Master's Thesis of Public Health

Predicting Coronary Artery Disease Risk using Polygenic Risk Scores and Clinical Variables in the East Asian Population

동아시아 인구 집단에서의 다유전자 위험점수와 임상변수를 이용한 관상동맥질환 위험예측

February 2023

Graduate School of Public Health
Seoul National University
Public Health Major

Yuree Chung

Predicting Coronary Artery Disease Risk using Polygenic Risk Scores and Clinical Variables in the East Asian Population

Name of Examiner Sungho Won

Submitting a master's thesis of
Public Health
December 2022

Graduate School of Public Health
Seoul National University
Public Health Major

Yuree Chung

Confirming the master's thesis written by
Yuree Chung
December 2022

Chair Joohon Sung (Seal)
Vice Chair Woojoo Lee (Seal)
Examiner Sungho Won (Seal)

Abstract

Background: Coronary Artery Disease (CAD) is a disease in which the coronary arteries that supply blood to the heart are narrowed by atherosclerosis or blood clots, resulting in myocardial ischemia or myocardial infarction due to poor blood supply. CAD has a high heritability of 40% to 60%, and multiple investigation has been conducted to provide CAD with large-scale genetic data for non-Hispanic whites. However there have been no studies for the east Asian population including the Korean population.

Objective: In our study, we would like to calculate PRS for CAD in the East Asian populations using different PRS calculation methods including meta-PRS. Furthermore, through an integrated model that considers both PRS and clinical markers, we intend to predict a CAD risk considering both genetic and clinical factors.

Methods: We considered the summary statistics from BioBank of Japan (BBJ) as a reference data and those were used to calculate the weights of each SNP used for calculating PRS of 71,009 Korean samples from KoGES data. Then we calculated the PRS_{CAD} using five different method, and the statistical method was chosen with highest AUC in predicting CAD. Furthermore, we selected 8 traits related to CAD and by using the meta-PRS, and built the prediction model with meta-PRS.

Results: We found that PRS and meta-PRS had a high odds ratio (OR 1.32, 95%CI 1.26-1.39), (OR 1.35, 95%CI 1.29-1.42). Net reclassification improvement for both were 0.072±0.0127 and 0.088±0.0135 respectively. The PRS score calculated by the LDpred-auto had a significant improvement in predictive ability compared to the model with only existing clinical variables (AUC: from 0.780 to 0.785, P=0.0003).

Conclusion: In this study, the genetic effect in predicting coronary artery disease in the East Asian population group was confirmed by analyzing the effect of adding PRS to the existing clinical variable predicting coronary artery disease risk.

Keywords: Coronary Artery Disease(CAD), Polygenic Risk Scores(PRS), meta-PRS, LDpred, Ridge Regression, Elastic Net Regression, East Asian Population

Student Number: 2021-21804

Table of Contents

List of Tables

List of Figures

List of Supplementary Materials

Ⅰ. Introduction

Cardiovascular disease (CVD) which is one of the leading causes of death globally (WHO, 2022) has a high heritability of 40% to 60%[1], and various genetic studies have been conducted to identify the genetic factors. Until now, more than 100 genome-wide significant loci associated with CAD have been found with GWAS. However, the effect size of each risk allele is not large, and which is a characteristic of polygenic diseases where several genetic factors affect the specific phenotypic outcome in general. Therefore, most of the recent studies use polygenic risk scores that can consider the effects of multiple SNPs which do not have a significant standard to identify more precise genetic risks.

It was reported that an integrated risk tool using both genetic risks and clinical risks showed a 5.9% net reclassification improvement[2]. Predicting genetic risk for CAD may help develop methods used for preventive interventions such as risk reduction, behavior modification, or pharmacologic treatment. Previous studies showed that the genetic effect summarized through PRS and the environmental effect of lifestyle factors contribute independently to CAD and related diseases (atrial fibrillation, ischemic stroke, hypertension, etc.)[3]. The onset of CAD is known to be related to traits such as hypertension, dyslipidemia, type 2 diabetes, and lifestyle factors including smoking behavior, exercising, and eating habits[4]. According to the GWAS research on the Korean population, it was found the leading SNPs of CAD in patients having hypertension(17q25.3/CBX8-CBX4 rs1550676), type 2 diabetes (17q25.3/RPTOR rs139293840), and dyslipidemia(rs79166762)[5]. In the case of lifestyle variables such as

smoking initiation, a whole-genome sequencing (WGS)-based GWAS performed in a Chinese population has shown thirteen SNPs of the RFTN1 gene[6]. The rs139753473 from RFTN1 and six other suggestively significant loci from the CSMD1 gene were also associated with cigarettes per day (CPD) in an independent.

The mainstream CAD risk prediction models are considering a combination of clinical risk factors and genetic risk factors, which reports a better performance than traditional tools such as PCE alone. However, the performance improvement due to polygenic risk is not significant enough despite the high heritability[7]. There might be three possible explanations for this: limited genomic scope, limited sample sizes, and ancestry-specific differences[8].

Possible explanation for limited genomic scope arises from the fact that the CAD outbreak is affected by multiple pathways. Although PRS was calculated using all range of SNPs related to CAD in some of the previous studies, many genes related to indirect pathological pathway of CAD outbreak were excluded from calculating PRS. Therefore, these un-optimized models might have lacked in making a precise prediction model for subjects having high CAD risks. As an effort to seek a solution for this, in 2018, the concept of meta-PRS emerged, an approach where multiple PRSs of related traits are combined into one meta-score (meta-PRS) to overcome the limitations of the previous research[7]. Meta-PRS has shown a better performance in Ischemic Stroke prediction than the previously known method of LDpred, and also supported by a meta-analysis of 979,286 participant data in predicting CAD (PRS_{LDpred} HR= 1.46, Meta-PRS HR=1.67)[9].

Moreover, in the non-white population, the quality and quantity of prediction based on Caucasian population are significantly low[10]. Here, in our study, we would like to improve the performance of the

prediction model in the Korean population by using similar ethnic origins in the East Asian population (Japanese population) as reference data for calculating genetic risks. Overall, the objective of this study is to calculate the PRS for CAD in East Asian populations including Korean population and use PRS as an indicator for improving the classification of a high-risk group. Therefore, through an integrated model including both PRS and clinical markers, we intend to develop a CAD prediction model that considers both genetic and clinical factors in the East Asian population.

Ⅱ. Materials and methods

1. Study Participants

Prediction model building with PRS and its evaluation require three different dataset for base(reference), validation, and test dataset (Figure 1), and we considered the BioBank of Japan (BBJ)[11,12,13,14,15] data, the Korean Genome and Epidemiology Study (KoGES) data and the Gene-environmental interaction and phenotype (GENIE), respectively.

BBJ project was started at the Institute of Medical Science, the University of Tokyo in 2003. BBJ data consist of around 212,453 subjects with disease cases consisting of 47 various diseases, and these subjects were recruited from 12 medical institutes in Japan. Korean Genome and Epidemiology Study (KoGES) dataset was collected by the Korean Center for Disease Control and Prevention based on the Korean population. KoGES data consist of KoGES Ansan and Ansung study (KARE), the KoGES health examinee (HEXA) study, and the KoGES cardiovascular disease association study (CAVAS)[11]. KoGES dataset consisted of 81,902 participants recruited from the

national health examinee registry, and they aged more than 40-year-old at the baseline. The baseline years are 2001, 2004, and 2005 for KARE, HEXA, and CAVAS, respectively. The dataset contains participants' medico-pharmacologic history, and physical examination (weight, height, BMI, BP, etc.), and provided fasting blood samples to measure blood lipid(TG, TC, HDL-C, etc).

Test data was derived from the Gene-environmental interaction and phenotype (GENIE) provided by the Gangnam Health Center of Seoul National University Hospital[12]. The original data provided by the hospital was named Health and Prevention Enhancement (H-PEACE), but around 2,000 samples were added and used for this study. Here, we will just write the name of the whole data (H-PEACE + GENIE) as GENIE. It contains longitudinally observed measures of participants' medico-pharmacologic history, physical examination, and blood lipid and glucose levels after fasting. GENIE cohorts were followed up during 2003-2017. For this study, blood samples and the main covariate data were available for 9,348 participants, independent of subjects in the KoGES dataset.

