2022 년 2 월

석사학위 논문

Disease risk prediction incorporating genetic variants from association analysis with disease phenotype and endophenotype

조선대학교 대학원

글로벌바이오융합학과

김 윤 태

Disease risk prediction incorporating genetic variants from association analysis with disease phenotype and endophenotype

질병 표현형과 내적표현형 연관 유전변이를 활용한 질병위험예측

2022 년 2 월 25 일

조선대학교 대학원

글로벌바이오융합학과

김 윤 태

Disease risk prediction incorporating genetic variants from association analysis with disease phenotype and endophenotype

지도교수 김 정 수

이 논문을 이학 석사학위 신청 논문으로 제출함

2021 년 10 월

조선대학교 대학원

글로벌바이오융합학과

김 윤 태

김윤태의 석사학위논문을 인준함

위원장 조선대학교 교 수 김 석 준 (인)

위 원 조선대학교 교 수 김 정 수 (인)

위 원 조선대학교 교 수 김 규 민 (인)

2021 년 12 월

조선대학교 대학원

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Disease risk prediction incorporating genetic variants from association analysis with disease phenotype and endophenotype

Yoontae Kim

Advisor: Prof. Jungsoo Gim, Ph.D.

Department of Integrative Biological Sciences

Graduate School of Chosun University

Genome-wide association study (GWAS) using vast amounts of genetic data generated by high-throughput technologies such as microarray and next generation sequencing discovered new genes and paths that were previously unknown, and expected that disease prediction, diagnosis, and treatment would develop.

However, in the case of complex diseases, the performance of models made of Polygenic Risk Score (PRS), which is commonly used, shows low performance of around 0.6 area under the curves (AUCs). This indicates that various heterogeneity of the disease cannot be considered with the existing GWAS using the phenotype of the disease. As an alternative, it is expected to improve performance by using endophenotype to find new genetic variations related to diseases and develop disease prediction models.

Alzheimer's disease (AD) initially accumulates amyloid beta protein in the brain, and then hyperphosphorylation of Tau protein occurs, resulting in neurodegeneration and ultimately dementia. Early diagnosis is essential for AD because symptoms appear after neurodegeneration.

However, an accurate early diagnosis of AD is only a method of measuring the amount of amyloid beta or tau by positron emission tomography or cerebrospinal fluid. These methods are expensive, time-consuming, and invasive. Therefore, there is a need for an early diagnosis method that is inexpensive and less time-consuming and non-invasive or a screening method to find a high-risk group for Alzheimer's disease.

In this study, GWAS using phenotype and structural magnetic resonance imaging, an endophenotype of AD, was conducted using Korean AD data obtained from Gwang-ju alzheimer's & related dementias cohort. Then a new PRS model was developed using genetic variants found through this.

Three types of models were developed, using genetic variants known to be associated with AD through existing GWAS analysis (AD Gene model), a model using genetic variants related to newly discovered genes (Novel Gene model), and a combination of genetic variants related to AD and newly discovered genes (AD & Novel Combine model). For real-world application of the model, accuracy, sensitivity, and specificity at optimal threshold were evaluated as performance, and the best model was the AD Gene model.

When measuring performance with two independent validation datasets, the AD Gene model has an average AUC of 72%, an average accuracy of 66.5%, an average sensitivity of 79.5%, and an average specificity of 53%. In addition, we measured cumulative proportion of amyloid positivity incidence across the age, two groups predicted as amyloid positive and negative showed a clear separation. The ability of separation was clearer in mild cognitive impairment subjects but also valid in subjects with the cognitive normal. From these results, it was confirmed that the performance of the model was improved by using the endophenotype, and the average sensitivity was 79.5%, it is expected that this model can be applied to screening tests that classify AD high-risk groups.

I. INTRODUCTION

I-1. Importance and limitation of GWAS

Through the Human Genome Project (HGP) conducted from 1990 to 2003, most of the human genome could be read, and it was expected that diagnosis, prevention, and treatment for many diseases would develop (Collins et al., 2003). Using vast amounts of genomic information obtained by the development of high-throughput technologies such as microarray and next generation sequencing (NGS) after HGP, genome-wide association study (GWAS) has discovered numerous single-nucleotide polymorphisms (SNPs) associated with various diseases over the past 16 years. And it could discover new genes and pathways that were previously unknown.

Many studies have been conducted to develop disease prediction, diagnosis, and treatment by applying new genes and pathway information identified by GWAS. Since most of the SNPs resulting from GWAS exist in the non-coding area and linkage disquilibrium (LD) exists, studies to screen which of the numerous mutations are causal mutations and studies to determine what causal variants function in biological mechanisms are needed (Michael D. Gallagher, 2018). These studies will allow us to find the SNPs that have a decisive effect on the disease among the numerous SNPs and to find fundamental disease treatments based on a more accurate pathophysiological understanding. Another post GWAS study is model development for disease diagnosis and prediction. Models using the commonly used polygenic risk score (PRS) show the risk of getting a disease by combining the effects of numerous SNPs affecting the disease with one score. An individual's PRS is an immutable value set at birth, so we can use it to prevent diseases.

However, the genomic prediction of complex disease is generally performing poorly, around 0.6 to 0.7 area under the curves (AUCs). (Louis Lello, 2019) One of these causes is that there are numerous complex actions that occur in the intermediate process, such as transcriptomics,

protomics, and metabolomics, until genetic variation determines the phenotype of the disease, and it is difficult to predict these effects only with phenotype information. Therefore, it would be helpful to add intermediate information for more accurate prediction.

I-2. Imaging GWAS

Existing GWAS has limitations in predicting complex disease, so studies have been conducted to find SNPs related to endophenotypes of disease. For brain diseases such as alzheimer's disease (AD) and parkinson disease (PD), GWAS using brain imaging information was performed, and additional SNPs that were not found in conventional GWAS were found (Zhiyuan Xu, 2017). This means that new SNPs that existing GWAS has not discovered using phenotype can be discovered as endophenotypes.

In the case of brain diseases, even in the same patient, there is heterogeneity that causes abnormalities in different brain regions. But through imaging GWAS, SNPs associated with each brain region can be found. So it is expected to be possible to predict diseases using mutations considering the heterogeneity of brain diseases.

I-3. Alzheimer’s disease

AD is the most common degenerative brain disease that causes dementia and gradually develops, and cognitive function deterioration, including memory, progresses. In general, AD is known to cause hyperphosphorylation of tau as amyloid beta accumulates in the brain, resulting in neurodegeneration and ultimately dementia. During the accumulation of amyloid beta and p-tau, no symptoms appear, making it difficult to detect the disease early, and even if symptoms appear later and detect the disease, treatment is impossible other than slowing the symptoms. Therefore, it is most important to detect AD early to remove amyloid beta or prevent phosphorylation of tau.

