Importance of GWAS risk loci and clinical data in predicting asthma using machine-learning approaches

Si-Qiao Liang
Guangxi Medical University First Affiliated Hospital
Jian-Xiong Long
Guangxi Medical University
Jingmin Deng (ldyyy666@163.com)
Guangxi Medical University First Affiliated Hospital https://orcid.org/0000-0003-0189-5371
Xuan Wei
Guangxi Medical University First Affiliated Hospital
Mei-Ling Yang
Guangxi Medical University
Shao-Jie Tang
Xi'an University of Posts and Telecommunications
Hua-jiao Qin
Guangxi Medical University
Jian-Peng Zhou
Guangxi Medical University
Hai-Li Li
Guangxi Medical University

Research

Keywords: Asthma, GWAS-supported loci, clinical data, machine learning

DOI: https://doi.org/10.21203/rs.3.rs-21271/v1

Abstract

Asthma is a serious immune-mediated respiratory airway disease. Its pathological processes involve genetics and the environment, but it remains unclear. To understand the risk factors of asthma, we combined genome-wide association study (GWAS) risk loci and clinical data in predicting asthma using machine-learning approaches. A case–control study with 123 asthma patients and 100 healthy controls was conducted in Zhuang population in Guangxi. GWAS risk loci were detected using polymerase chain reaction, and clinical data were collected. Machine-learning approaches (e.g., extreme gradient boosting [XGBoost], decision tree, support vector machine, and random forest algorithms) were used to identify the major factors that contributed to asthma. A total of 14 GWAS risk loci with clinical data were analyzed on the basis of 10 times of 10-fold cross-validation for all machine-learning models. Using GWAS risk loci or clinical data, the best performances were area under the curve (AUC) values of 64.3% and 71.4%, respectively. Combining GWAS risk loci and clinical data, the XGBoost established the best model with an AUC of 79.7%, indicating that the combination of genetics and clinical data can enable improved performance. We then sorted the importance of features and found that the top six risk factors for predicting asthma were rs3117098, rs7775228, family history, rs2305480, rs4833095, and body mass index. Asthma-prediction models based on GWAS risk loci and clinical data can accurately predict asthma and thus provide insights into the disease pathogenesis of asthma. Further research is required to evaluate more genetic markers and clinical data and predict asthma risk.

Background

Asthma is an immune-mediated respiratory airway disease, characterized by cough, wheezing, chest tightening, shortness of breath, and so on. The Global Initiative for Asthma (GINA) has reported that the prevalence of asthma is increasing in many countries, and approximately 339 million people are affected worldwide[1]. Over recent decades, various studies have focused on the relationship between genetics and/or environment with asthma, but its etiology remains unclear. Although the heritability estimate of asthma has reached 80%[2], the etiology cannot simply be explained by genetic factors. The mechanisms underlying environmental effects on genetics also play an important role. The methods for gene discovery in asthma start from candidate gene association studies to family-based genome-wide linkage analyses and are followed by genome-wide association studies (GWAS). GWAS of asthma have dominated in recent years, providing bias-free discovery of novel risk loci[3]. Since the first GWAS of child asthma reported in 2007, 83 papers on asthma or asthma-related traits are reported in the GWAS catalog until January 27, 2020 (https://www.ebi.ac.uk/gwas/). Among these papers, more than 1000 loci are associated with asthma or asthma-related traits. However, most of the loci are located in introns or in the intergenic regions, and few loci are located in the functional regions such as exons and non-coding regions. Fourteen loci in the functional regions are associated with asthma. Five missense variants (*CDHR3* rs6967330[4], *TLR1* rs4833095[5], *GSDMA* rs7212938[5], *GSDMB* rs2305480[6], and *GSDMA* rs3894194[6]), three regulatory region variants (*IL5* rs4143832[7], *HLA-DQB1* rs7775228[8], and *TACR1* rs7588010[9]), three upstream gene variants (*IL33* rs928413[4], *HLA-DRA* rs9268516[10], and *TSLP*

rs1837253[11]), two non-coding transcript exon variants (*NOTCH4* rs404860[8] and *BTNL2* rs3117098[8]), and one 5'UTR variant (*IL1RL1* rs3771180[11]) are reported.

