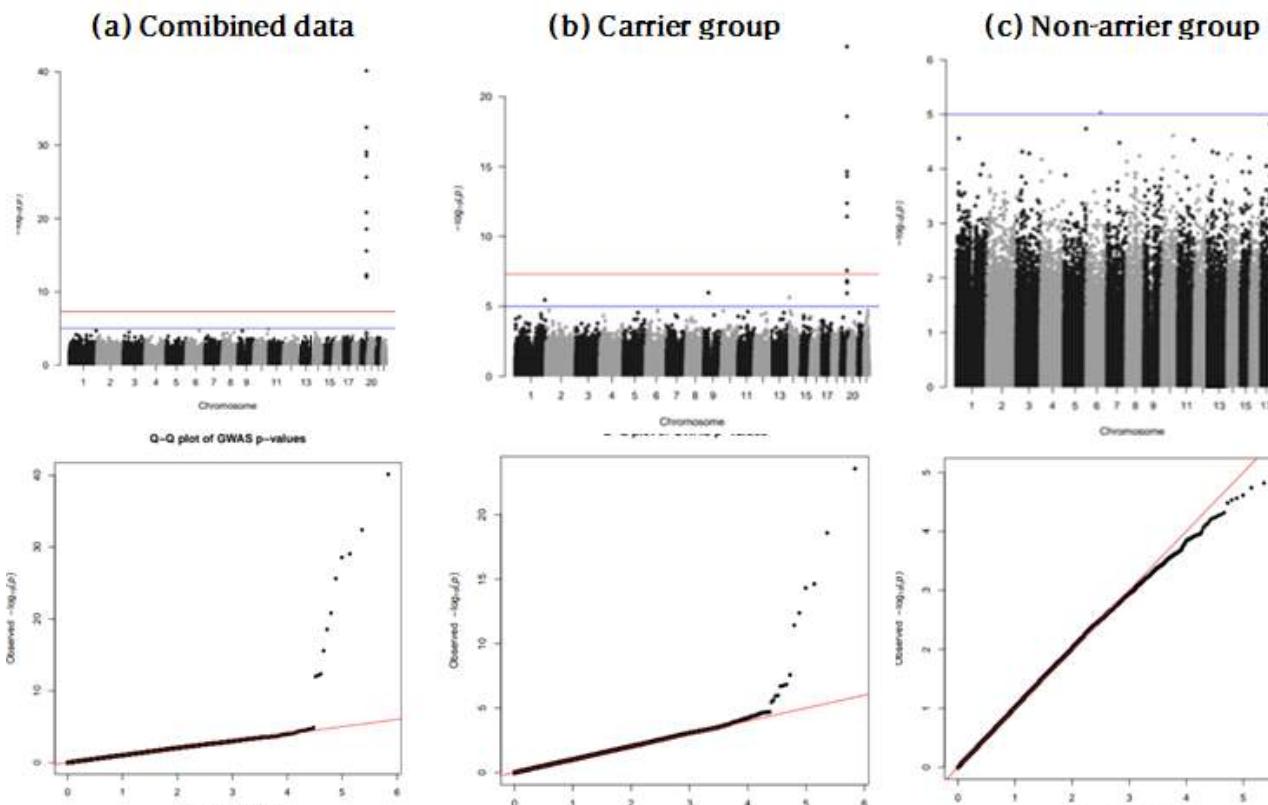


Figure 3. Manhattan plot (up) and Q–Q plot of combined (left) and separate (right) analysis