Currently, methods for diagnosing AD include examining postmortem brain tissue and checking the amount of amyloid beta and tau accumulated in the brain through positron emission tomography (PET) images or cerebrospinal fluid (CSF). In addition, there are methods of checking brain atrophy with magnetic resonance imaging (MRI) and checking cognitive dysfunction through neuropsychological assessments such as mini-mental state exam (MMSE). Methods such as MRI and neuropsychological tests are difficult to use for early diagnosis because only patients who have already developed neurodegeneration can know. PET or CSF can be diagnosed at the asymptomatic stage, and although it is highly accurate, it is an invasive method, and it takes a lot of money and time, so a cheaper and faster diagnosis method or a screening method that can primarily classify AD high-risk groups is needed.

I-4. Genomic prediction of AD

Numerous GWAS analyses associated with AD have been performed, and many SNPs associated with AD have been identified accordingly (Shea J Andrews, 2020). Creating an AD prediction model using these SNPs will help early diagnosis by predicting AD risk with genome data obtained from blood or buccal swab. According to Valentina Escott Price's papers, well-known for the development of the AD PRS model, AUC performs around 70-80%. However, there is an overfitting problem, so if it is corrected, the performance will be further reduced. In addition, large-scale GWAS is mainly a study of caucasians, so SNPs found in GWAS results are also the result of caucasian standards, which can lead to poor performance if predicted for East Asians.

Therefore, in this study, Korean AD data were obtained and GWAS analysis was conducted using AD diagnosis information and subcortical volume, an endophenotype, and a PRS-based AD prediction model was developed using the results. In addition, many studies use AUC values to measure the performance of predictive models, which are not good indicators for real

world application, so performance evaluation was evaluated by measuring accuracy, sensitivity, and specificity at optimal thresholds for each model.

II. MATERIALS AND METHOD

II-1. Dataset description

All analyses in this study were conducted using data from GARD (Gwangju Alzheimer's & Related Dementias) cohort research center. The GARD cohort research center is a single longitudinal research center established to develop clinical, imaging, genes, and other biomarkers for early diagnosis and tracking of dementia. This research protocol was approved by the Institutional Review Committee of Chosun University Hospital. All volunteers or certified guardians for the cognitive impaired agreed in advance prior to participation.

II-1.1. Characteristics of the discovery samples

In this study, genetic variation characteristics for predicting Alzheimer's risk were examined using four study samples (2 samples for training and 2 samples for test). First, in order to discover genetic variations associated with atrophy by brain region taken with magnetic resistance images (MRI), subjects with both structural MRI (sMRI) and genetic information were extracted from the GARD database, of which 3303 subjects were finally analyzed in consideration of conditions over the age of 60. The basic characteristics of the sample are shown in Table II-1.1

In order to focus on subcortical areas related to AD among brain areas, data from 1401 subjects whose diagnostic information was clearly diagnosed as cognitive normal (CN) and Dementia were extracted. The extracted data showed relatively balanced characteristic values for each diagnostic group except for age, and the specific characteristic values are shown in Table II-1.2.

The second sample is a Diagnosis-GWAS sample consisting of 1980 (AD 990) subjects to perform case-control GWAS analysis. The case group defined the diagnostic information of

GARD Database as one of preclinical AD, prodromal AD, AD Dementia, diagnosed as amyloid PET positive, and over 60 years old, and the control group defined the diagnostic information as CN and over 73 years old. The characteristics of the detailed Diagnosis-GWAS sample are shown in Table II-1.3.

II-1.2. Characteristics of the validation samples

The performance of the predictive model was evaluated using two different samples for validation, including amyloid PET positive/negative information, a more accurate pathological diagnosis information of AD. For reference, the validation sample is independent of the two MRI-GWAS sample and the Diagnosis-GWAS sample described above. The case (control) group definition of the validation sample was based on amyloid positive (negative), and the age criterion of the case group was defined as over 60, the age criterion of the control group was over 73, and the second sample was over 71. The application of different age criterion in the control group is to balance the number of subjects in the case/control group, and both criterion is judged to be acceptable age. The characteristics of the two verification samples are shown in Table II-1.4

Table II-1.1. Demographics of sMRI-GWAS sample

sMRI GWAS	**sMRI sample (n=3303)**
Age, mean (SD)	73.52 (0.1)
Female, (%)	1922 (58.2)
ICV, mean (SD)	1441113 (2676.76)
APOE e4 alleles (0/1/2, (%))	2428/831/44 (73.5/25.2/1.3)

Table II-1.2. Demographics of a sample for finding AD-associated subcortical regions

	Control Group (n=1195)	**Case Group (n=206)**	**P-value (t-test or prop-test)**
Age, mean (SD)	72.94 (0.16)	76.47 (0.44)	8.33×10^{-13}
Female, (%)	713 (59.7)	115 (55.8)	0.3378
ICV, mean (SD)	1442192 (4292.79)	1426214 (11331.49)	0.1597

Table II-1.3. Demographics of Diagnosis-GWAS & Training sample

Diagnosis-GWAS & Training	**Control Group (n=990)**	**Case Group (n=990)**	**P-value (t-test or prop-test)**
Age, mean (SD)	78.05 (0.11)	74.37 (0.22)	$< 2.2 \times 10^{-16}$
Female, (%)	624 (63)	526 (53.1)	9.966×10^{-6}
APOE e4 alleles (0/1/2, %)	772/203/6 (78.7/20.7/0.6) NAs : 9	570/351/68 (57.6/35.5/6.9) NAs : 1	-
APOE e4 carrier, (%)	209 (21.3)	419 (42.4)	$< 2.2 \times 10^{-16}$

Table II-1.4. Demographics of validation samples

	1st validation data			2nd validation data		
	Control Group (n=133)	**Case Group (n=135)**	**P-value (t-test or prop-test)**	**Control Group (n=128)**	**Case Group (n=130)**	**P-value (t-test or prop-test)**
Age, mean (SD)	78.28 (3.6)	74.92 (5.27)	4.668×10^{-9}	75.3 (4.06)	70.96 (6.49)	7.566×10^{-10}
Female, (%)	63 (46.7)	60 (45.1)	0.8945	69 (53.9)	65 (50)	0.6148
APOE e4 alleles (0/1/2/, (%))	108/26/0 (80.6/19.4/0) The number of NAs : 1	41/81/11 (30.8/60.9/8.3)	-	94/32/1 (74.0/25.2/0.8) The number of NAs : 1	44/75/11 (33.8/57.7/8.5)	-
APOE e4 carrier, (%)	26 (19.4)	94 (69.2)	7.35×10^{-16}	33 (26.0)	86 (66.2)	2.423×10^{-10}

II-2. Genotyping & QC

Genotyping used microarray using K-chip produced for Koreans, and PLINK version 1.9 was used for quality control (QC). Subjects with a sample call rate of 95% or less and heterogeneity greater than average were removed. Multidimensional scaling (MDS) analysis was performed to remove subjects who deviated a lot from the cluster. Through the Identity By Descent (IBD) calculation, subjects with P(IBD=2)>0.9 were removed with only one subject left. Sex inconsistency or sex uncertainty subjects were also removed. SNPs with a call rate of 95% or less, a hard-Weinberg Equilibrium (HWE) of 1e-6 or less, and a minor allele frequency (MAF) of 0.01 or less were removed.