Most of these loci are located in immune-related genes, such as *TLR1*, *GSDMA*, *IL5*, and *HLA-DQB1*, which are consistent with immune responses in asthma. Several loci have been replicated and associated with asthma in China (rs6967330[12] and rs928413[13]), United Kingdom (rs7212938 and rs3894194[14]), Korea (rs7212938[15]), Slovenia (rs2305480[16]), and Sweden (rs2305480 and rs3771180[17]). Furthermore, two loci (rs1837253 and rs3117098) have been replicated and associated with asthma in the Guangxi Zhuang population in our previous studies[18, 19]. This finding suggests that GWAS risk loci, particularly those located in the functional regions, play an important role in the genetics of asthma.

To date, many studies of candidate genes or environmental risk factors are conducted, which have often excluded the consideration of the gene-environment interactions on asthma. The observable features (phenotype) of asthma, such as clinical features and underlying mechanisms (endotype), are complex and represent a series of host–environment interactions that occur over different spatial scales. However, environmental exposure factors (e.g., exposure to tobacco smoke and allergies, body mass index [BMI], vitamin D levels, and air pollutants) may alter gene activity and expression without changing the underlying DNA sequence and increase the risk of asthma. In recent years, several Mendelian randomization studies have confirmed that higher BMI[20], smoking[21], and low linoleic acid[22] (which is believed to suppress immune responses) may increase asthma risk. Therefore, the risk prediction model for asthma requires a combination of genetic and environmental information.

The traditional method, which sets *P* values as the “gold standard” of statistical validity[23], cannot meet the requirements of the present multiple data types and high accuracy of risk prediction. Machine-learning approaches (e.g., extreme gradient boosting [XGBoost], decision tree [DT], support vector machine [SVM], and random forest [RF] algorithms) can improve the accuracy of risk prediction by automatically learning from experience. Combining GWAS-based genetic and environment stratification, Antonucci et al. obtained a reliably cognitive deficit-stratified model for schizophrenia[24]. Li et al. constructed a SVM model predictive of essential hypertension based on six environmental factors and three SNPs (single nucleotide polymorphism)[25].

The present study is the first report on combining GWAS risk loci and clinical data to predict asthma using machine-learning approaches. In addition, we conducted a case control study of 123 patients with asthma and 100 healthy controls and detected 14 GWAS risk loci located in the functional regions. We collected clinical data, which were easy to collect from the records of patients with asthma. Machine-learning approaches were used to build an asthma risk prediction model by combining GWAS risk loci and clinical data.

Subjects And Methods

Subject ascertainment and collection of clinical data

A total of 123 patients with asthma and 100 healthy controls were recruited from the first affiliated hospital of Guangxi Medical University in China, from February 2010 to August 2016. All subjects were from Zhuang population and permanently residing in Guangxi Province, China. Patients with asthma were diagnosed by at least two respiratory physicians in accordance with the guidelines of GINA. The controls were also evaluated by respiratory physicians using a self-report questionnaire, which included general conditions and medical history, but excluded those who had a history of lung diseases, asthma, rhinitis, or other allergic diseases. Clinical data, including gender, age, height, weight, family history, exposure to tobacco smoke, and allergies, were collected from electronic medical records. The same information was also collected in controls from the health examination center. The present study was approved by the Institutional Ethics Committee of Guangxi Medical University (Approval Number: 2013-KY-GuiKe-053), according to the Helsinki Declaration of Human Rights. All subjects were informed and wrote an informed consent.