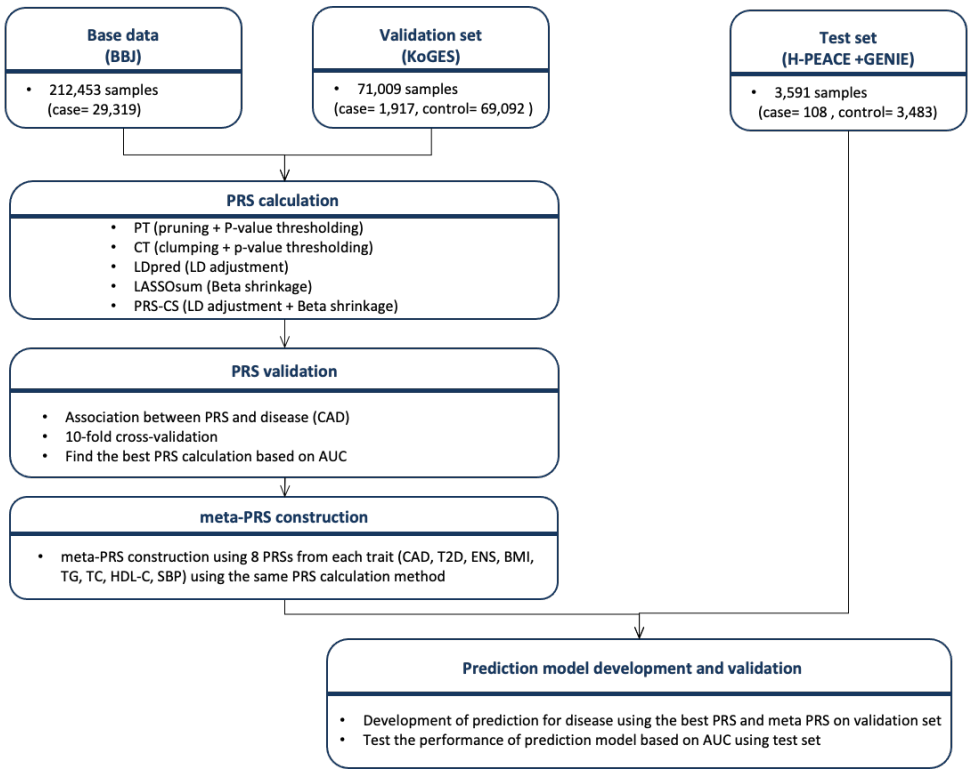

Figure 1. Flowchart of PRS prediction model development and validation

2. Genotyping, Quality Controls and Imputation

For KoGES and GENIE data, all participants were genotyped using a Korean Chip and variant calling was conducted with K-medoid algorithm to minimize the batch effect [13, 14]. Genotypes were quality-controlled to improve the genotype accuracy via the pipeline shown in Figure 2. We excluded subjects with missing genotype rate > 0.05, with 0.2 < homozygosity chrX < 0.8, and heterozygosity rate > mean ± 3 std. Then, any SNPs were filtered out if missing rates>0.05, or P-value for Hardy-Weinberg equilibrium < 10^{-5}. With the remaining SNPs and subjects, genotypes were imputed using the NARD imputation server using 1000 Genomes data as a reference penal[15]. Finally, we eliminated SNPs with low R-squared (Rsq<0.3) and multiallelic SNPs.

All QC process was performed using PLINK and ONETool[16, 17]. As a result, 71,686 subjects with 17,527,243 SNPs genotyped and 9,348 subjects with 15,948,813 SNPs genotyped remained for KoGES and GENIE data, respectively (Figure 2). Additional QC for calculating PRS was performed by following the process from the previous study: pruning out SNPs with Rsq >0.5 was conducted through PLINK[18]. 10 Principal Components(PCs) from this genotyped data were obtained for each data set using PLINK.

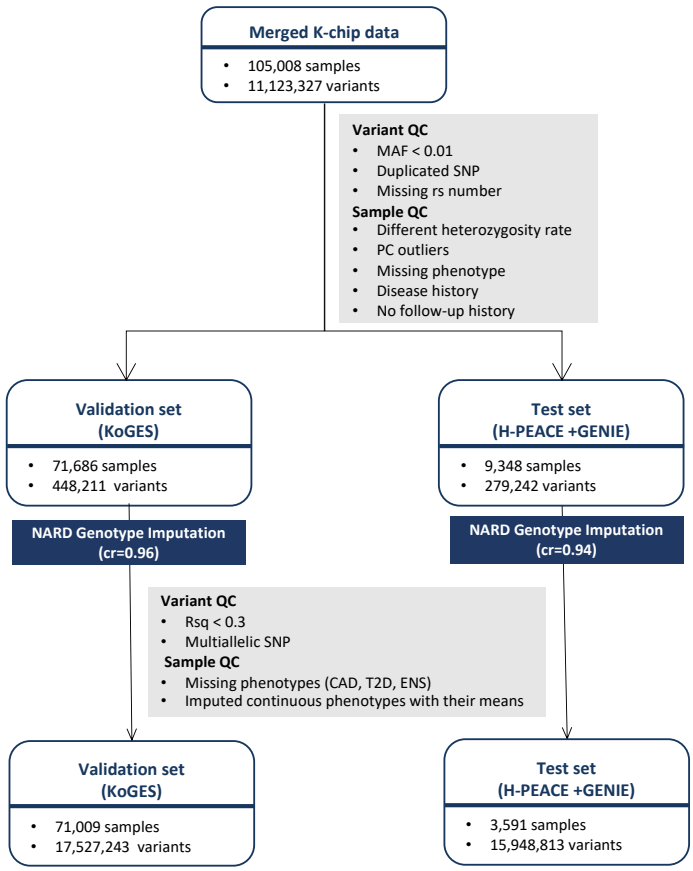

Figure 2. Flowchart of Genotyping, Quality Controls and Imputation of KoGES and GENIE data

3. Ascertainment of the Target Disease, Clinical Variables, and Lifestyle Variables

Operational definition of outcome

In this study, we have obtained summary statistics of CAD from the Jenger Riken Institute, which utilized Myocardial Infarction and Angina Pectoris BBJ summary statistics to get CAD summary statistics. Following the definition from BBJ, in KoGES and GENIE, CAD cases are composed of CAD, Myocardial Infarction, Myocardial Ischemia, and Angina Pectoris. The variables used in the model are coded as follows. The dependent variable is coded as binary, people with CAD are 1 and normal people are 0.

Clinical and lifestyle variables

Type 2 Diabetes(T2D) and Smoking Behavior(ENS) are regarded as categorical variables. People with T2D are 1 and normal people are 0. For smoking behavior, BBJ provides two types of GWAS summary statistics, Cigarettes per day (CPD) and Ever-never smoked (ENS). Since the criteria of questionnaires for KARE were different from HEXA and CAVAS cohorts, we have re-defined the smoking status as "Have you smoked more than 400 cigarettes?" and classified those who have not smoked more than 400 cigarettes as non-smokers and those who have smoked more than 400 cigarettes as smokers. For hypertension, systolic blood pressure (SBP) and diastolic blood pressure(DBP) are provided in BBJ summary statistics. For dyslipidemia, total cholesterol(TC), triglycerides(TG), and high-density lipoprotein cholesterol(HDL-C) are provided. However, GWAS summary statistics for hypertension and dyslipidemia are not provided. Therefore, instead of using binary phenotype data for each symptom, we have used continuous biochemical data for each variable given. In

particular, for dyslipidemia, TC ≥ 230 mg/dL, HDL-C< 40 mg/dL, and TG≥ 200 mg/dL are criteria for diagnosis in the Korean population. For AGE is the age of subjects itself, without any transformation, and SEX is coded as 1 to male and 2 to female. For BMI, we have excluded 4 samples whose height and weight are both missing and imputed the missing value as the average. After imputation, we calculated BMI as weight in kilograms divided by height in meters. In addition, samples with both TC and HDL-C missing are excluded and the missing values for TG are imputed with its mean.

4. Calculating Polygenic Risk Scores

PRS is the marginal effects of susceptible SNPs and is calculated as the weighted sum of risk alleles where the weights are coefficients of simple logistic regression. To construct PRS, two procedures are required, variable selection and coefficients of selected SNPs. PRS is useful for disease prediction because it is computationally efficient. There are many different methods to calculate PRS considering different interactive effects of genomic structure on SNPs. However, since it has been reported that the LDpred method showed the best performance in CAD prediction[1], we have compared the performance of six methods (LDpred-inf, LDpred-auto, P+T, C+T, lassosum, PRS-CS) of calculating PRS to see which method has the best performance in the validation set. Considering the LD structure of our genotype data, we calculated the LD score from KoGES Kchip data following the protocol from the previous study[19].