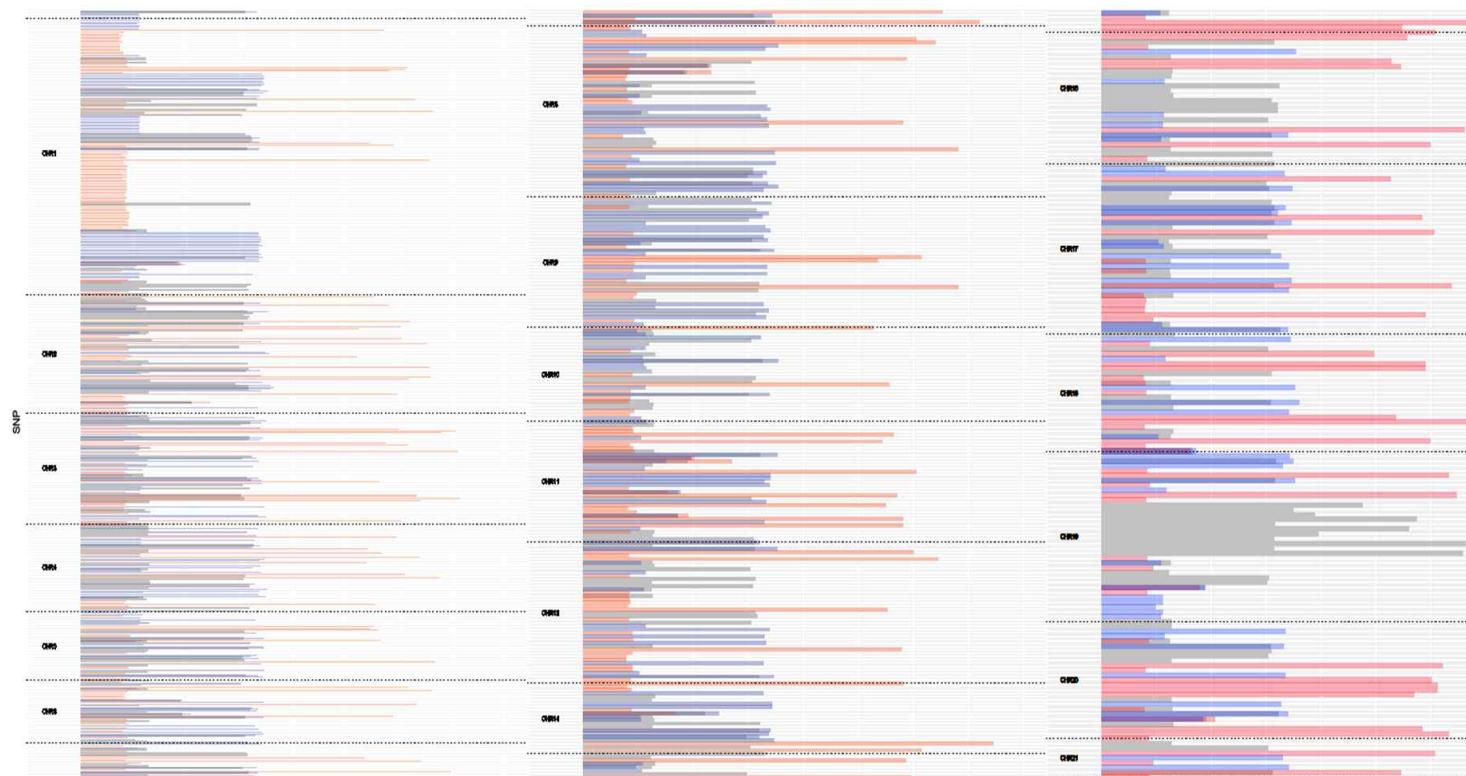


Figure 4. Odds ratio of the largest 100 SNPs in the combined data(grey), carrier group(red) and non-carrier(blue) group