II-3. MRI acquisition & QC

MRI was taken at Chosun University Hospital, and 0.8 mm sagittal MPRAGE images were obtained under conditions of TE 2.143 ms, TR 2300 ms, TI 900 ms, FoV 256x256, 9 filp angle, 178 slices, matrix 320x320 with 3T MR scanner (Skyra, Siemens). The T1-weighted image was processed using an automatic processing stream to reconstruct the three-dimensional cortical surface model of FreeSurfer software (version 5.3). Briefly, processing includes motion correction, T1 image normalization, non-brain tissue removal, Talairach transformation, cortical WM and GM structural division, strength normalization, tessellation and topology correction of GM/WM boundaries, etc (Tetiana Gorbach, 2017).

II-4. Feature selection

Feature selection was carried out using 5-fold Cross Validation. For each CV, a logistic regression model or linear regression model and a mixed model were performed, and P-value and best linear unbiased prediction (BLUP) values were measured. Age and sex were corrected with covariates, and in the case of models using subcortical volume, the number of APOE e4

alleles and ICV (Intracranial Volume) was additionally corrected. In the case of sMRI-GWAS, the subcortical region associated with AD was selected and only that area was used for analysis and standardized volume values were used. For each CV, 1000 SNPs were selected from the lowest P-value and 1000 SNPs from the largest absolute value of BLUP, and statistics for these SNPs were estimated in models using all samples.

To find the AD specific subcortical regions, a stepwise logistic regression model was performed with 1401 subjects extracted from the sMRI-GWAS sample. Six of the 16 subcortical regions (right lateral ventrical, right pallidum, left hippocampus, right hippocampus, right amygdala, left nucleus accumben) were included in the final model. (Table II-4.1) Age, sex, and ICV were corrected for covariates.

As a result of sMRI-GWAS analysis, 19074 SNPs were selected based on P-value, and 17598 SNPs were selected based on BLUP. As a result of the Diagnosis-GWAS analysis, 3463 SNPs were selected based on P-value and 3257 SNPs based on BLUP. In the two results, the total number of SNPs excluding duplication is BLUP 20163 and P-value 22340, respectively.

Subcortical volumes were estimated using FreeSurfer (version 5.3), P-value was estimated using Plink (version 1.9), BLUP was estimated using GCTA (version 1.92), and all analyses were conducted by R Software (version 3.6).

Table II-4.1. Stepwise logistic regression result for finding AD-associated subcortical regions

Step Result	Estimate	Std.Error	z value	Pr(>|z|)
Intercept	15.43	2.172	7.102	1.23×10^{-12}
RLatVent	5.625×10^{-5}	1.196×10^{-5}	4.703	2.56×10^{-6}
Lhippo	-0.001535	2.864×10^{-4}	-5.360	8.31×10^{-8}
Rhippo	-0.001398	3.040×10^{-4}	-4.598	4.27×10^{-6}
Ramyg	-0.001295	5.664×10^{-4}	-2.287	0.0222
Laccumb	-0.002351	0.001088	-2.160	0.0307
Rpal	8.017×10^{-4}	3.783×10^{-4}	2.119	0.0341
AGE	-0.08083	0.02049	-3.945	7.97×10^{-5}
SEX2	-0.3643	0.2214	-1.646	0.0998

II-5. PRS generation and predictability modelling

PRS was generated using two values: the Likelihood Ratio (LR) and β. Previous studies use clumping or pruning for LD correction, which ignores the effects of other SNPs other than the tagging SNPs. Therefore, we divided the associated SNPs by gene and weighted each SNP considering MAF to represent the effects of SNPs associated with one gene as an integrated value and used them for PRS calculation.

The LR-based PRS consists of a total of three steps, and the first is to calculate the LR value for each SNP, and the ratio between the control group and the case group for each genotype of the k-th SNP is obtained. The second step is to calculate the LR value for each gene, and to sum all of the LR values for each k-th SNP associated with the nth gene by multiplying the weight w considering MAF. In the last step, the final LR PRS is calculated by multiplying all the LR values for each gene obtained in the second step.

STEP 1 : $LR_{SNP_k} = \frac{Pr(SNP_k's\ genotype\ in\ case)}{Pr(SNP_k's\ genotype\ in\ control)}$

STEP 2 : $LR_{gene_n} = \sum_{i=1}^{k} w_i * LR_{SNP_i}$ $\left(w_i = \frac{\frac{1}{MAF_i}}{\sum_{i=1}^{k}\frac{1}{MAF_i}}\right)$

STEP 3 : $LR\ PRS = \prod_{i=1}^{n} LR_{gene_i}$

In the equation, k denotes the number of SNPs associated with the nth gene, and n denotes the number of genes included in the model.

PRS based on β also consists of a total of three steps, and the first step is to estimate the β for each SNP, and to estimate the effect size of the minor allele of the k-th SNP by performing logistic regression on the phenotype using a discovery sample. Age and sex were corrected as covariates, and β were NA-treated when the p-value of β_{SNP_k} was less than 0.05. The second step is to calculate the β for each gene, summing all of the β of k SNPs associated with the n-th gene by multiplying the β by the number of minor

alleles (g_{SNP_k}) and the weight w considering MAF. In the last step, the final β PRS is calculated by summing all the β values for each gene obtained in the second step.

STEP 1 : $\text{Phenotype} \sim \alpha + g_{SNP_k} * \beta_{SNP_k} + age * \beta_{age} + sex * \beta_{\text{sex}} \rightarrow \beta_{SNP_k}$

STEP 2 : $\beta_{gene_n} = \sum_{i=1}^{k} w_i * g_{SNP_i} * \beta_{SNP_i}$ $\left(w_i = \frac{\frac{1}{MAF_i}}{\sum_{i=1}^{k}\frac{1}{MAF_i}}\right)$

STEP 3 : $\beta\, PRS = \sum_{i=1}^{n} \beta_{gene_i}$

In the equation, k denotes the number of SNPs associated with the n-th gene, and n denotes the number of genes included in the model.