DNA isolation and SNP genotyping

DNA was extracted from peripheral venous blood (2 mL) of each subject using a DNA extraction kit (provided by Tiangen, Shanghai, China) following the instructions. Primer design (Primer 3 Online), synthesis, and SNP genotyping were conducted by Shanghai BioWing Applied Biotechnology Company (http://www.biowing.com.cn/). The primer sequences of the 14 GWAS risk loci are listed in Table 1. All these SNPs were genotyped using the polymerase chain reaction (PCR)/ligase detection reaction assay. Multiplex PCR method was used for the amplification of target DNA sequences. The reaction conditions included 2 min of initial denaturation at 95 °C, then 40 cycles of denaturation at 94 °C for 30 s, 90 s of annealing at 53 °C, 30 s of extension at 65 °C, and a final 10 min of extension at 65 °C. To determine whether the reaction was successful, 2 μL of each product was run in a 3.0% agarose gel.

The ligation reaction for each subject was conducted in a total volume of 10 μL, including 1× NEB Taq DNA ligase buffer 1 μL, 2 pmol/μl of each probe mix 1 μL, Taq DNA ligase 0.05 μL, ddH2O 4 μL, and multi-PCR product 4 μL. The ligase detection reaction was performed at 95 °C for 2 min, then 40 cycles at 94 °C for 15 s, and at 50 °C for 20 s. The fluorescent ligase detection reaction product was characterized by the sequencer PRISM 3730 (ABI). Approximately 5% of the DNA samples were added into the total samples under blind conditions to assess the quality of SNP genotyping. The concordance rate was 100%.

Data preprocessing and machine-learning approaches

The procedure for preprocessing data was as follows:

(1) For the genotype of GWAS risk loci, wild type is set to “0”; heterozygous is set to “1”, and homozygous is set to “2”.

(2) For BMI categories, light weight (BMI<18.5) is set to “0”; normal weight (18.5≤BMI<24.0) is set to “1”, and overweight (BMI≥24.0) is set to “2”

(3) For family history of asthma, exposure to tobacco smoke, and allergies, “no” is set to “0”, and “yes” is set to “1”.

Missing records were no more than 10% of all features, and we used mode data to fill in the missing records. The traditional method (chi-square test) was used to compare the difference between cases and controls with SPSS 16.0.

All modeling processes were programmed on PyCharm software using Python version 3.7.4 on an IntelR Core™ i7-9850H Central Processing Unit with 16 GB RAM @2.60GHz laptop. Using the module cross_val_score[26], evaluation metrics were calculated to evaluate the predictive power of the models, including area under the curve (AUC), receiver operator characteristic curve, accuracy score, precision score, recall score, and f1_score. We evaluated all models using 10-fold cross-validation repeated 10 times. Machine-learning approaches including XGBoost[27], DT[28], SVM[29], and RF[30] algorithms were selected as classifiers to identify the importance of features.

Results

GWAS risk loci genotype and clinical data of study subjects

A total of 123 patients with asthma (50 males and 73 females) and 100 health controls (52 males and 48 females) were included in the present study. The median age and range of the asthma group were 27.9 years and 22–67 years, respectively, and the controls were 38.8 years and 18–71 years, respectively. GWAS risk loci genotype and clinical data of cases and controls, which were used as risk features for asthma classifications, are listed in Table 2. When using the chi square test to compare the proportion of these risk features, four positive features (rs3117098, rs1837253, BMI, and family history) were significantly different between cases and controls ($P<0.05$).

Machine-learning model performance and comparison

Four different machine-learning models and five evaluation metrics were performed using python, and the results are listed in Table 3. Using GWAS risk loci or clinical data, the best performances were an AUC of 64.3% and 71.4%, respectively. Using the four positive features of the traditional method, the best performance was demonstrated with an AUC of 70.2%. When combining GWAS risk loci and clinical data, the XGBoost established the best model with an AUC of 79.7%, indicating that the combination of genetics and clinical data could obtain a better performance.

Important features of asthma prediction

The XGBoost models, which showed the best predictive performances, were selected as the final predictors. To capture the informative risk features of asthma, splitting node algorithm was used, which showed the number of each important splitting node (feature) in trees. A high F score indicated that the corresponding feature was important. As listed in Figure 1, the top six risk factors in predicting asthma were rs3117098, rs7775228, family history, rs2305480, rs4833095, and BMI.