Pruning and Thresholding (P+T) refers to the strategy of first applying informed LD pruning with an R-square threshold of 0.2 and subsequently applying p-value thresholding, where the p-value

threshold is optimized over a grid concerning prediction accuracy in the validation data. Here, we only included SNPs with p-value<10^{-5} for pruning, and the LD threshold for pruning was set to 0.5. **Clumping and thresholding (C+T)** is a widely used method to derive polygenic scores[20]. The significance threshold for index SNPs and secondary significance threshold for clumped SNPs is set to 10^{-5}, and the LD threshold for clumping was set to 0.5 as well. **Lassosum** is a method for computing LASSO/Elastic-Net estimates of a linear regression problem given summary statistics from GWAS, accounting for Linkage Disequilibrium (LD), via a reference panel[21]. **PRS-CS** utilizes a high-dimensional Bayesian regression framework, and is distinct from previous work by placing a continuous shrinkage (CS) prior on SNP effect sizes, which is robust to varying genetic architectures, provides substantial computational advantages, and enables multivariate modeling of local LD patterns[22]. Lastly, **LDpred** is a Bayesian PRS that estimates posterior mean causal effect sizes from GWAS summary statistics by assuming a prior for the genetic architecture and LD information from a reference panel. Unlike P+T, LDpred has the desirable property that its prediction accuracy converges to the heritability explained by the SNPs as the sample size grows[19]. A key feature of LDpred is that it relies on GWAS summary statistics, which are often available even when raw genotypes are not. There are three options for the LDpred method (inf, grid, and auto). **LDpred-inf** (using GWAS summary statistics) is analogous to genomic BLUP (using raw genotypes) because it assumes the same prior. In **LDpred-auto**, we have calculated SNP heritability (h^2) from LD score regression, and then set initial vector p which ranges from 10^{-4} ~ 0.9 divided into either maximum of 80 or the number of cores. Then, we filtered outlier predictions and averaged the remaining predicted values of PRS.

Construction of meta-PRS

Meta-PRS was generated by integrating the eight optimal trait-specific PRSs following after the results from the previous study of meta PRS on the Chinese population[8]. The predictive performance of meta-PRS scores considering the genetic effects of various traits and PRS scores calculated by the LDpred method was compared through two models: logistic regression and ridge regression.

(1) Calculate PRS for each trait using the fixed optimal PRS method and save the weights of each SNP used for PRS calculation in the train set.

(2) 10-fold cv Elastic-net regression for the following model to find the best lambda:
CAD ~ age + sex + PC1~10 + standardized 8 PRS s

(3) Using the optimized hyperparameter (alpha and lambda) in ridge regression, obtain the beta for each PRS and save the mean, std, and effect size for each PRS.

(4) Comparing two different models with PRS_{CAD} and meta-PRS in AUC

We conducted an elastic-net logistic regression with 10-fold cross-validation using the R package *'glmnet'* to fit a parameter lambda for Ridge regression. Adjusting age, sex, and 10 PC scores, we assessed the association between the eight optimal PRSs and CAD in the training set and then obtained the effect size of the eight PRSs, which was used for the weight of meta-PRS calculation. Finally, the meta-PRS for CAD was constructed by summing the standardized optimal trait-specific PRSs weighted by adjusted estimates β_1, .., β_8 derived from the ridge regression model. The meta-PRS can be calculated via a weighted sum by using genotype data,

$$\text{metaPRS}_\text{i} = \sum_{i=1}^{m} \frac{\beta_1}{\sigma_1} x_1 + \cdots + \frac{\beta_i}{\sigma_i} x_i$$

where m is the total number of traits, σ_1,…, σ_8 are the empirical standard deviations of each of PRS in the training set, β_1, …, β_8 are the effect sizes for the i^{th} PRSs from the regression, and x_i is the PRS_i centered to zero with the mean of PRS in i^{th} trait. β_i was considered to be zero for the i^{th} trait if the PRS was not included in the variable selection. All this standardization information was saved to apply to the test data. Furthermore, we have considered a different combination of eight PRSs via penalized regression model to derive the best meta-PRS. Ridge regression methods were used to calculate the weight of each PRS in the meta-PRS calculation.

Elastic-net and Ridge regression

Elastic-net complements the ridge and lasso by penalizing with both l_1 and l_2 norm[23]. This has the effect of effectively shrinking the coefficients and setting some coefficients to zero.

Ridge regression is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters. In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias.

In this study, we employed Ridge logistic regression to model the associations between the 8 PRSs and CAD, adjusting for sex, age, and 10 genetic PCs. Also, we have calculated meta-PRS using selected PRS as a result of 10-fold CV elastic-net regression. The best model, in terms of the highest cross-validated AUC (area under receiving-operating characteristic curve), was selected as the final model and held fixed for validation in the rest of the data. The final adjusted coefficients for each PRS in the penalized logistic regression are compared with the univariate estimates.

Generation of trait-specific polygenic risk scores

Eight trait-specific PRSs (CAD, SBP, T2D, TC, TG, HDL-C, ENS, and BMI) were separately constructed by summing the number of corresponding risk alleles (0, 1, or 2) for each subject, weighted by the effect size of variants on the corresponding trait. Variants were chosen among the common SNPs with BBJ to calculate PRS. A detailed list of studies for CAD GWAS analysis in BBJ is available in Table S1.

Assuming the effect size of optimal PRS_{CAD} should be the largest in meta-PRS for CAD, we have fixed the PRS calculation method to one that showed the best AUC for predicting CAD. Therefore, optimized PRS for each trait was calculated based on trait-specific summary statistics from the large-scale BBJ GWAS in East Asian ancestry. The distribution of eight PRS is shown as a boxplot in Figure S1. Each optimal PRS was standardized by calculating the z-score (zero mean, unit standard deviation).

5. Statistical analysis

To compare the prediction accuracy, we first used PRS(PRS_{CAD}) which was calculated through the LDpred-auto in KoGES data using BBJ as reference data. Figure 3 illustrates the whole model comparisons. Model I (M1) is a null model which has age and sex as variables. Model II (M2) is a model to show the effect of PRS_{CAD} and Model II-prime (M2') shows the pure effect of PRS_{CAD} without age and sex. Model III (M3) checks whether PC scores should be included in the model to obtain a better performance or not. Model IV (M4) shows the effect of meta-PRS and Model IV-prime (M4') shows the pure effect of meta-PRS without age and sex. In addition, the final models contain PRS, seven CAD-related traits (BMI, T2D, ENS, SBP, TG, TC,

HDL-C), and other covariates (age, sex, PC_1,…, PC_{10}). PRS is put into the CAD prediction model as the main variable, and the logistic regression analysis method is used to analyze and test the significance of genetic risk factors in CAD. Accuracies of the disease prediction models were assessed via 10-fold cross-validated AUC. The final models were evaluated in the test data.

To compare the clinical usefulness of PRS, we have divided the study population into three risk groups (high, medium, and low) and compared the AUC of the integrated model and the p-value of PRS in each group. Moreover, we have calculated Net Reclassification Improvements(NRI)[24] in the total population of KoGES and GENIE using the R package *'nricens'*. For the risk difference-based NRI calculation, the cutoff value of risk difference was specified as 0.02, where UP and DOWN are defined as $p_{new} - p_{standard} > \delta$ and $p_{standard} - p_{new} > \delta$, respectively. $p_{standard}$ and p_{new} are predicted individual risks from a standard and a new prediction model, respectively, and δ corresponds to the cutoff. Interval estimation is based on the percentile bootstrap method.

Simple Models (without clinical variables)

M1 : CAD ~ age + sex

M2 : CAD ~ age + sex + **PRS_{CAD}**

M2': CAD ~ **PRS_{CAD}**

M3 : CAD ~ age + sex **+ PC_1+…+PC_{10} + PRS_{CAD}**

M4 : CAD ~ age + sex **+ metaPRS**

M4': CAD ~ **metaPRS**

Integrated Models (with scaled clinical variables)

M5 : CAD ~ age + sex + metaPRS + traditional clinical variables

M6 : CAD ~ age + sex + PRS_{CAD} + traditional clinical variables

Figure 3. Model Comparison simple models and integrated models

Ⅲ. Results

1. Descriptive Statistics

Descriptive statistics analysis for each population was done after the final subject QC. We excluded samples with either one of the binomial variables(CAD, T2D, ENS) missing, and then imputed the missing values for the continuous variables(TC, TG, HDL-C, BMI, SBP) with their means. Along with QCed genotype data, we had 71,686 samples in KoGES with 4,008,884 SNPs, and 9,847 samples in GENIE with 3,708,789 SNPs as final data for analysis.

