Comparative Analysis of Asthma Prediction Model using Statistical and Artificial Intelligence Algorithms in KoGES Data

Yongjun Choi[1], Junho Cha[2], Taehee Woo[1], Sungkyoung Choi[1,3*]

[1]Department of Applied Artificial intelligence, Hanyang University (ERICA), Ansan, Korea
[2]Department of Molecular Life Science, Hanyang University (ERICA), Ansan, Korea
[3]Department of Applied Mathematics, Hanyang University (ERICA), Ansan, Korea, *Corresponding author.

Abstract

- Asthma is a complex disease that affect lungs. The interaction of various genetic factors can lead to asthma in adults as well as one of the most common complex diseases in children. Predicting the occurrence of asthma diseases can prevent its onset for those high-risk groups.
- In this study, we aimed to construct prediction models for asthma using genome-wide association study (GWAS) and clinical factors.
- We performed logistic regression analysis under an additive genetic model, adjusting for age, sex, smoke, body mass index, allergy, and 10 principal components(PCs).
- Then, we compared the performance of prediction models constructed using penalized methods, and ensemble methods.
- Application of our model to asthma in the Korean Genome and Epidemiology Study (KoGES) data shows that genetic factors provide valuable information on the variation of asthma diseases and improve prediction performance.

Introduction

- A number of GWAS studies have identified Single Nucleotide polymorphisms (SNPs) associated with asthma diseases. However, predicting the occurrence of asthma with GWAS discovered problem.[1,2]
- The main goal is to build asthma prediction model and to compare the performance of model.
- In this study, we used cohort data consisting of Health examines study (HEXA), Cardiovascular disease association study (CAVAS), Korea association resource study (KARE) cohort of KoGES and used genetic data, Korean chips (KORV1.1).
- We analyzed the performance of our models according to genetic variables and demographic variables (sex, age, smoke, body mass index, allergy, and 10 PCs).
- To evaluate for the model performance, we computed the Area Under the Curve (AUC) of prediction models by cohorts (HEXA, CAVAS, and KARE cohorts) of KoGES data.

Materials & Methods

- **Quality control**
  - Quality control filters were used, including minor allele frequency (MAF) ≤ 0.05, missing genotype call ratio ≥ 0.05, removing insert-deletion (INDEL) and multiallelic SNPs, and Hardy-Weinberg Equilibrium (HWE) $P > 10^{-5}$.
  - And then, we removed missing values from epidemiological factors (age, sex, smoke, body mass index, and allergy).
  - After filtering, Total 5,416,280 SNPs remained from Korean chips for HEXA, CAVAS, and KARE cohorts.
  - Among the filtered SNPs, we selected genetic variables by top 50, 200, 1,000, 2,500 SNPs through logistic regression model.
  - Demographic variables show with Table 1. For each HEXA, CAVAS, and KARE cohorts, we extract genetic data and epidemiological data of Korean chips and KoGES data.

Table 1. Demographic variables for HEXA, CAVAS, and KARE cohorts

	HEXA		CAVAS		KARE	
	Normal	Asthma	Normal	Asthma	Normal	Asthma
# of Samples	57,459	975	2,908	95	5,308	112
Sex (Male/Female)	19,924(98.6%)/ 37,535(98.2%)	283(1.4%)/ 692(1.8%)	1,164(96.9%)/ 1,744(96.8%)	37(3.1%)/ 58(3.2%)	2,563(98.5%)/ 2,745(97.4%)	39(1.5%)/ 73(2.6%)
Age[a]	53.8±8.0	55.4±8.4	55.4±7.8	57.9±7.8	51.5±8.5	53.3±7.9
BMI	23.9±2.9	24.3±3.2	24.5±3.0	25.5±3.4	24.6±3.0	25.0±3.5
Smoke status[b] (No/Yes)	42,070(98.3%)/ 15,389(98.4%)	721(1.7%)/ 254(1.6%)	2,123(97.0%)/ 785(97.2%)	72(3.0%)/ 23(2.8%)	3,173(97.8%)/ 2,135(98.1%)	71(2.2%)/ 41(1.9%)
Allergy status (No/Yes)	53,642(98.7%)/ 3,817(93.9%)	727(1.3%)/ 248(6.1%)	2,695(97.3%)/ 213(91.0%)	74(2.7%)/ 21(9.0%)	5,015(98.3%)/ 293(91.8%)	86(1.7%)/ 26(8.2%)

BMI, body mass index
[a]**Means ± standard deviation (SD)**
[b]**Smoke status (No: never smoker / Yes: former smoker + current smoker)**

- **Penalized and ensemble methods**
  - Penalized methods improve prediction model performance by shrinkage of coefficients tuning parameters to nonzero.[3] Ensemble methods also enhance the model performance by creating multiple models and combining models to produce improved results.
  - We used Ridge, Least absolute shrinkage and selection operator (LASSO), Elastic-Net (Enet), and Smoothly clipped absolute deviation (SCAD) regression to build asthma prediction model of penalized methods.[4,5]
  - Ensemble model was constructed using Support Vector Machine (SVM), Random Forest (RF), and Boosting. [6,7,8]