Many studies use AUC to measure the performance of the model, which is not a suitable method for applying the model in the real world. Therefore, a method of setting an optimal threshold for the model and evaluating the performance of the results of actually predicting the disease was used. β PRS and LR PRS in each model were calculated for each subject of training (discovery) sample, and each PRS was divided into percentiles. Each quantile was used as a threshold to calculate the accuracy, sensitivity, and specificity of each threshold, and the optimal threshold with the highest accuracy was selected and used as the diagnostic criterion for the validation sets.

Three types of models were developed: the AD Gene model, the Novel Gene model, and the AD & Novel Combine model. The AD Gene model used only SNPs associated with genes reported by previous GWAS. The Novel Gene model used SNPs associated with unknown genes in the previous GWAS, and the AD & Novel combine model used a combination of AD gene SNPs and Novel gene SNPs.

III. RESULTS

III-1. Direct application of GWAS finding in Korean AD

In order to determine whether known AD related SNPs are also applied to Korean data, the PRS model (Caucasian AD SNP model) was developed with SNPs known as previous GWAS and then the performance was measured. When the performance of the Caucasian AD SNP model was measured in two independent validation sets used in this study, the average performance of the LR PRS model was AUC 70%, accuracy 63%, sensitivity 62%, and specificity 64%, and the average performance of the β PRS model was AUC 74% , accuracy 67% and specificity 69%. (Table III-1.1, Figure III-1.1) The LR PRS model showed poor overall performance. The β PRS model showed good performance, but low performance in the training set, and the model's performance was unstable depending on the dataset. In addition, sensitivity is important for screening AD high-risk groups but sensitivity was measured to be low. Through these results, it was confirmed that the PRS model was developed using only the SNP known as the previous GWAS and applied to Koreans, showing poor performance.

Table III-1.1 Performance of Caucasian AD SNP model in validation sets

Model	Type	SNP	Gene	Best threshold	Training			1st Validation			2nd Validation			AUC (Train/1st Val/2nd Val)
					ACC	SEN	SPC	ACC	SEN	SPC	ACC	SEN	SPC	
Caucasian AD SNP	LR	84	49	0.3546	0.6323	0.5323	0.7323	0.6754	0.6917	0.6593	0.5814	0.5385	0.625	0.678/0.734/0.670
	BETA	84	49	0.6167	0.6374	0.5273	0.7475	0.7127	0.6992	0.7259	0.6473	0.6385	0.6562	0.668/0.772/0.708

A Performance of LR based Reported AD SNP model

	True Aβ+	True Aβ-		
Predict Aβ+	92	46	138	PPV 0.67
Predict Aβ-	41	89	130	NPV 0.68
	133	135	268	
	SEN 0.69	SPC 0.66	ACC 0.68	

1st validation set

	True Aβ+	True Aβ-		
Predict Aβ+	70	48	118	PPV 0.59
Predict Aβ-	60	80	140	NPV 0.57
	130	128	258	
	SEN 0.54	SPC 0.63	ACC 0.58	

2nd validation set

B Performance of β based Reported AD SNP model

	True Aβ+	True Aβ-		
Predict Aβ+	93	37	130	PPV 0.72
Predict Aβ-	40	98	138	NPV 0.71
	133	135	268	
	SEN 0.70	SPC 0.73	ACC 0.71	

1st validation set

	True Aβ+	True Aβ-		
Predict Aβ+	83	44	127	PPV 0.65
Predict Aβ-	47	84	131	NPV 0.64
	130	128	258	
	SEN 0.64	SPC 0.66	ACC 0.65	

2nd validation set

Figure III-1.1 Performance of Caucasian AD SNP model in validation sets. A is the confusion matrix and performance table of the Caucasian AD SNP model based on LR PRS, and B is the confusion matrix and performance table of the Caucasian AD SNP model based on β PRS. The confusion matrix shows the result of predicting Aβ (Predict Aβ in row) for each validation set and the actual Aβ information (True Aβ in column), and the table outside the confusion matrix shows the sensitivity (SEN), specificity (SPC), accuracy (ACC), and positive prediction value (PPV) for the negative prediction value (NPV).

III-2. Performance of the three models in validation sets

In this study, GWAS analysis using phenotype and endophenotype was conducted to develop three types of models using LR and β based PRS. (Figure III-2.1) As a result of measuring the performance of the three models in two independent validation datasets, the Novel Gene model performed well in the training set, but the performance was very poor in the validation set. (Table III-2.1) In the case of the LR based AD Gene model, the average performance in the two validation sets was good with AUC 72% and accuracy 66% and sensitivity 80%, and specificity 53%. (Table III-2.1, FigureIII-2.2) The LR-based AD & Novel Combine model also showed better performance than the LR based AD Gene model in the training set, with average performance AUC 70%, accuracy 64%, sensitivity 79%, and specificity 50%. (Table III-2.1, FigureIII-2.3)

Both the LR-based AD Gene model and the LR based AD & Novel Combine model showed good performance, but the LR based AD & Novel Combine model showed much lower performance in the second validation set than other results, so the LR based AD Gene model was selected as the best model because it showed more stable performance.