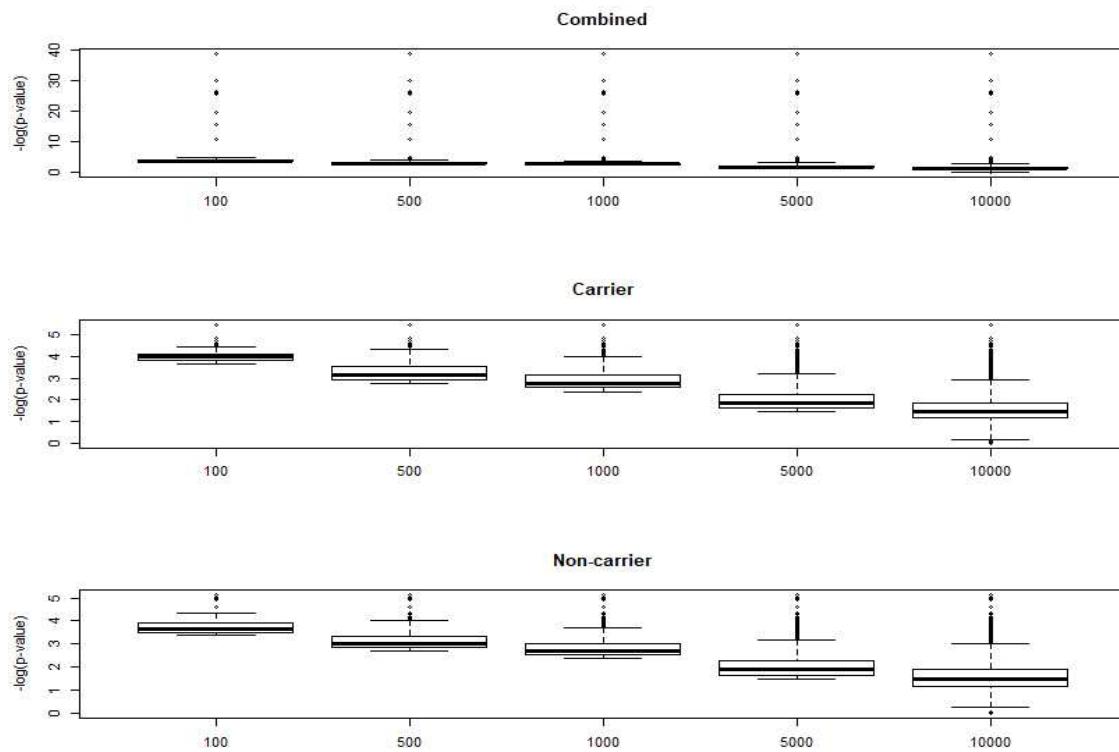


Figure 5. Negative log p-value of selected top SNPs for combined, carrier and non-carrier group

	Model	# of SNP	Penalty	AUC
	Y~APOE4+APOE2			0.6183
	Y~APOE4+APOE2 +AGE+SEX			0.6782
Unpenalized Regression	Y~APOE4+APOE2 +AGE+SEX+PRS	100	Ridge	0.6520
		500	EN0.2	0.6236
		1,000	EN 0.4	0.6182
		5,000	EN 0.6	0.6363
		1,0000	EN 0.8	0.6334
		100	Lasso	0.6051
		500	Ridge	0.5956
		500	EN 0.4	0.5917
		500	EN 0.6	0.5911
Penalized Regression	Y~APOE4+APOE2 +AGE+SEX +top SNP	1,000	EN 0.8	0.5882
		1,000	Lasso	0.5911
		5,000	Ridge	0.6059
		5,000	EN0.2	0.5885
		5,000	EN 0.4	0.5870
		5,000	EN 0.6	0.5803
		5,000	EN 0.8	0.5826
		5,000	Lasso	0.5805
		10,000	Ridge	0.6128
		10,000	EN0.2	0.5803
	Y~APOE4+APOE2 +AGE+SEX +top SNP	1,000	EN 0.4	0.5754
		1,000	EN 0.6	0.5749
		1,000	EN 0.8	0.5818
		1,000	Lasso	0.5867
		5,000	Ridge	0.6659
		5,000	EN0.2	0.6360
		5,000	EN 0.4	0.6308
		5,000	EN 0.6	0.6277
		5,000	EN 0.8	0.6257
		5,000	Lasso	0.6238
	Y~APOE4+APOE2 +AGE+SEX +top SNP	10,000	Ridge	0.6671
		10,000	EN0.2	0.6486
		10,000	EN 0.4	0.6405
		10,000	EN 0.6	0.6376
		10,000	EN 0.8	0.6351
		10,000	Lasso	0.6342