Descriptive characteristics of KoGES were calculated in CAD case and control groups (Table 1). CAD cases are 1917 (2.7%) among 71,009 samples. Among CAD case groups, 52.1% were men and there was more percentage of people who have Type 2 diabetes (17.8%) and have smoked more than 400 cigarettes(38.9%) than control groups(6.7% and 25.9% respectively). The case group had a lower level of HDL-cholesterol (mean: 48.2, std: 11.9) than the control group. Descriptive characteristics for GENIE are attached in Table S2. Pearson correlation tests among different clinical variables were conducted in KoGES. Figure S2 shows the correlation plot of 11 variables in KoGES and Figure S3 shows the result plot of the correlation coefficient with p-value<0.05 using *R*. As a result, mean blood pressure(MBP) and DBP were excluded for the further analysis because the correlation coefficients with SBP were larger than 0.7 (0.962 and 0.764 respectively).

Table 1. Descriptive characteristics of KoGES

Risk factor	Controls	Cases
Sample size, N	69092 (97.3)	1917 (2.7)
Male, (%)	24522 (35.5)	998 (52.1)
Body mass index, kg/m^2	24 (2.9)	24.9 (2.9)
Total cholesterol, mg/dl	197.5 (35.3)	175.3 (37.7)
High-density lipoprotein cholesterol, mg/dl	52.2 (13.2)	48.2 (11.9)
Triglycerides, mg/dl	130 (88.8)	131.5 (80.6)
Systolic blood pressure, mmHg	122.7 (15.4)	125.1 (15)
Diastolic blood pressure, mmHg	76.5 (10.1)	76.6 (9.8)
Type 2 diabetes (%)	4607 (6.7)	358 (18.7)
Ever smokers(>400 cigarettes), (%)	17888 (25.9)	745 (38.9)

Values are mean (Standard deviation) or N (%)

2. Selection of PRS calculation method

In this study, we have calculated performance accuracy among different PRS toward CAD. LDpred showed the best performance in the simple logistic regression. Therefore, we have compared the AUC between two LDpred methods in a null model(M1) and M5. The M1 performance in the train set is compared with a null model which has all the covariates and clinical variables for logistic regression. As a result, the LDpred-auto method has the best significant improvement in predicting CAD in the logistic model from AUC: 0.780 to AUC: 0.785 (p-value<0.000) (Table 2 and Table 3). Therefore, we have selected the LDpred-auto method to calculate PRS for the other 7 traits to calculate meta-PRS.

Table 2. Results of simple logistic regression of CAD ~ PRS_{CAD}

method	beta	p-value	AIC	AUC
C+T	0.053	0.000	16302	0.545
P+T	0.041	0.28	16335	0.505
lassosum	1.159	0.000	16235	0.567
LDpred-inf	0.648	0.000	16257	0.545
LDpred-auto	0.572	0.000	16206	0.578
PRS-CS	0.620	0.000	16246	0.564

Table 3. Comparing the AUC of Integrated model with PRS_{CAD}

method	Null AUC	Model AUC	p-value of PRS
LDpred-auto	0.780	0.785	0.000
LDpred-inf	0.780	0.782	0.000
P+T	0.780	0.780	0.726

3. Construction of meta-PRS for CAD

Genetic correlations

The results of simple logistic regression from six different PRS calculation methods are provided in Table 4. This presents genetic and environmental correlations between the PRSs of CAD-related traits using LDpred-auto in univariate CAD analysis. Significantly positive genetic correlations were found between PRS_{CAD} and CAD (β=0.552, p-value<0.000), between PRS_{SBP} and CAD (β=0.741, p-value<0.000), and between PRS_{TC} and CAD (β=0.361, p-value<0.005). In addition, Pearson correlation test was done among 9 different PRS calculated through LDpred-auto and excluded one of the variables which have the correlation coefficient larger than 0.7. Figure S4 shows the result plot from the test and Figure S5 shows the results with p-value<0.05.

As a result, we have excluded PRS for DBP because the correlation with SBP was 0.728. Furthermore, the results for correlation test among clinical variables and PRSs are in Figure S6.

Table 4. Simple logistic regression model for each PRS

Logistic regression model		
Y(CAD)	**PRS_i only**	
PRS_i	**beta**	**p-value**
PRS.CAD	0.552	0.000
PRS.BMI	0.161	0.081
PRS.T2D	-0.099	0.014
PRS.SBP	0.741	0.000
PRS.ENS	0.846	0.165
PRS.TG	-0.021	0.797
PRS.TC	0.361	0.004
PRS.HDL	-0.621	0.35

Results of penalized regression for meta-PRS construction

We calculated two meta-PRS from the results of Ridge regression and elastic net regression. As a result of 10-fold CV elastic-net regression, minimum lambda which had the highest AUC was chosen for the ridge regression. 8 PRSs (PRS_{CAD}, PRS_{BMI}, PRS_{SBP}, PRS_{ENS}, PRS_{T2D}, PRS_{TC}, PRS_{TG}, PRS_{HDL}) were used for the construction of $metaPRS_8$ in the ridge regression, and 3 PRSs (PRS_{CAD}, PRS_{SBP}, PRS_{T2D}) were chosen for the construction of $metaPRS_3$ as a result of elastic-net regression.

The effect sizes of each PRS are in Table 5 and Table 6. The same effect size for each PRS was used for the construction of PRS in the test data. For unmatching SNPs from the train data in the test data(~10,000), we have substituted the risk alleles as the expected value (2*MAF of each SNP) in the KoGES data.

Table 5. Result of Ridge regression model with all 8 PRS

Y(CAD) ~ Age + sex + PC1~10 + 8 PRS		
PRS for each trait	**beta (weight for meta-PRS)**	**p-value**
PRS.CAD	0.241	0.000
PRS.BMI	0.021	0.439
PRS.T2D	-0.040	0.249
PRS.SBP	0.091	0.000
PRS.ENS	0.034	0.278
PRS.TG	-0.027	0.255
PRS.TC	0.043	0.082
PRS.HDL	-0.019	0.463

Each PRS is standardized with its mean and std. regression model is adjusted with age, sex, and pc scores.

Table 6. Result of 10-fold CV elastic-net regression model

PRS for each trait	**beta (weight for meta-PRS)**
PRS.CAD	0.2216
PRS.BMI	0.0000
PRS.T2D	-0.0002
PRS.SBP	0.0565
PRS.ENS	0
PRS.TG	0
PRS.TC	0
PRS.HDL	0

Each PRS is standardized with its mean and std. Regression model is adjusted with age, sex, and pc scores. (Lambda=0.001186663)

4. Integrated model using PRS and clinical variables

As a result of comparing the prediction accuracy between the traditional model with clinical variables and the integrated model with an additional genetic effect (Figure 3), the model with meta-PRS and clinical variables showed the best performance (AUC: 0.785). Though the simple model with meta-PRS showed better performance than the simple model with PRS_{CAD} (M2 AUC: 0.733, M4 AUC: 0.735, Delong test p-value<0.05), the integrated model with PRS_{CAD} showed comparable performance(AUC: 0.784, Delong-test p-value>0.05) with the integrated model with meta-PRS. Therefore, both integrated models with PRS_{CAD} and meta-PRS were selected as the final model. The comparison of the model performance with PRS_{CAD}, $metaPRS_{3,}$ and $metaPRS_{8}$ in KoGES is summarized in Table 7, Table 8, and Table 9.

In detail, the final model with PRS_{CAD} (M6) showed a significant effect on predicting CAD in KoGES data. Figure S7 is a forest plot that represents the Odds Ratio(OR) and the 95% Confidence Interval(95% CI)of each variable used in the logistic regression. Type 2 Diabetes(T2D) and Smoking Behavior(more than 400 cigarettes) also showed significant OR larger than 1 in predicting CAD(OR: 1.46, 95%CI: [1.28, 1.66], p-value<0.000, and OR: 1.23, 95%CI: [1.08, 1.41], p-value=0.005, respectively). Moreover, BMI has OR larger than 1 (OR: 1.37, 95%CI: [1.30, 1.43], p-value<0.000). In addition, females had less likelihood of CAD outbreak than males(OR: 0.85, 95%CI: [0.75, 0.98], p-value<0.05), and AGE has OR slightly larger than 1 (OR: 1.09, 95%CI: [1.08,1.10], p-value<0.000). Finally, Figure 4 shows the ROC curve for M1, M2, M4, and M5.

Further, the final model was tested in the GENIE data for validation, AUC: 0.732 and 0.731, respectively PRS_{CAD} and meta-PRS (Table 10).

Unlike in the KoGES data, the increment due to PRS was not statistically significant in the test data.