Discussions

In general, to collect all environmental exposure factors and detect all genetic information for asthma are difficult. As such, using the limited genetic and environmental exposure information to predict the risk of asthma is important. In the present study, we applied machine-learning approaches by combining GWAS risk loci and clinical data to build an accurate classifier for the prediction of asthma. The results showed that XGBoost established the best model with an AUC of 79.7% in predicting asthma, wherein rs3117098, rs7775228, family history, rs2305480, rs4833095, and BMI were the top six risk factors in this model.

A previous study selected SNPs from openSNP database and used machine-learning approaches to predict asthma. The results showed that asthma can be predicted with an AUC of 0.62 and 0.64 for RF-SVM and RF-K-nearest neighbor models, respectively[31]. Similar performance metric with an AUC of 64.3% was obtained in our study, when only GWAS risk loci were adopted as risk factors. Family-based studies showed that asthma was a heritable disease with heritability estimates of approximately 50%–60%[32]. Identified genetic variants associated with asthma in large-scale GWAS studies only accounted for a low fraction[3]. Introducing clinical data as environmental exposure factors was necessary to improve the accuracy of asthma-prediction models. AlSaad et al. used a real electronic health record dataset comprising 6159 asthma cases and 4912 controls to predict the risk of asthma and obtained the highest AUC of 0.831[33]. In contrary to using GWAS risk loci or clinical data alone, we combined GWAS risk loci and clinical data and obtained a more accurate asthma-prediction models with an AUC of 79.7%. The accuracy of our study was lower than that of AlSaad et al. [33] probably because the sample size of our study was relatively small, and we selected fewer clinical data. Collecting the clinical data for control samples was difficult, particularly for the clinical indicators which were not checked. Therefore, larger, more comprehensive data collection and better design research were required to verify our results.

Meanwhile, we found that the top six risk factors in predicting asthma were rs3117098, rs7775228, family history, rs2305480, rs4833095, and BMI. However, in contrary to the results obtained by traditional methods, several loci without remarkable differences appeared in the top risk factors based on XGBoost modeling. XGBoost considered the interactions of risk factors, whereas the traditional method only performed direct genotype–phenotype association testing. XGBoost method was evidently more efficient than traditional methods and can use more information for the construction of a classification model. Boosting is a popular ensemble technique in which new models are added to adjust the errors made by the prior models. Models were added recursively until no remarkable improvements can be observed. Gradient boosting is an algorithm in which new models are created for predicting the residuals of previous models and then combined for the final prediction. When adding new models, a gradient descent algorithm was used to minimize the loss. The XGBoost model was widely used for diagnosis classification[34, 35], treatment effect[36, 37], and prognosis evaluation[38, 39] in different diseases.

Several important limitations were found in the present study. First, pulmonary function and clinical indicators were particularly related to asthma development. Clinical data in controls were not included in our clinical feature set because of the lack of such data. Further research should focus on gathering more

clinical data to improve the diagnostic value with asthma. Second, our work included only Zhuang population from a single center in China. The prediction models might not be suitable for other ethnicities or districts. Third, although we used 10-fold cross-validation for processing, the small sample size still affected the accuracy and stability of the model.

Conclusions

Our study combined GWAS risk loci and clinical data for the first time to construct asthma-prediction models. We have obtained an asthma-prediction model with a higher accuracy based on the XGBoost method, which may provide insights into the pathogenesis of asthma. Further study is required to evaluate more genetic markers and clinical data and predict asthma risk.

Abbreviations

GWAS: genome-wide association study

XGBoost: extreme gradient boosting

AUC: area under the curve

GINA: Global Initiative for Asthma

BMI: body mass index

DT: decision tree

SVM: support vector machine

RF: random forest

SNP: single nucleotide polymorphism

PCR: polymerase chain reaction

Declarations

Ethics approval and consent to participate

The present study was approved by the Institutional Ethics Committee of Guangxi Medical University (Approval Number: 2013-KY-GuiKe-053). All subjects were informed and wrote an informed consent.