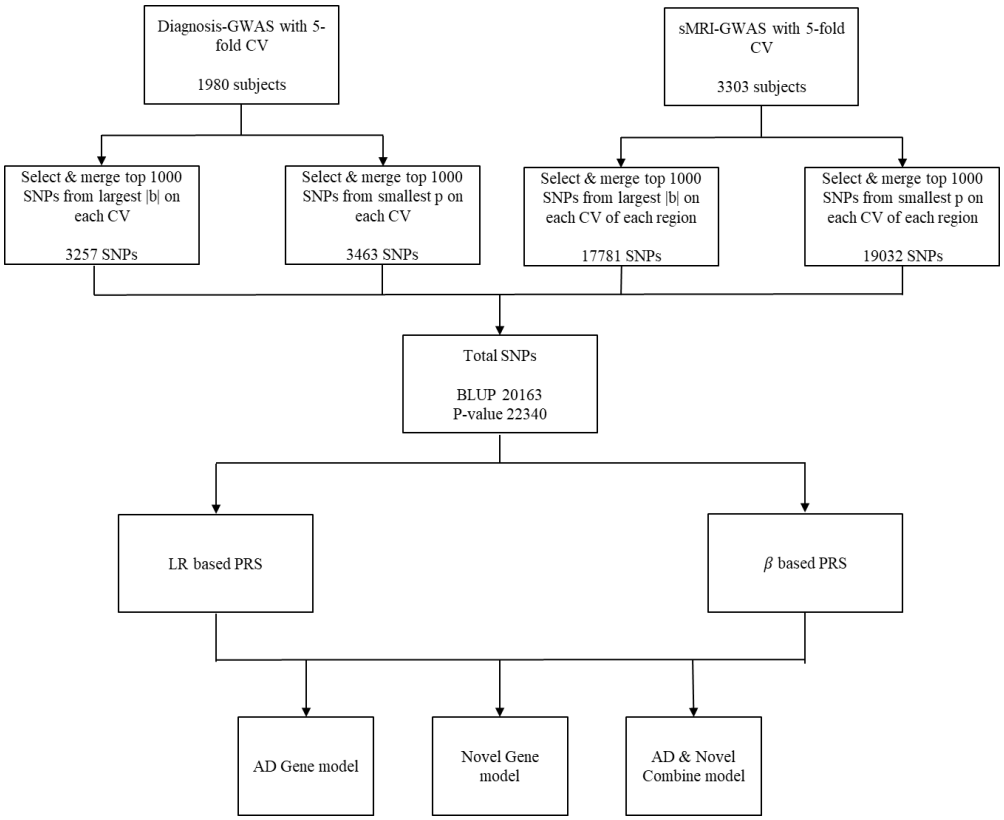

Figure III-2.1 Overall workflow of the analysis. It shows the overall flow of analysis. First, Diagnosis-GWAS and sMRI-GWAS are performed by applying 5 fold cross validation. Next, the p-value and BLUP values are measured for each CV to select the most significant 1000 SNPs. The selected SNPs are then combined to create an AD SNP pool, and then an AD Gene model, Novel Gene model, and AD & Novel Combine model are developed using a PRS based on LR and β.

Table III-2.1 Performance of the three models in validation sets

Model	Type	SNP	Gene	Best threshold	Training			1st Validation			2nd Validation			AUC (Train/1st Val/2nd Val)
					ACC	SEN	SPC	ACC	SEN	SPC	ACC	SEN	SPC	
Novel Gene	BLUP LR	193	183	-0.2346	0.698	0.7879	0.6081	0.5373	0.5564	0.5185	0.469	0.5385	0.3984	0.766/0.544/0.492
	BLUP BETA	43	41	-0.1396	0.6561	0.7859	0.5263	0.5336	0.5789	0.4889	0.4845	0.5846	0.3828	0.725/0.540/0.490
	Pval LR	194	160	-0.5505	0.7429	0.7828	0.703	0.5336	0.5188	0.5481	0.4961	0.5615	0.4297	0.811/0.523/0.514
	Pval BETA	40	37	-0.5962	0.6934	0.6535	0.7333	0.5261	0.4211	0.6296	0.5698	0.5385	0.6016	0.767/0.502/0.577
AD Gene	Pval LR	183	35	0.2746	0.6611	0.7212	0.601	0.6754	0.8045	0.5481	0.6512	0.7923	0.5078	0.718/0.746/0.693
	Pval BETA	60	20	0.6323	0.6419	0.5919	0.6919	0.7015	0.7594	0.6444	0.624	0.6923	0.5547	0.696/0.733/0.664
AD & Novel Combine	Pval LR	191	43	-0.0136	0.6859	0.7758	0.596	0.6828	0.8195	0.5481	0.6008	0.7538	0.4453	0.758/0.743/0.660
	Pval BETA	63	23	0.231	0.6611	0.6212	0.701	0.7015	0.7744	0.6296	0.624	0.7	0.5469	0.721/0.737/0.667

A Performance of LR based AD Gene model

	True Aβ+	True Aβ-		
Predict Aβ+	107	61	168	PPV 0.64
Predict Aβ-	26	74	100	NPV 0.74
	133	135	268	
	SEN 0.80	SPC 0.55	ACC 0.68	

1st validation set

	True Aβ+	True Aβ-		
Predict Aβ+	103	63	166	PPV 0.62
Predict Aβ-	27	65	92	NPV 0.71
	130	128	258	
	SEN 0.79	SPC 0.51	ACC 0.65	

2nd validation set

B Performance of β based AD Gene model

	True Aβ+	True Aβ-		
Predict Aβ+	101	48	149	PPV 0.68
Predict Aβ-	32	87	119	NPV 0.73
	133	135	268	
	SEN 0.76	SPC 0.64	ACC 0.70	

1st validation set

	True Aβ+	True Aβ-		
Predict Aβ+	90	57	147	PPV 0.61
Predict Aβ-	40	71	111	NPV 0.64
	130	128	258	
	SEN 0.69	SPC 0.55	ACC 0.62	

2nd validation set

Figure III-2.2 Performance of AD Gene model in validation sets. A is the confusion matrix and performance table of the LR PRS based AD Gene model, and B is the confusion matrix and performance table of the β PRS based AD Gene model. The confusion matrix shows the result of predicting Aβ (Predict Aβ in row) for each validation set and the actual Aβ information (True Aβ in column), and the table outside the confusion matrix shows the sensitivity (SEN), specificity (SPC), accuracy (ACC), and positive prediction value (PPV) for the negative prediction value (NPV).

A Performance of LR based AD & Novel combine model

	True Aβ+	True Aβ-		
Predict Aβ+	109	61	170	PPV 0.64
Predict Aβ-	24	74	98	NPV 0.76
	133	135	268	
	SEN 0.82	SPC 0.55	ACC 0.68	

1st validation set

	True Aβ+	True Aβ-		
Predict Aβ+	98	71	169	PPV 0.58
Predict Aβ-	32	57	89	NPV 0.64
	130	128	258	
	SEN 0.75	SPC 0.46	ACC 0.60	

2nd validation set

B Performance of β based AD & Novel combine model

	True Aβ+	True Aβ-		
Predict Aβ+	103	50	153	PPV 0.67
Predict Aβ-	30	85	115	NPV 0.74
	133	135	268	
	SEN 0.77	SPC 0.63	ACC 0.70	

1st validation set

	True Aβ+	True Aβ-		
Predict Aβ+	91	58	149	PPV 0.61
Predict Aβ-	39	70	109	NPV 0.64
	130	128	258	
	SEN 0.70	SPC 0.55	ACC 0.62	

2nd validation set

Figure III-2.3 Performance of AD & Novel Combine model in validation sets. A is the confusion matrix and performance table of the LR PRS based AD & Novel Combine model, and B is the confusion matrix and performance table of the β PRS based AD & Novel Combine model. The confusion matrix shows the result of predicting Aβ (Predict Aβ in row) for each validation set and the actual Aβ information (True Aβ in column), and the table outside the confusion matrix shows the sensitivity (SEN), specificity (SPC), accuracy (ACC), and positive prediction value (PPV) for the negative prediction value (NPV).