Table 7. Model Comparison(AUC) of PRS_{CAD} and $metaPRS_8$ (ridge)

	Model AUC		De Long test p-value
M2' vs M1	0.576	0.724	0.000
M4' vs M1	0.580	0.724	0.000
M1 vs M2	0.724	0.733	0.000
M1 vs M4	0.724	0.735	0.000
M2 vs M3	0.733	0.734	0.446
M2 vs M4	0.733	0.735	0.048
M4 vs M5	0.735	0.785	0.000

Table 8. Model Comparison(AUC) of PRS_{CAD} and $metaPRS_3$ (elastic-net)

	Model AUC		De Long test p-value
M2' vs M1	0.579	0.724	0.000
M1 vs M2	0.724	0.733	0.000
M2 vs M3	0.733	0.734	0.446
M2 vs M4	0.733	0.735	0.014
M2 vs M5	0.733	0.784	0.000
M4 vs M5	0.735	0.785	0.000

Table 9. Final Model CAD ~ optimal PRS + age + sex + BMI + T2D + SBP + TC + TG + HDL

	M6 (PRS_{CAD})			M5 (meta-PRS)		
$metaPRS_i$	**p-value**	**OR**	**AUC**	**p-value**	**OR**	**AUC**
$metaPRS_3$	0.000	1.323	0.784	0.000	1.340	0.785
$metaPRS_8$	0.000	1.321	0.784	0.000	1.348	0.785

Table 10. AUC of the final models in the test data(GENIE)

			Model AUC	
M1	**vs**	**M2**	0.7047	0.7052
MC	**vs**	**M5**	0.7310	0.7318
MC	**vs**	**M6**	0.7310	0.7311

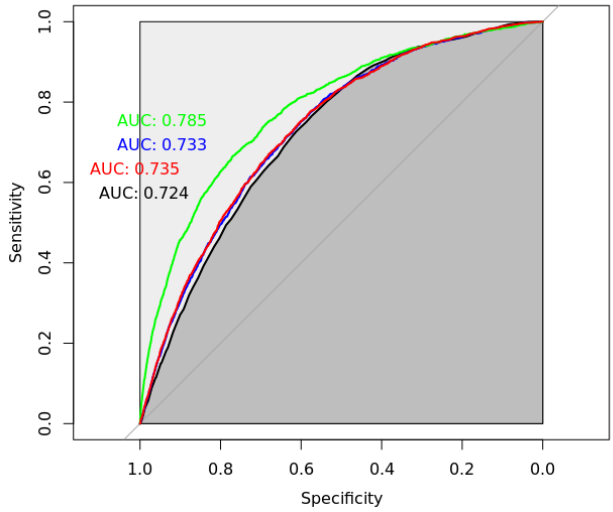

Figure 4. ROC curve for M1, M2, M4, and M5 in KoGES M1(CAD ~ age + sex) had AUC of 0.724, M2(CAD ~ PRS_{CAD} + age + sex) had AUC of 0.735, M4(CAD ~ $metaPRS_8$+ age + sex) had AUC of 0.733, and M5(CAD ~ $metaPRS_8$ + age + sex + BMI + T2D + SBP + TC + TG + HDL) showed the highest AUC of 0.785. Note: All continuous variables were scaled using their SD.

5. Classification of CAD risks

The AUC of the integrated model and p-value of PRS in three risk groups(high, medium, and low) are in Table 11. Instead of only comparing AUC in different risk groups, we also compared NRI in the total study population (KoGES + GENIE). NRI of the integrated model with PRS_{CAD} compared to the traditional model was 0.072±0.0127 and NRI of the integrated model with meta-PRS was 0.088±0.0135. Both were positive values which means PRS helped classify CAD cases into a high-risk group. Between the two models, NRI is 0.0179 which means adding meta-PRS had a more effect on classification than $PRS_{CAD.}$

Table 11. Results of risk classification of PRS in KoGES

	PRS risk-group	AUC	OR of PRS	p-value
PRS_{CAD}	low	0.750	1.184	0.883
	medium	0.777	1.190	0.320
	high	0.805	1.625	0.000
$metaPRS_8$	low	0.751	1.017	0.137
	medium	0.777	1.101	0.073
	high	0.807	1.381	0.000

Ⅳ. Discussion

This study supports that PRS as a genotypic effect can be used as a useful tool to predict CAD outbreaks in accordance with traditional clinical variables such as hypertension, obesity, diabetes, dyslipidemia, smoking status, etc. Furthermore, we were able to see the difference in the effect of integrated risks on CAD by risk group according to the level of PRS score. Through this study, it was found that the PRS score calculated by the LDpred-auto method had a statistically significant improvement in predictive ability compared to the model with only existing clinical variables in KoGES.

The major strengths of the current study include our PRS and meta-PRS included most of the variants that underlay CAD risk capturing the full spectrum of genomic variants. Here, to derive a PRS for CAD, we used the current large GWAS not only of CAD but also of CAD-related traits from BBJ. Finally, it was the first attempt to construct meta-PRS in the Korean population, which enabled us to comprehensively evaluate the combination of polygenic risk and traditional clinical risk. Several studies examined whether the genome-wide CAD PRSs improved risk prediction beyond the PCE in European ancestry populations[25]. Using a risk threshold of 7.5%, the addition of the polygenic risk score to pooled cohort equations resulted in an overall net reclassification improvement of 4.0% [95% CI, 3.1% to 4.9%]) in the UK biobank. In our study, adding the PRS to the traditional model yielded an increment of about 0.005 in AUC and an NRI of 0.0772 at a risk threshold of 2% in the KoGES data. We also demonstrated that the meta-PRS provided statistically significant yet modest discrimination. In our analysis, we observed a comparable

level of risk (OR 1.32, 95% CI 1.26-1.39), supporting that the meta-PRS may serve as a risk-enhancing factor for CAD.

However, the general clinical utility of PRS and meta-PRS in CAD risk reclassification was uncertain in the test data, which contained OR of 1 in its 95%CI. This infers that there might have been several limitations in our study. The main reason might be originated from the discrepancies among three different groups used for PRS calculation, BBJ(Japanese origin), train data (KoGES, Korean origin), and test data (GENIE, Korean origin). Therefore, in further studies, meta-GWAS of the East Asian population for the reference data can be used to calculate PRS.

The result implies that M2 which only has age and sex as its explanatory variables has a pretty high AUC compared to other models with more variables. As we can see here, age is the most important risk driver in the risk prediction model. The effect of age might has resulted in overestimation or underestimation of risk of CAD, whereas genetic risk is age-independent and can be determined early in life. Our findings highlight the concept that PRS may provide complementary information to better stratify CAD risks and inform clinical decision-making for primary prevention. Further research on the models without age will be done to support this idea.

In this model, PRS for CAD had a high odds ratio compared to other well-known clinical variables. However, PRSs for other variables did not show significant OR for CAD outcomes. In particular, all three phenotypes related to dyslipidemia did not show consistent effect signs in different combinations of variables.

There might be possible explanations for this. First, this might be due to the high percentage of medication for hyperlipidemia in the Korean population. Causal association between low-density lipoprotein-cholesterol (LDL-C) and Ischemic Heart Disease(IHD) was observed in the previous Mendelian Randomization analysis, but high-density lipoprotein-cholesterol (HDL-C) and triglyceride (TG) did not show causal association with IHD[26]. However, a direct measure of LDL-C was not obtained in some subgroups of KoGES, and Friedewald formula[27] to calculate LDL-C from TG, TC, and HDL-C is known to be inaccurate in case of high triglyceride (>400 mg/dL). Another explanation might be that dyslipidemia has a long-term pathology throughout one's lifetime[28]. Hence, considering the time effect of each variable might have been crucial in this type of follow-up cohort data.

Secondly, in the case of smoking behavior, PRS_{ENS} showed consistently significant and high OR as the phenotype of smoking behavior itself also showed the same direction of effect size. However, this effect might be overstated due to the misclassification followed by different definitions of current smokers in BBJ and three KoGES cohorts. Some past smokers who have smoked more than 400 cigarettes might be classified as current smokers or current smokers who have not smoked more than 400 cigarettes yet as non-smokers.

Overall, it has been demonstrated that the high genetic risk of CAD may be mitigated by statin use and healthy lifestyle in both primary and secondary prevention and that individuals at high genetic risk were found to derive the greatest benefit from the therapeutic intervention[29]. The randomized controlled trials focusing on individuals at intermediate or high clinical risk, especially Korean, are required to

confirm the clinically meaningful benefit and the cost-effectiveness of polygenic risk stratification for CAD.