Consent for publication

Not applicable.

Availability of data and material

The data for the study are available from the corresponding authors on reasonable request.

Funding

This work is supported by the Guangxi Natural Science Foundation (Grant no. 2017GXNSFAA198104), Young and Middle Teachers Basic Capacity Improvement Project of Guangxi higher education institutions (Grant no. 2017KY0101 and Grant no. 2018KY0137). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the article.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank Miss. Deeya Patel for helping us to read the full manuscript and modify grammar, who comes from Tulane University, New Orleans, LA, USA, Email: dpatel9@tulane.edu

Author Contributions

Jing-Min Deng proposed the project and provided supervision. Si-Qiao Liang and Jian-xiong Long are responsible for project implementation, data collection, and manuscript drafting. Xuan Wei and Mei-Ling Yang provided clinical guidance and performed the manuscript editing. Shao-Jie Tang provided part of statistical analysis and revised the manuscript. Hua-jiao Qin, Jian-Peng Zhou, Hai-Li Li conducted part of the data collection. All authors reviewed the manuscript in its final form.

References

1. Global Initiative for Asthma (GINA). *The global strategy for asthma management and prevention.* Updated 2018. http://www.ginasthma.org. 2019.
2. Los, H., G.H. Koppelman, and D.S. Postma. *The importance of genetic influences in asthma.* Eur Respir J, 1999. **14**(5): p. 1210-1227.
3. Kim, K.W. and C. Ober. *Lessons Learned From GWAS of Asthma.* Allergy Asthma Immunol Res, 2019. **11**(2): p. 170-187.
4. Bonnelykke, K., P. Sleiman, K. Nielsen, et al. *A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations.* Nat Genet, 2014. **46**(1): p. 51-55.
5. Ferreira, M.A., M.C. Matheson, C.S. Tang, et al. *Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype.* J Allergy Clin Immunol, 2014. **133**(6): p. 1564-1571.

6. Moffatt, M.F., I.G. Gut, F. Demenais, et al. *A large-scale, consortium-based genomewide association study of asthma.* N Engl J Med, 2010. **363**(13): p. 1211-1221.
7. Gudbjartsson, D.F., U.S. Bjornsdottir, E. Halapi, et al. *Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction.* Nat Genet, 2009. **41**(3): p. 342-347.
8. Hirota, T., A. Takahashi, M. Kubo, et al. *Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population.* Nat Genet, 2011. **43**(9): p. 893-896.
9. Yucesoy, B., K.M. Kaufman, Z.L. Lummus, et al. *Genome-Wide Association Study Identifies Novel Loci Associated With Diisocyanate-Induced Occupational Asthma.* Toxicol Sci, 2015. **146**(1): p. 192-201.
10. Ramasamy, A., M. Kuokkanen, S. Vedantam, et al. *Genome-wide association studies of asthma in population-based cohorts confirm known and suggested loci and identify an additional association near HLA.* PLoS One, 2012. **7**(9): p. e44008.
11. Torgerson, D.G., E.J. Ampleford, G.Y. Chiu, et al. *Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations.* Nat Genet, 2011. **43**(9): p. 887-892.
12. Leung, T.F., M.F. Tang, A.S.Y. Leung, et al. *Cadherin-related family member 3 gene impacts childhood asthma in Chinese children.* Pediatr Allergy Immunol, 2019.
13. Chen, J., J. Zhang, H. Hu, et al. *Polymorphisms of RAD50, IL33 and IL1RL1 are associated with atopic asthma in Chinese population.* Tissue Antigens, 2015. **86**(6): p. 443-447.
14. Marinho, S., A. Custovic, P. Marsden, et al. *17q12-21 variants are associated with asthma and interact with active smoking in an adult population from the United Kingdom.* Ann Allergy Asthma Immunol, 2012. **108**(6): p. 402-411 e9.
15. Yu, J., M.J. Kang, B.J. Kim, et al. *Polymorphisms in GSDMA and GSDMB are associated with asthma susceptibility, atopy and BHR.* Pediatr Pulmonol, 2011. **46**(7): p. 701-708.
16. Zavbi, M., P. Korosec, M. Flezar, et al. *Polymorphisms and haplotypes of the chromosome locus 17q12-17q21.1 contribute to adult asthma susceptibility in Slovenian patients.* Hum Immunol, 2016. **77**(6): p. 527-534.
17. Ullemar, V., P.K. Magnusson, C. Lundholm, et al. *Heritability and confirmation of genetic association studies for childhood asthma in twins.* Allergy, 2016. **71**(2): p. 230-238.
18. Sun, Y., X. Wei, J. Deng, et al. *Association of IL1RL1 rs3771180 and TSLP rs1837253 variants with asthma in the Guangxi Zhuang population in China.* J Clin Lab Anal, 2019. **33**(6): p. e22905.
19. Liang, S.Q., J.M. Deng, X. Wei, et al. *Association of GWAS-supported noncoding area loci rs404860, rs3117098, and rs7775228 with asthma in Chinese Zhuang population.* J Clin Lab Anal, 2019: p. e23066.
20. Granell, R., A.J. Henderson, D.M. Evans, et al. *Effects of BMI, fat mass, and lean mass on asthma in childhood: a Mendelian randomization study.* PLoS Med, 2014. **11**(7): p. e1001669.
21. Skaaby, T., A.E. Taylor, R.K. Jacobsen, et al. *Investigating the causal effect of smoking on hay fever and asthma: a Mendelian randomization meta-analysis in the CARTA consortium.* Sci Rep, 2017. **7**(1): p. 2224.