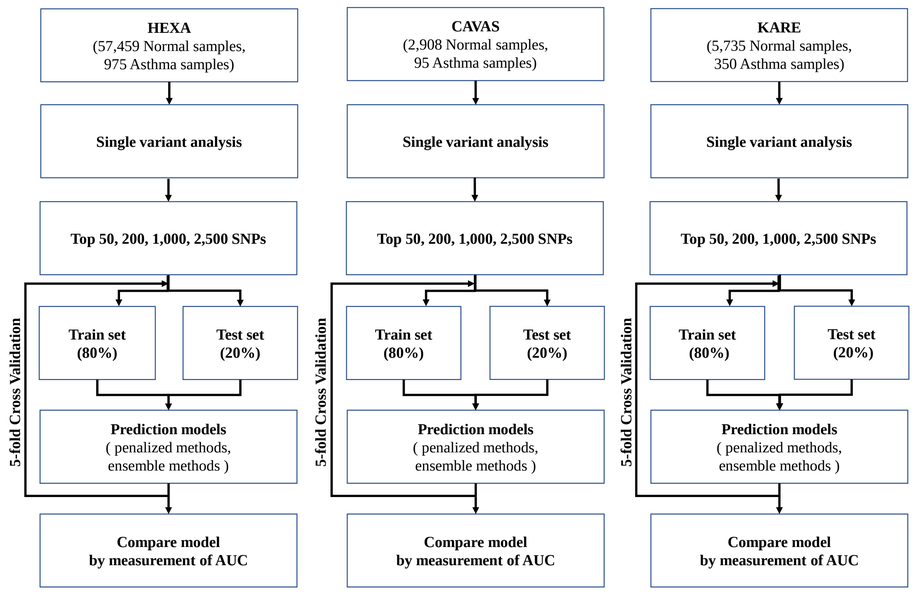

Figure 1. Flow chart of the asthma prediction model construction.

Results & Discussions

- **Results**
  - Covariates : sex, age, smoke, body mass index, allergy, and 10 PCs
  - When we compared the performance of the models, we found that Ridge, Lasso, Enet, SCAD, SVM, RF, and Boosting showed good performance in predicting asthma. it can be confirmed that the performance improved as the number of SNPs increased in Table 2.

Table 2. Comparison of test set AUC of prediction methods

Group	Penalized methods				Ensemble methods		
HEXA cohort							
	AUC by methods				AUC by methods		
# of SNPs	**Ridge**	**LASSO**	**Enet**	**SCAD**	**SVM**	**RF**	**Boosting**
50	0.70210	0.71221	**0.71298**	0.71150	0.50636	0.54502	0.68686
200	0.74295	**0.75050**	0.75041	0.74941	0.51552	0.54409	0.69780
1,000	0.87910	0.88457	**0.88530**	0.88294	0.53020	0.54989	0.68389
2,500	0.93383	**0.93662**	**0.93662**	0.93288	0.57957	0.53902	0.68287
CAVAS cohort							
	AUC by methods				AUC by methods		
# of SNPs	**Ridge**	**LASSO**	**Enet**	**SCAD**	**SVM**	**RF**	**Boosting**
50	0.81836	0.82927	0.82936	**0.83137**	0.51589	0.60974	0.78920
200	**0.94229**	0.94113	0.94156	0.94147	0.61187	0.69774	0.82785
1,000	**0.99559**	0.99204	0.99204	0.90231	0.80010	0.67667	0.81286
2,500	**0.98849**	0.98513	0.98513	0.82688	0.85290	0.73917	0.81347
KARE cohort							
	AUC by methods				AUC by methods		
# of SNPs	**Ridge**	**LASSO**	**Enet**	**SCAD**	**SVM**	**RF**	**Boosting**
50	0.77311	**0.78105**	0.77989	0.78086	0.54952	0.58292	0.71370
200	0.87613	0.88040	**0.88101**	0.88042	0.59777	0.61203	0.75143
1,000	**0.96063**	0.95975	0.95981	0.93895	0.73424	0.64931	0.72376
2,500	**0.98603**	0.97945	0.97945	0.90124	0.82988	0.63277	0.69578

- **Discussions**
  - Our result show that genetic factors provide significant information on the variation of asthma diseases. these results may indicate that the prediction performance of asthma is improved. Therefore, this study could have a positive impact on the early diagnosis of asthma diseases.
  - Furthermore, we will find and compare the performance of the asthma prediction model with a set of SNPs constructed on a p-value basis.

References

1.Shigemizu D, Abe T, Morizono T, Johnson TA, Boroevich KA, Hirakawa Y, Ninomiya T, Kiyohara Y, Kubo M, Nakamura Y *et al*: **The construction of risk prediction models using GWAS data and its application to a type 2 diabetes prospective cohort**. *PLoS One* 2014, **9**(3):e92549.

2.Bijanzadeh M, Mahesh PA, Ramachandra NB: **An understanding of the genetic basis of asthma**. *The Indian journal of medical research* 2011, **134**(2):149.

3.Won S, Choi H, Park S, Lee J, Park C, Kwon S: **Evaluation of Penalized and Nonpenalized Methods for Disease Prediction with Large-Scale Genetic Data**. *Biomed Res Int* 2015, **2015**:605891.

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

5.Fan J, Li R: **Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties**. *Journal of the American Statistical Association* 2001, **96**(456):1348-1360.

6.Opitz D, Maclin R: **Popular ensemble methods: An empirical study**. *Journal of artificial intelligence research* 1999, **11**:169-198.

7.Rokach L: **Ensemble-based classifiers**. *Artificial intelligence review* 2010, **33**(1):1-39.

8.Polikar R: **Ensemble based systems in decision making**. *IEEE Circuits and systems magazine* 2006, **6**(3):21-45.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.2018R1C1B6008277). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01343, Artificial Intelligence Center(Hanyang University ERICA)). This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (2019M3E5D3073365). This study was conducted with bioresources from National Biobank of Korea, the Korea Disease Control and Prevention Agency, Republic of Korea (KBN-2020-106).