III-3. Performance of the best model in varying symptom

In order to confirm that the best model works well even in different AD stages, the predictive performance of the best model was confirmed with subjects with amyloid PET results. 1,020 subjects aged 60 or older with PET results and genetic information were used, of which 415 had mild cognitive impairment (MCI) diagnosis and 283 had CN diagnosis. Figure III-3.1 shows cumulative proportion of amyloid positivity incidence across the age, two groups predicted as amyloid positive and negative showed a clear separation. The ability of separation was clearer in MCI subjects but also valid in subjects with the CN. As a result of this, it was confirmed that the best model worked well regardless of the stage of the disease.

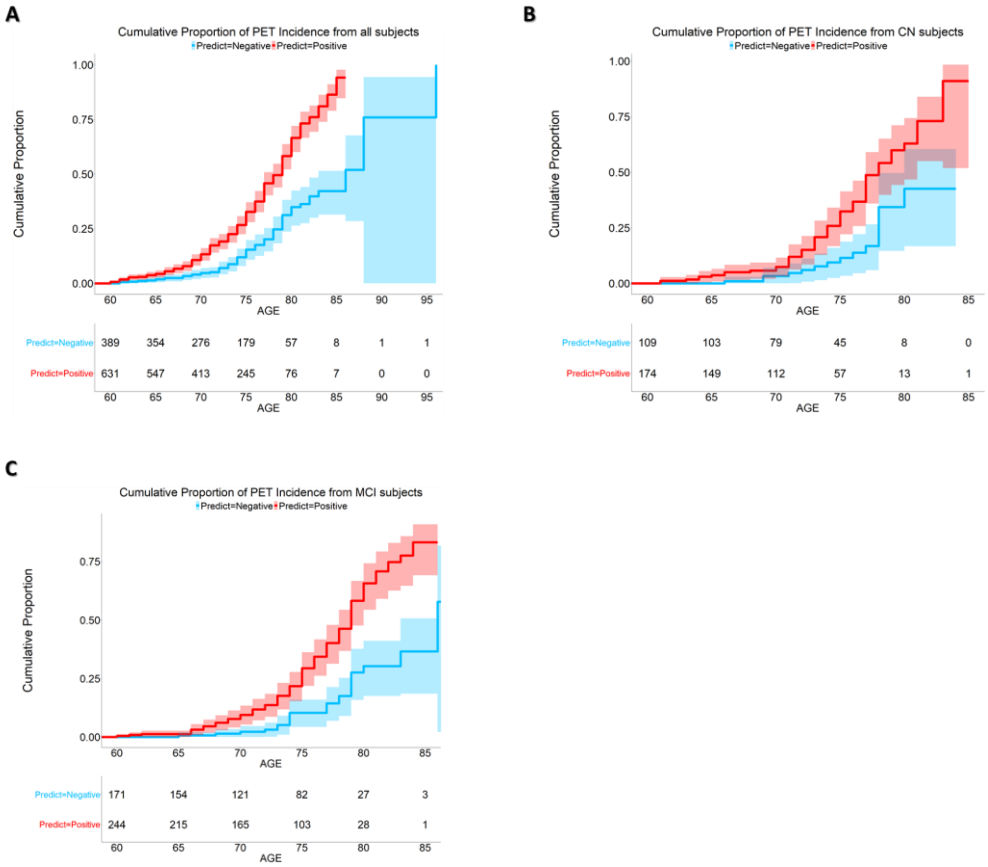

Figure III-3.1 Prediction performance of the best model in varying symptom. In the graph, the x-axis represents age and the y-axis represents cumulative proportion of amyloid positivity incidence. The blue line is the group predicted as Aβ negative, the red line is the group predicted as Aβ positive, and the area around the line is the 95% confidence interval. The table below the graph shows the total number of subjects within each group that have not yet been diagnosed with positive at the age of the x-axis. A is a graph for all patients with Aβ information (1020 subjects), B is for patients diagnosed with CN (283 subjects), and C is for patients diagnosed with MCI (415 subjects).

IV. DISCUSSION

Although the performance of PRS based models using genetic variants obtained from conventional GWAS studies was about 60% AUC, the Aβ prediction PRS model for AD prediction developed by us showed good performance of 70% AUC on average in two independent trials. In addition, as a result of checking performance based on accuracy, sensitivity, and specificity of real world application, it was confirmed that amyloid beta-positive patients were well selected with an average sensitivity of 79%. And the best model divides the amyloid positive group and the amyloid negative group well regardless of the disease stage. Therefore, we demonstrated that additional information such as endophenotype improves the performance of the model.

These results suggest that although insufficient for accurate early diagnosis of AD, the best model can be used as the first test method for screening AD high-risk groups. It is expected that more efficient early AD diagnosis will be possible through accurate AD diagnosis through amyloid PET or CSF testing if AD high-risk group is selected using this PRS model.

In this study, previously known AD-related SNPs were Caucasian based GWAS results, so SNPs related to Korean AD patients were selected using Korean data from the GARD cohort. However, because the sample size is small, GWAS with a larger sample size is needed to discover Korean AD SNPs and obtain the exact effect size of SNPs.

Previous studies have shown that there are various heterogeneity in AD (Zhang B, 2021). However, the current PRS model does not take into account the various heterogeneity of AD, and the next study expects to perform better by dividing AD into several subgroups and then finding genetic variants related to specific AD subgroups to predict AD risk.

V. FURTHER STUDY

As mentioned in Discussion, in order to develop a PRS model considering the heterogeneity of AD in the future study, some analyses were conducted for find regional atrophy patterns of the brain using longitudinal sMRI, amyloid PET, and diagnosis data from GARD database.

First, the average atrophy rate for each brain region was calculated using a subcortical volume that corrected sex and ICV for each subject, and PET positive groups were divided into subgroups using several clustering groups using several clustering methods. After that, linear regression analysis was performed to confirm that the average atrophy rate by brain region was significantly different between the PET negative and the CN-diagnosed cognitive unimpairment (CU) group and the PET positive subgroup. In addition, linear regression was performed for each disease status within the subgroup to confirm the difference in average atrophy rate by brain region according to disease status.

In Table V-1, group 1 showed a significant difference from CU in most regions. Depending on the disease state, it can be seen that atrophy occurs rapidly in more regions as it progresses from CN to MCI, and it can be seen that there is little difference between CU and atrophy speed in Dem stage. group2 showed a significant difference only in left caudate, contracted faster than CU in four areas in CN, showed a significant difference only in left caudate in MCI state, and did not show a significant difference from CU in Dem. (Table V-1)

These results confirmed from Korean AD data that there are heterogeneity in which the rate of atrophy varies from region to region of the brain, and considering this, it can be expected that the predictive model will perform better.