Table 4. Predictive accuracy for combined data

	Model	# of SNP	Penalty	Carrier AUC	Non-carrier AUC	Meta AUC
Unpenalized Regression	Y~APOE4 +APOE2			0.5588	0.5281	0.6157
	Y~APOE4 +APOE2 +AGE+SEX			0.5800	0.5943	0.6767
	Y~APOE4 +APOE2 +AGE +SEX +PRS	100		0.5450	0.5916	0.6551
		500		0.5635	0.5899	0.6270
		1,000		0.5441	0.5914	0.6233
		5,000		0.5126	0.5804	0.5978
		1,0000		0.5387	0.5692	0.5917
			Ridge	0.5517	0.5943	
			EN0.2	0.5547	0.5838	
		100	EN 0.4	0.5514	0.5888	0.6288
			EN 0.6	0.5548	0.5844	
			EN 0.8	0.5451	0.5842	
			Lasso	0.5410	0.5811	
Penalized Regression	Y~APOE4 +APOE2 +AGE +SEX +PRS +topSNP	500		0.5478	0.5883	
			EN0.2	0.5436	0.5767	
			EN 0.4	0.5501	0.5843	0.6174
			EN 0.6	0.5565	0.5858	
			EN 0.8	0.5565	0.5845	
			Lasso	0.5533	0.5839	
			Ridge	0.5862	0.5922	
			EN0.2	0.5666	0.5907	
		1,000	EN 0.4	0.5637	0.5888	0.6741
			EN 0.6	0.5654	0.5825	
			EN 0.8	0.5696	0.5884	
			Lasso	0.5774	0.5850	
		5,000		0.5652	0.5887	
			EN0.2	0.5501	0.5874	
			EN 0.4	0.5520	0.5891	0.6324
			EN 0.6	0.5563	0.5888	
			EN 0.8	0.5626	0.5904	
			Lasso	0.5689	0.5889	
			Ridge	0.5612	0.5956	
			EN0.2	0.5414	0.5886	
			EN 0.4	0.5495	0.5879	0.6410
			EN 0.6	0.5607	0.5884	
			EN 0.8	0.5662	0.5865	
			Lasso	0.5704	0.5849	

Table 5. Predictive accuracy for separated data

Model	# of SNP	AUC
Y ~ APOE4+APOE2+AGE+SEX +top SNP	100	0.5989
	500	0.6053
	1,000	0.6075
	5,000	0.6244

Table 6. Predictive accuracy of group lasso regression

## IV. DISCUSSION

This study supports that AD has different polygenic architectures according to APOE types. This implies that the genetic architecture of AD includes many common variants of small effects and the kinds and effects of its susceptible genes are different between APOE e4 carriers and non-carriers. First, the results of GWAS for combined data and separated data have shown that different kinds of SNPs affect AD with different effects. Second, we show that stratified analysis improves AUC over combined one. For extension of our analysis, we may identify people of high risk of AD without any APOE e4 alleles. That is, the suggested method can provide more variation in estimated risk in the population.

However, when we separate the groups, the sample size is small for prediction, 647 in carrier group and 1,725 in non-carrier group. We did 10-fold cross-validation to prevent overfitting problem, then only 60 samples were used

to train the model for the carrier group. Thus, group lasso method is adopted to solve the limit of the sample size and at the same time to estimate different coefficients for each group. The result was not good enough as the separated analysis and combined analysis. Group lasso selects or does not select a SNP simultaneously in both the carrier group and the non-carrier group. However, as we explore odds ratio from GWAS for each group, it would be better to grouping by SNP, but one of estimated coefficients could be zero within the selected SNPs. It means, even we predefined groups of predictors, we can select variables within the groups like lasso does.