Additionally, some other limitations should also be noted. First of all, though the sample size for the PRS calculation was large enough, the sample size of the test was not large. Next, baseline phenotyping according to a well-defined and standardized protocol was lacking due to different definitions of some traits. Accordingly, a more complex consideration of the relationship between clinical variables was unavailable, possibly resulting in the inconsistent effect of well-known risk factors on CAD.

In further studies, we would like to supplement the integrated model by using the Cox survival regression method. Since both train and test data used for this study are composed of follow-up cohort data, further study can examine the risk of development over time. This way, we can also consider the time effect of disease outbreak with respect to the genotype data and utilized the follow-up data for a more precise phenotype of the sample population.

Ⅴ. Conclusion

This study has shown that adding PRS to the traditional prediction model has a significant impact on improving prediction accuracy. Moreover, the predictive performance of meta-PRS considering the genetic effects of various traits and PRS scores calculated by the existing LDpred method was compared through logistic regression. In addition, CAD risk group classification according to PRS calculated by each method was conducted. However, in this study, the validation was not successful to show the additive effect on the performance of PRS in test data. This might be because the two population have different features. In particular, calculating PRS is affected by the characteristic of the population such as MAF, so considering a more general feature of the Korean population is needed.

On top of that, we assessed the various models to KoGES data consisting of all Koreans. The optimal PRS for CAD risk is calculated, and the genetic effect in CAD is investigated by using it as a risk predictor. In addition, a disease prediction model including clinical variables related to the disease is created to confirm the acquired effects of clinical variables. Finally, we have presented a CAD prediction model using these variables to model a more accurate prediction system for CAD risk.

The expected effect of this study is to compare the performance of the CAD prediction model using only existing clinical variables and the PRS-added model to confirm the clinical usefulness of PRS in CAD prediction in the East Asian population.

Reference

1. Vinkhuyzen AA, Wray NR, Yang J, Goddard ME, Visscher PM. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet*. 2013;47:75-95. doi:10.1146/annurev-genet-111212-133258

2. Riveros-Mckay F, Weale ME, Moore R, et al. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ Genom Precis Med*. Apr 2021;14(2):e003304. doi:10.1161/CIRCGEN.120.003304

3. Said MA, Verweij N, van der Harst P. Associations of Combined Genetic and Lifestyle Risks With Incident Cardiovascular Disease and Diabetes in the UK Biobank Study. *JAMA Cardiology*. 2018;3(8):693-702. doi:10.1001/jamacardio.2018.1717

4. Hajar R. Risk Factors for Coronary Artery Disease: Historical Perspectives. *Heart Views*. Jul-Sep 2017;18(3):109-114. doi:10.4103/heartviews.Heartviews_106_17

5. Song Y, Choi JE, Kwon YJ, et al. Identification of susceptibility loci for cardiovascular disease in adults with hypertension, diabetes, and dyslipidemia. *J Transl Med*. Feb 25 2021;19(1):85. doi:10.1186/s12967-021-02751-3

6. Li M, Chen Y, Yao J, et al. Genome-Wide Association Study of Smoking Behavior Traits in a Chinese Han Population. *Front Psychiatry*. 2020;11:564239. doi:10.3389/fpsyt.2020.564239

7. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol*. Oct 16 2018;72(16):1883-1893. doi:10.1016/j.jacc.2018.07.079

8. Lu X, Liu Z, Cui Q, et al. A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study. *Eur Heart J*. May 7 2022;43(18):1702-1711. doi:10.1093/eurheartj/ehac093

9. Agbaedeng TA, Noubiap JJ, Mofo Mato EP, et al. Polygenic risk score and coronary artery disease: A meta-analysis of 979,286 participant data. *Atherosclerosis*. Sep 2021;333:48-55. doi:10.1016/j.atherosclerosis.2021.08.020

10. Dikilitas O, Schaid DJ, Kosel ML, et al. Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. *Am J Hum Genet*. May 7 2020;106(5):707-716. doi:10.1016/j.ajhg.2020.04.002

11. Kim Y, Han BG. Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium. *Int J Epidemiol*. Apr 1 2017;46(2):e20. doi:10.1093/ije/dyv316

12. Lee C, Choe EK, Choi JM, et al. Health and Prevention Enhancement (H-PEACE): a retrospective, population-based cohort study conducted at the Seoul National University Hospital Gangnam Center, Korea. *BMJ Open*. Apr 19 2018;8(4):e019327. doi:10.1136/bmjopen-2017-019327

13. Moon S, Kim YJ, Han S, et al. The Korea Biobank Array: Design and Identification of Coding Variants Associated with Blood Biochemical Traits. *Sci Rep*. Feb 4 2019;9(1):1382. doi:10.1038/s41598-018-37832-9

14. Seo S, Park K, Lee JJ, Choi KY, Lee KH, Won S. SNP genotype calling and quality control for multi-batch-based studies. *Genes Genomics*. Aug 2019;41(8):927-939. doi:10.1007/s13258-019-00827-5

15. Yoo S-K, Kim C-U, Kim HL, et al. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Medicine*. 2019/10/22 2019;11(1):64. doi:10.1186/s13073-019-0677-z

16. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. Sep 2007;81(3):559-75. doi:10.1086/519795

17. Song YE, Lee S, Park K, Elston RC, Yang HJ, Won S. ONETOOL for the analysis of family-based big data. *Bioinformatics*. Aug 15 2018;34(16):2851-2853. doi:10.1093/bioinformatics/bty180

18. Ye Y, Chen X, Han J, Jiang W, Natarajan P, Zhao H. Interactions Between Enhanced Polygenic Risk Scores and Lifestyle for Cardiovascular Disease, Diabetes, and Lipid Levels. *Circ Genom Precis Med*. Feb 2021;14(1):e003128. doi:10.1161/circgen.120.003128

19. Vilhjalmsson BJ, Yang J, Finucane HK, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet*. Oct 1 2015;97(4):576-92. doi:10.1016/j.ajhg.2015.09.001

20. Privé F, Vilhjálmsson BJ, Aschard H, Blum MGB. Making the Most of Clumping and Thresholding for Polygenic Scores. *The American Journal of Human Genetics*. 2019/12/05/ 2019;105(6):1213-1221. doi:https://doi.org/10.1016/j.ajhg.2019.11.001

21. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol*. Sep 2017;41(6):469-480. doi:10.1002/gepi.22050

22. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. Apr 16 2019;10(1):1776. doi:10.1038/s41467-019-09718-5

23. Zou H, 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.

24. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. Jan 30 2008;27(2):157-72; discussion 207-12. doi:10.1002/sim.2929

25. Elliott J, Bodinier B, Bond TA, et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA*. Feb 18 2020;323(7):636-645. doi:10.1001/jama.2019.22241

26. Lee SH, Lee JY, Kim GH, et al. Two-Sample Mendelian Randomization Study of Lipid levels and Ischemic Heart Disease. *Korean Circ J*. Oct 2020;50(10):940-948. doi:10.4070/kcj.2020.0131

27. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem*. Jun 1972;18(6):499-502.

28. Brunner FJ, Waldeyer C, Ojeda F, et al. Application of non-HDL cholesterol for population-based cardiovascular risk stratification: results from the Multinational Cardiovascular Risk Consortium. *Lancet*. Dec 14 2019;394(10215):2173-2183. doi:10.1016/S0140-6736(19)32519-X

29. Whayne TF, Jr., Saha SP. Genetic Risk, Adherence to a Healthy Lifestyle, and Ischemic Heart Disease. *Curr Cardiol Rep*. Jan 10 2019;21(1):1. doi:10.1007/s11886-019-1086-z

30. Koyama S, Ito K, Terao C, et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat Genet*. Nov 2020;52(11):1169-1177. doi:10.1038/s41588-020-0705-3

31. Kanai M, Akiyama M, Takahashi A, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat Genet*. Mar 2018;50(3):390-400. doi:10.1038/s41588-018-0047-6

32. Suzuki K, Akiyama M, Ishigaki K, et al. Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat Genet*. Mar 2019;51(3):379-386. doi:10.1038/s41588-018-0332-4

33. Akiyama M, Okada Y, Kanai M, et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat Genet*. Oct 2017;49(10):1458-1467. doi:10.1038/ng.3951

34. Kanai M, Ulirsch J, Karjalainen J, et al. *Insights from complex trait fine-mapping across diverse populations*. 2021.