Table V-1.1 Result of linear regression between CU and PET^{+} subgroup

A Result of linear regression between CU and PET+ subgroup

Region		LLatVent	RLatVent	Lthal	Rthal	Lcaud	Rcaud	Lput	Rput	Lpal	Rpal	Lhippo	Rhippo	Lamyg	Ramyg	Laccumb	Raccumb
Group1 (54)	P-value	1.64×10^{-6}	2.09×10^{-9}	2.00×10^{-5}	0.000492	0.001577	0.790627	7.21×10^{-7}	4.79×10^{-8}	0.794336	0.007511	0.013109	0.002522	0.001339	0.00845	0.884043	0.755582
	FDR	6.58×10^{-6}	3.34×10^{-8}	6.40×10^{-5}	0.001313	0.003155	0.847291	3.85×10^{-6}	3.83×10^{-7}	0.847291	0.012018	0.017479	0.004483	0.00306	0.012291	0.884043	0.847291
Group2 (69)	P-value	0.092685	0.087836	0.008431	0.193084	2.43×10^{-6}	0.048373	0.014568	0.014634	0.168735	0.452666	0.024828	0.892281	0.61207	0.484198	0.553251	0.092028
	FDR	0.164773	0.164773	0.058538	0.280849	3.88×10^{-5}	0.128995	0.058538	0.058538	0.269977	0.595936	0.079451	0.892281	0.652875	0.595936	0.632287	0.164773

B Result of linear regression between CU and each disease status in Group1

Region		LLatVent	RLatVent	Lthal	Rthal	Lcaud	Rcaud	Lput	Rput	Lpal	Rpal	Lhippo	Rhippo	Lamyg	Ramyg	Laccumb	Raccumb
CN^{+} (10)	P-value	0.007803	2.03×10^{-5}	0.002604	0.060633	0.58818	0.642453	2.70×10^{-5}	4.56×10^{-7}	0.964333	0.259522	0.180679	0.249941	0.011444	0.532177	0.001849	0.599169
	FDR	0.020809	0.000144	0.008333	0.121265	0.684764	0.685283	0.000144	7.29×10^{-6}	0.964333	0.377486	0.321207	0.377486	0.026157	0.684764	0.007394	0.684764
MCI^{+} (22)	P-value	6.03×10^{-7}	7.11×10^{-9}	0.000163	0.00069	0.003743	0.882252	0.007342	5.48×10^{-5}	0.685568	0.159533	0.009786	0.000659	0.002381	0.006935	0.472106	0.2728
	FDR	4.83×10^{-6}	1.14×10^{-7}	0.000652	0.00184	0.007487	0.882252	0.011747	0.000292	0.731273	0.212711	0.014235	0.00184	0.005442	0.011747	0.53955	0.335754
Dem^{+} (12)	P-value	0.050701	0.002041	0.114387	0.295472	0.248547	0.189018	0.025852	0.037044	0.211644	0.275396	0.894914	0.814053	0.307065	0.032028	0.506645	0.694158
	FDR	0.162243	0.03265	0.305032	0.40942	0.40942	0.40942	0.148176	0.148176	0.40942	0.40942	0.894914	0.868324	0.40942	0.148176	0.623563	0.793323
MCI to CN (10)	P-value	0.261711	0.302702	0.147944	0.09683	0.00074	0.059238	0.001936	0.000684	0.182029	0.001294	0.171634	0.305241	0.608066	0.77744	0.305244	0.123861
	FDR	0.348851	0.348851	0.291247	0.258213	0.005924	0.189562	0.007743	0.005924	0.291247	0.006901	0.291247	0.348851	0.648603	0.77744	0.348851	0.283111

C Result of linear regression between CU and each disease status in Group2

Region		LLatVent	RLatVent	Lthal	Rthal	Lcaud	Rcaud	Lput	Rput	Lpal	Rpal	Lhippo	Rhippo	Lamyg	Ramyg	Laccumb	Raccumb
CN+ (11)	P-value	0.436657	0.616911	0.000837	0.131029	0.007746	0.141991	0.000827	0.174007	0.002713	0.330674	0.926523	0.740439	0.040713	0.865584	0.811397	0.557635
	FDR	0.698652	0.822548	0.006692	0.324551	0.030985	0.324551	0.006692	0.348015	0.014468	0.587864	0.926523	0.911309	0.130282	0.92329	0.92329	0.811105
MCI+ (31)	P-value	0.174958	0.098077	0.708604	0.607258	1.88×10^{-6}	0.115118	0.296366	0.049974	0.385655	0.809653	0.071242	0.565238	0.284174	0.863982	0.356255	0.314806
	FDR	0.466553	0.368377	0.809833	0.747394	3.01×10^{-5}	0.368377	0.559655	0.368377	0.560952	0.863629	0.368377	0.747394	0.559655	0.863982	0.560952	0.559655
Dem+ (11)	P-value	0.367553	0.528517	0.270911	0.548955	0.094375	0.446501	0.841672	0.48481	0.650996	0.992417	0.273162	0.80308	0.027406	0.270351	0.170586	0.00922
	FDR	0.731939	0.731939	0.62437	0.731939	0.503331	0.731939	0.897783	0.731939	0.801225	0.992417	0.62437	0.897783	0.219246	0.62437	0.62437	0.147522
MCI to CN (11)	P-value	0.769018	0.7018	0.013141	0.72242	0.892228	0.387529	0.112573	0.000581	0.010943	0.093179	0.001277	0.261466	0.565268	0.215958	0.273365	0.706786
	FDR	0.820286	0.820286	0.052564	0.820286	0.892228	0.620046	0.300194	0.00929	0.052564	0.298173	0.010216	0.485983	0.820286	0.485983	0.485983	0.820286

VI. 초 록

질병 표현형과 내적표현형 연관 유전변이를 활용한 질병위험예측

김 윤 태

지도교수 : 김 정 수

글로벌바이오융합학과

조선대학교 대학원

인간 게놈 프로젝트 이후 마이크로어레이와 차세대 염기서열 분석법같은 대량신속처리 기술이 발전하면서 생성된 방대한 양의 유전체 자료를 활용한 전장유전체연관분석을 통해 이전에는 알지 못했던 새로운 유전자와 경로들을 발견하였고, 이를 통해 질병 예측, 진단, 치료법이 발전할 것이라고 기대하였다.