In further studies, we require to understand the specific genetic factors that comprise the polygenic component by each APOE groups. To capture the susceptible SNPs in each group, stratified GWAS are needed and so that large sample size will be required to pass the genome-wide significant threshold.

## REFERENCE

- Bird, T.D. Alzheimer disease overview. 2015.
- Chatterjee, N., Shi, J. and Garc  a-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* 2016;17(7):392–406.
- Corder, E., *et al.* Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature genetics* 1994;7(2):180–184.
- Delaneau, O., Coulonges, C. and Zagury, J.-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics* 2008;9(1):540.
- Desikan, R.S., *et al.* Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS medicine* 2017;14(3):e1002258.
- Escott-Price, V., *et al.* Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* 2015;138(12):3673–3684.
- Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 2001;96(456):1348–1360.
- Gim, J., *et al.* Improving Disease Prediction by Incorporating Family Disease History in Risk Prediction Models with Large-Scale Genetic Data. *Genetics* 2017;207(3):1147–1155.
- Hoerl, A.E. and Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*

1970;12(1):55–67.

Howie, B. and Marchini, J. IMPUTE2. In.: 2011.

Jiang, S., *et al.* A Systems View of the Differences between APOE  $\epsilon$ 4 Carriers and Non-carriers in Alzheimer's Disease. *Frontiers in aging neuroscience* 2016;8.

Lambert, J.-C., *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics* 2013;45(12):1452–1458.

Lee, S.H., *et al.* Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Human molecular genetics* 2012;22(4):832–841.

Meier, L., Van De Geer, S. and Bühlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2008;70(1):53–71.

Ridge, P.G., *et al.* Alzheimer's disease: analyzing the missing heritability. *PloS one* 2013;8(11):e79771.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996:267–288.

Won, S., *et al.* Evaluation of penalized and nonpenalized methods for disease prediction with large-scale genetic data. *BioMed research international* 2015;2015.

Yang, J., *et al.* GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 2011;88(1):76–82.

Zhang, Z., *et al.* Improving the accuracy of whole genome

prediction for complex traits using the results of genome wide association studies. *PloS one* 2014;9(3):e93017.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005;67(2):301–320.

# 국문초록

## 다유전적 위험 구조을 고려한 알츠하이머 질병 예측

서 수 진  
서울대학교 보건대학원  
보건학과 보건통계전공

**배경:** 알츠하이머 질병은 다유전적 구조를 가지고 있는 것으로 알려져 있는데, 이는 작은 효과를 가지는 수많은 단일 염기 다형성(SNP)이 동시에 작용하여 알츠하이머에 영향을 미치는 것을 의미한다. 또한 이전 연구에 따르면 APOE e4 보유자와 그렇지 않은 사람의 유전적 발병 요인은 다르다고 알려져 있다. 알츠하이머에 대한 APOE e4의 효과가 매우 크기 때문에, APOE e4 보유자와 그렇지 않은 사람을 같이 분석하게 되면, APOE 유전자 외 다른 유전적 요인을 파악하기 어렵다.

**목적:** APOE 유형에 따라 알츠하이머에 미치는 다유전적 요인이 다른지를 확인하고자 한다. 다유전적 구조를 고려하기 위하여 벌점화 회귀 모형을 이용하여 알츠하이머를 예측한 후, 다유전적 위험 점수의 방법과 예측 성능을 비교하고자 한다. 또한 APOE e4의 여부에 따라 데이터를 나눠서 분석하는 계층화 방법과 그렇지 않은 방법을 비교하고자 한다.