Supplementary Figures and Tables

Table S1. Sources of BBJ summary statistics used for PRS calculation
Sources of summary statistics used for each trait-specific PRS construction and matched SNPs of KoGES used for PRS LDpred-auto calculation

Trait	Source	Types	Ancestry	Sample size (case / control)	Method	Reference	Common SNPs (KoGES)
CAD	BBJ (Riken)	GWAS	Japanese	212,453 (29,319/183,134)	Meta-analysis	Koyama et al.[30]	510,458
SBP	BBJ	GWAS	Japanese	145,505	GWAS	Kanai et al.[31]	510,462
DBP	BBJ	GWAS	Japanese	145,515	GWAS	Kanai et al.[31]	510,462
T2D	BBJ	GWAS	Japanese	177,415 (45,383/132,032)	Meta-analysis	Suzuki et al.[32]	526,674
BMI	BBJ	GWAS	Japanese	163,835	GWAS	Akiyama et al.[33]	510,562
TC	BBJ	GWAS	Japanese	135,808	GWAS	Kanai et al.[31]	513,041
TG	BBJ	GWAS	Japanese	111,667	GWAS	Kanai et al.[31]	513,041
HDL-C	BBJ	GWAS	Japanese	74,970	GWAS	Kanai et al.[31]	513,041
Smoking (ENS)	BBJ	GWAS	Japanese	88,277	GWAS	Malik et al.[34]	510,462

PRS, polygenic risk score; GWAS, genome-wide association study; BP, blood pressure; SBP, systolic BP; DBP, diastolic BP; T2D, type 2 diabetes; BMI, body mass index; HDL-C, high density lipoprotein cholesterol; TC, total cholesterol; TG, triglyceride; ENS, ever-never smoked; Common SNPs are used for PRS calculation.

Table S2. Descriptive characteristics of GENIE

Characteristics (N=3,591)	***Mean* (*SD*) or *n (%)***
Demographic data	
Age	46.26 (10.43)
Sex	
Female	1526 (42.5)
Male	2065 (57.5)
Disease history	
CAD	108 (3.0)
T2D	119 (3.3)
ENS	1357 (37.8)
Anthropometric data	
BMI	23.15 (3.15)
SBP	115.35(13.34)
Blood Lipid levels	
TC	193.47 (31.46)
TG	109.17 (75.44)
HDL-C	53.34 (11.33)

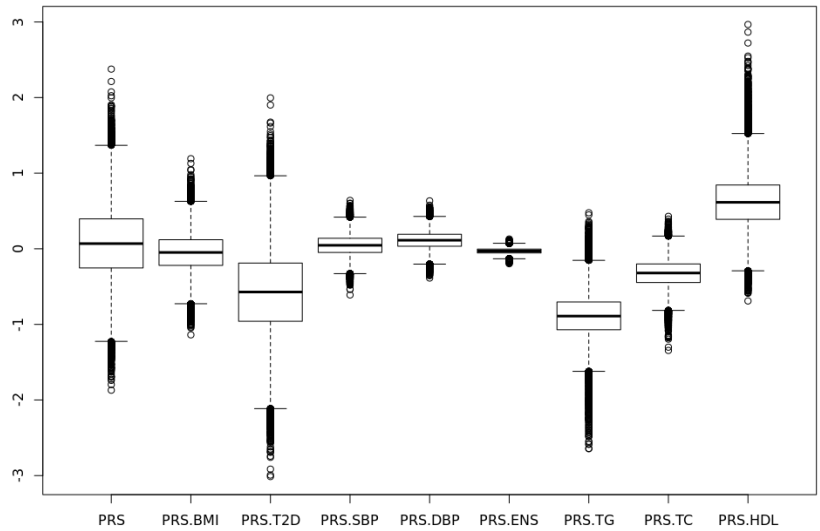

Figure S1. Boxplot for PRS of 9 traits in KoGES

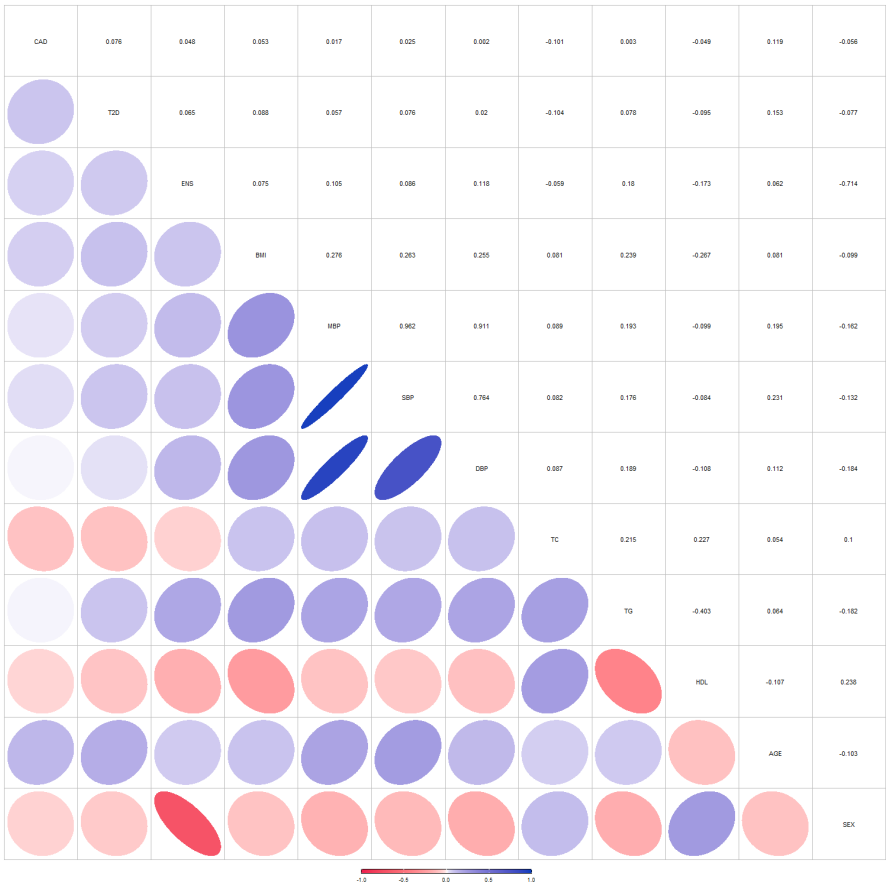

Figure S2. Correlation Plot of 11 variables in KoGES This figure represents the sign and size of the correlation coefficient calculated through Pearson's correlation test. Since MBP, SBP, and DBP have high correlation coefficients and the sign is positive, we only included SBP in our model which is clinically more important and useful. Also, SEX and ENS have a correlation of -0.714; however, since both variables are crucial in their relationship with CAD in previous studies, we included both variables in our model.

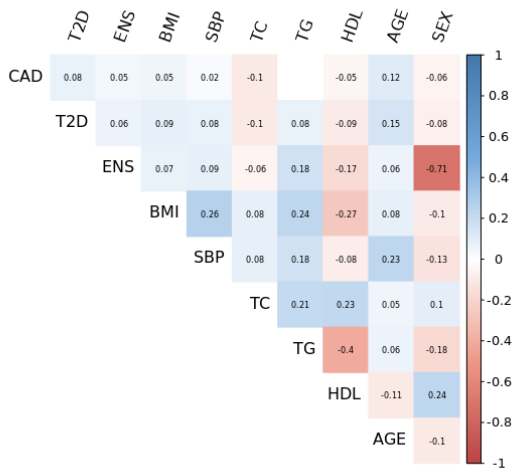

Figure S3. Pearson correlation coefficients(p-value<0.05) of clinical variables in KoGES

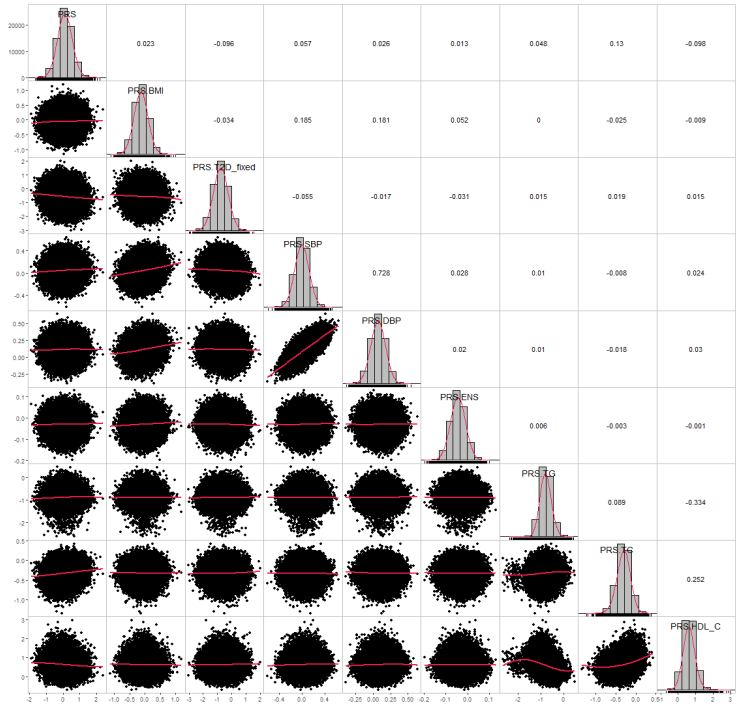

Figure S4. Correlation plot of trait-specific PRSs in KoGES Correlation coefficients and p-values were estimated from the Pearson correlation test for each pair of PRSs. Likely, SBP and DBP have a correlation coefficient of 0.726. PRS, polygenic risk score; CAD, coronary artery disease; BP, blood pressure; BMI, body mass index; T2D, type 2 diabetes; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides.