그러나 복합질환의 경우 일반적으로 많이 사용되는 다중유전자위험점수로 만든 모델의 성능을 보았을 때 곡선하면적 0.6대의 낮은 성능을 보인다. 이는 질병의 표현형을 활용한 기존의 전장유전체연관성 연구로는 질병의 다양한 이질성을 고려하지 못한다는 것을 보여준다. 이에 대한 대안으로 표현형 외에 내적표현형을 활용하여

새로운 질병연관 유전변이를 찾아내어 질병예측모델을 만들면 성능이 향상될 것이라 기대된다.

알츠하이머병은 초기에 뇌에 아밀로이드베타단백질이 축적되고, 이후에 타우 단백질의 과인산화가 일어나 신경퇴화가 일어나 인지기능이 저하되고 궁극적으로 치매증상을 띄게 되는 질병이다. 알츠하이머병은 신경퇴화가 일어난 뒤에 증상이 나타나기 때문에 조기진단이 필수인 질병이다. 하지만 현재 알츠하이머병을 진단할 수 있는 방법은 아밀로이드 양전자 단층촬영이나 뇌척수액을 추출하여 아밀로이드베타의 양을 측정하는 방법 뿐이다. 이러한 방법은 비용과 시간이 많이 들고, 침습적인 방법이기 때문에 비용과 시간이 적게들면서 비침습적인 조기진단법이나 알츠하이머병 고위험군 대상자를 찾아내기 위한 선별검사방법이 필요하다.

본 연구는 광주치매코호트연구단에서 한국인 알츠하이머병 데이터를 사용하여 표현형을 활용한 전장유전체연관성연구를 통해 알아낸 유전변이와 알츠하이머병의 내적표현형인 구조적 자기공명영상에서 추출한 겉질밑 부피을 활용한 전장유전체연관성연구를 통해 얻은 유전변이를 이용하여 알츠하이머병 위험예측을 위한 새로운 다중유전자위험점수 모델을 만들었다.

총 3가지 모델을 만들어 성능을 측정하였으며 기존 전장유전체연관분석을 통해 알츠하이머병과 연관되어 있다고 알려진 유전자와 관련된 유전변이들을 사용한 모델(AD Gene model), 기존에 알려지지 않았던 새롭게 발견된 유전자에 관련된 유전변이들을 사용한 모델(Novel Gene model), 알츠하이머병과 연관된 유전자와 관련된 유전변이들과 새롭게 발견된 유전자와 관련된 유전변이들을 합쳐서 사용한 모델(AD & Novel Combine model)이다. 모델의 실제 적용을 위해 최적의

임계값에서의 정확도, 민감도, 특이도를 성능으로 평가하였고, 최적의 모델은 “AD Gene” 모델이었다.

“AD Gene” 모델은 독립적인 두 개의 검증 데이터 세트로 성능을 측정했을 때 평균 곡선하면적은 72%이며 평균 정확도는 66.5%, 평균 민감도는 79.5%, 평균 특이도는 53% 이다. 이러한 결과로부터 내적표현형을 활용하면 모델의 성능이 향상된다는 것을 확인할 수 있었으며, 평균 민감도가 79.5%로 준수한 성능을 보였다. 또한 알츠하이머병 환자 중 모든 증상이 존재하는 집단과 무증상단계 집단과 경도인지장애 집단에서 아밀로이드 베타 예측 결과를 바탕으로 두 개의 집단으로 나눠서 각 그룹별 아밀로이드 베타 양성 누적발생률을 확인 했을 때 두 집단이 잘 나뉘는 것을 확인할 수 있었다. 따라서 본 모델이 알츠하이머병의 고위험군을 구분하는 선별검사에 적용할 수 있을 것이라 기대한다.

VII. REFERENCES

Andrews, S. J., Fulton-Howard, B., & Goate, A. (2020). Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *The Lancet Neurology*, *19*(4), 326-335.

Chaudhury, S., Brookes, K. J., Patel, T., Fallows, A., Guetta-Baranes, T., Turton, J. C., Guerreiro, R., Bras, J., Hardy, J., & Francis, P. T. (2019). Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. *Translational psychiatry*, *9*(1), 1-7.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, *300*(5617), 286-290.

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., Badarinarayan, N., GERAD/PERADES, consortia, I., & Morgan, K. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain*, *138*(12), 3673-3684.

Escott-Price, V., Myers, A. J., Huentelman, M., & Hardy, J. (2017). Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Annals of neurology*, *82*(2), 311-314.

Ferreira, D., Verhagen, C., Hernández-Cabrera, J. A., Cavallin, L., Guo, C.-J., Ekman, U., Muehlboeck, J.-S., Simmons, A., Barroso, J., & Wahlund, L.-O. (2017). Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific reports*, *7*(1), 1-13.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., & Klaveness, S. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341-355.

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The post-GWAS era: from association to function. *The American Journal of Human Genetics*, *102*(5), 717-730.

Giri, M., Zhang, M., & Lü, Y. (2016). Genes associated with Alzheimer's disease: an overview and current status. *Clinical interventions in aging*, *11*, 665.

Goedert, M., & Spillantini, M. G. (2006). A Century of Alzheimer's Disease. *Science*, *314*(5800), 777-781.

Gorbach, T., Pudas, S., Lundquist, A., Orädd, G., Josefsson, M., Salami, A., de Luna, X., & Nyberg, L. (2017). Longitudinal association between hippocampus atrophy and episodic-memory decline. *Neurobiology of aging*, *51*, 167-176.

Grothe, M. J., Levin, F., Teipel, S. J., & Habes, M. (2020). Disentangling neurodegeneration subtypes of Alzheimer's disease using data-driven methods: Neurobiological subtypes of Alzheimer's disease. *Alzheimer's & Dementia*, *16*, e037183.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, *9*(1), 1-9.

Lane, C., Hardy, J., & Schott, J. (2018). Alzheimer's disease. *European journal of neurology*.

Lello, L., Raben, T. G., Yong, S. Y., Tellier, L. C., & Hsu, S. D. (2019). Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Scientific reports*, *9*(1), 1-16.

Loos, R. J. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, *11*(1), 1-3.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, *101*(1), 5-22.

Vogel, J. W., Young, A. L., Oxtoby, N. P., Smith, R., Ossenkoppele, R., Strandberg, O. T., La Joie, R., Aksman, L. M., Grothe, M. J., & Iturria-Medina, Y. (2021). Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nature Medicine*, *27*(5), 871-881.

Xu, Z., Wu, C., Pan, W., & Initiative, A. s. D. N. (2017). Imaging-wide association study: Integrating imaging endophenotypes in GWAS. *Neuroimage*, *159*, 159-169.

Zhang, B., Lin, L., Wu, S., & Al-Masqari, Z. H. (2021). Multiple Subtypes of Alzheimer's Disease Base on Brain Atrophy Pattern. *Brain Sciences*, *11*(2), 278.

국립보건연구원. (2018). 한국인칩사업 백서.