**방법:** 본 논문에서는 다유전적 위험 점수의 대안으로 벌점화 회귀 모형을 사용하였다. 다유전적 위험 점수와 달리 벌점화 회귀 모형에서는 여러 유전자를 개별의 변수로 간주하여 유전자들 간의 복합 작용을 고려할 수 있다. 또한 추정된 회귀 계수에 벌점을 부과함으로서, 샘플의 수보다 변수의 개수가 많은 경우에 적용가능하다는 장점이 있다. 특히, lasso, elastic-net 등과 같은

방법은 변수 추정과 동시에 변수 선택을 수행함으로 결과 해석을 용이하게 해 준다. Group-lasso 방법은 lasso의 확장된 방법으로서 사전에 정의한 그룹 수준에서 변수를 선택합니다. 본 논문에서는 APOE e4 보유 그룹과 그렇지 않은 그룹의 단일 염기 다형성을 그룹화한 후, group-lasso 방법을 적용하였다. 더불어 APOE e4 보유 여부에 따라 계층화함으로서 두 그룹의 매커니즘을 개별적으로 탐색하였다. 앞서 제시한 방법들을 치매국책연구단의 한국인 데이터에 적용하였다.

**결과:** 결합된 데이터와 계층화한 데이터 각각에 대해 전장 유전체 연관 분석(GWAS)을 한 후, P-value가 가장 낮은 100개 SNP의 오즈비를 비교하였다. 결합된 데이터에서는 대부분의 SNP이 APOE 유전자가 위치한 19번 유전체에 분포하고 있었다. APOE e4 보유 그룹의 경우, 특정 SNP의 강한 효과를 가지고 있으며, 그 외 다른 SNP들은 낮은 오즈비를 가지고 있었다. 반면 APOE e4 비보유 그룹에서는 대부분의 SNP들이 1.5 정도의 오즈비를 가지고 있는 것으로 추정되었다.

별점화 회귀 모형과 다유전적 위험 점수를 사용한 모형의 예측 성능을 비교하자면, 결합된 데이터와 계층화된 데이터 모두에서 별점화 회귀 모형의 AUC가 더 높게 나타났다. 결합된 데이터에서 별점화 회귀 모형의 AUC는 0.6671, 다유전적 위험 점수를 사용한 모형에서는 0.6571로 추정되었다. 계층화된 데이터에서 별점화 회귀 모형의 AUC는 0.6773, 다유전적 위험 점수를 사용한 모형에서는 0.6551이었다.

또한, 계층화한 방법이 알츠하이머 예측 성능 향상에 도움이 되는지를 살펴보았다. 모델에 사용한 SNP의 개수가 고정된 경우, 적은 수의 SNP을 사용하였을 때는 계층화된 방법의 성능이 더 우월하였지만, 많은 수의 SNP을 사용하였을 때는 결합된 방법의 성능이 더 좋게 나타났다. 그럼에도 불구하고, 가장 예측 성능이 좋았던 모델은 계층화된 데이터에 APOE e4 보유 그룹에서는 APOE, 나이, 성별 정보를 이용하고 비보유 그룹에서는 10,000개의 SNP을 이용한 모델에서 가장 높은 AUC 값을 볼 수 있었다 (AUC = 0.6773).

**결론:** 본 연구를 통해 APOE 유형에 따라 알츠하이머에 미치는 다유전적 요인이 다름을 확인할 수 있었다. 결합된 데이터와 계층화된 데이터에 대한 GWAS의 결과는 서로 다른 종류의 SNP들이 다른 효과로 알츠하이머에 영향을 미친다는 것을

보여준다. 이를 뒷받침 하는 결과로서 알츠하이머 예측 모형에서 계층화된 방법이 더 높은 AUC를 보여주었다. 또한 벌점화 회귀 모형이 기존의 다유전적 위험 점수를 이용한 방법의 성능이 더 좋은 것을 살펴볼 수 있었다. 즉, 벌점화 회귀 모형을 이용한 계층화 분석 방법은 전체 인구 집단의 알츠하이머 위험을 좀 더 설명할 수 있다고 판단할 수 있다.

**주요어:** 알츠하이머, 다유전적 구조, 예측, 벌점화 회귀모형, 릿지, 라쏘, 그룹라쏘

**학번:** 2016-24008