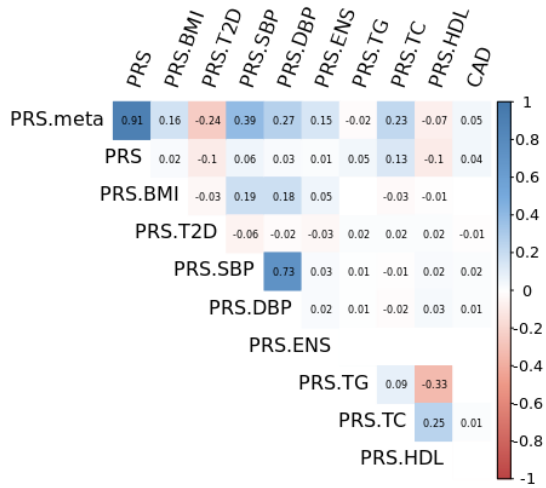

Figure S5. Pearson correlation coefficients(p-value<0.05) of PRSs in KoGES PRSs include 9 PRS, and metaPRS$_8$

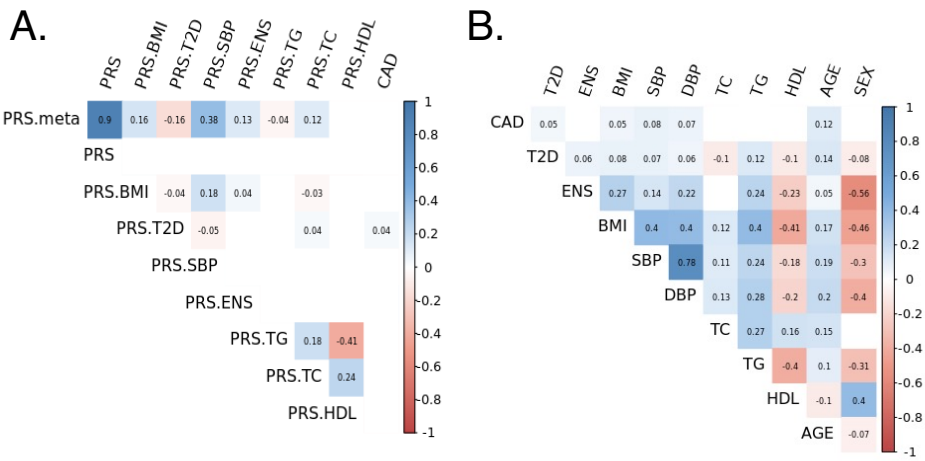

Figure S6. Pearson correlation test (A) Correlation test of 8 PRS, metaPRS$_8$, and CAD in GENIE (B) Correlation test of clinical variables in GENIE

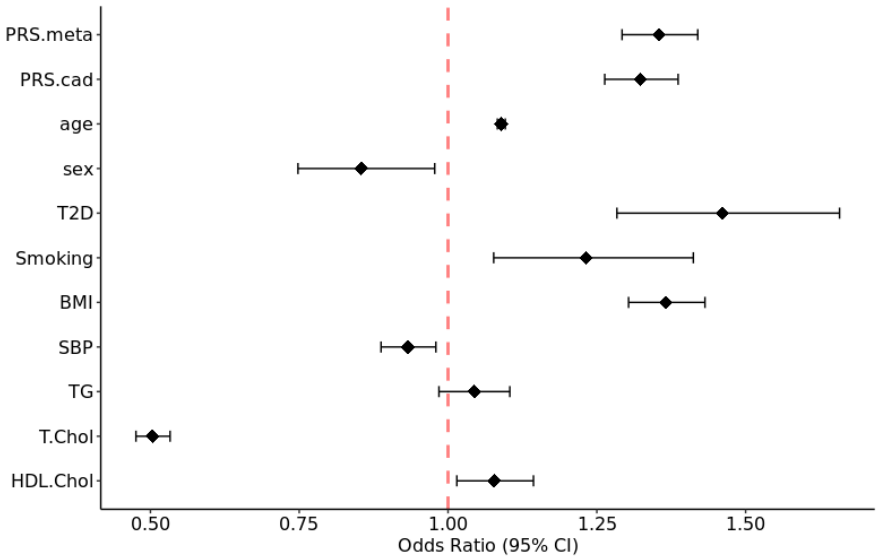

Figure S7. Forest Plot for Integrated Model (metaPRS$_8$ and PRS$_{CAD}$) in KoGES The forest plot represents the Odds Ratio and the 95% Confidence Interval of each component used in the logistic regression model.

국문초록

동아시아 인구집단에서의 다유전자 위험점수와 임상변수를 이용한 관상동맥질환 위험예측

정유리

서울대학교 보건대학원

보건학과 보건학전공

연구배경: 관상동맥질환(Coronary Artery Disease, CAD)은 심장에 혈액을 공급하는 관상동맥이 동맥경화증이나 혈전으로 좁아져 혈액 공급이 원활하지 않게 되어 심근허혈이나 심근경색이 발생하는 질환이다. CAD 는 유전율이 40%~60%로 높기 때문에 위험을 예측하는데 있어 유전적 요인을 고려하는 것이 중요하다. 연구자들은 이전 연구들의 다유전적 형질 질병의 성질을 고려하지 않은 한계를 극복하기 위해 메타 PRS 를 고안했으며 이는 LDpred 에 의해 계산된 단일 PRS 보다 CAD 예측에서 더 나은 성능을 보였다. 그러나 대부분의 연구는 백인 대상으로 진행되었으며, 특히 한국인을 포함한 동아시아 인구에 대한 연구는 불충분한 실정이다.

연구목표: 본 연구에서는 메타 PRS 방법을 포함한 다양한 PRS 계산 방법을 이용하여 동아시아 인구에서 CAD 에 대한 PRS 를 계산하고자 한다. 또한 PRS 와 임상 마커를 모두 고려한 통합 모델을 통해 유전적 요인과 임상적 요인을 모두 고려한 CAD 위험을 예측하고자 한다.

연구방법: 일본 바이오뱅크(BBJ)의 일본인에 대한 GWAS 결과 요약 통계량을 참고자료로 활용하여 KoGES 데이터의 한국인 71,009 명에 대한 PRS 를 계산하기 위한 각 SNP 에 대한 가중치를 계산하였다. 그리고 나서 5 가지 계산 방법을 통해 CAD 에 대한 PRS 를 계산하였으며, 제일 높은 AUC 를 보인 방법을 채택하였다. 여기에 더해, CAD 를 포함한 관련 형질 8 개를 선정하여 메타 PRS 를 산출하여 예측모형을 구축하였다.

결과: PRS 와 메타 PRS 가 높은 오즈비(OR 1.32, 95% CI 1.26-1.39), (OR 1.35, 95% CI 1.29-1.42)를 가지고 있다는 것을 보였으며, 두 가지 모두에 대한 순재분류 개선은 각각 0.072±0.0127 과 0.088±0.0135 였다. LDpred-auto 로 계산한 PRS 점수는 기존

임상변수만 있는 모형에 비해 유의미한 예측능력 향상을 보였다(AUC: 0.780~0.785, P=0.0003).

결론: 본 연구에서는 기존의 관상동맥질환 위험예측 임상 변수에 PRS를 추가한 효과를 분석함으로써 동아시아 인구집단에서 관상동맥질환 예측에 있어 유전적 효과를 확인하였다.

핵심어 : 관상동맥질환, 심혈관질환 위험 예측모형, 메타 다유전자 위험점수, 연관불균형 예측, 릿지 회귀, 엘라스틱넷 회귀

학번 : *2021-21804*