22. Zhao, J.V. and C.M. Schooling. *The role of linoleic acid in asthma and inflammatory markers: a Mendelian randomization study.* Am J Clin Nutr, 2019. **110**(3): p. 685-690.
23. Nuzzo, R. *Scientific method: statistical errors.* Nature, 2014. **506**(7487): p. 150-152.
24. Antonucci, L.A., G. Pergola, A. Pigoni, et al. *A Pattern of Cognitive Deficits Stratified for Genetic and Environmental Risk Reliably Classifies Patients With Schizophrenia From Healthy Control Subjects.* Biol Psychiatry, 2019.
25. Dongol Singh, S. and A. Shrestha. *Risk Factors Associated with Childhood Asthma - A Case Control Study.* Kathmandu Univ Med J (KUMJ), 2018. **16**(64): p. 290-295.
26. *Guido S., Müller A. (2016,October), Introduction to Machine Learning with Python, A Guide for scikit-learn, 123-145.*
27. Chen, T.; Guestrin, C. *XGBoost: A Scalable Tree Boosting System.* In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), San Franciso, CA, USA, 13–17 August 2016; p. 785–794.
28. Li, Linna, and Xuemin Zhang. *"Study of data mining algorithm based on decision tree."* 2010 International Conference On Computer Design and Applications. Vol. 1. IEEE, 2010.
29. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.) and random forest(RF)(Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
30. T.K. Ho. *Random decision forests.* Proceedings of the 3rd International Conference on Document Analysis and Recognition (1995), p. 278-282.
31. Gaudillo, J., J.J.R. Rodriguez, A. Nazareno, et al. *Machine learning approach to single nucleotide polymorphism-based asthma prediction.* PLoS One, 2019. **14**(12): p. e0225574.
32. Los, H., P.E. Postmus, and D.I. Boomsma. *Asthma genetics and intermediate phenotypes: a review from twin studies.* Twin Res, 2001. **4**(2): p. 81-93.
33. AlSaad, R., Q. Malluhi, I. Janahi, et al. *Interpreting patient-Specific risk prediction using contextual decomposition of BiLSTMs: application to children with asthma.* BMC Med Inform Decis Mak, 2019. **19**(1): p. 214.
34. Ogunleye, A.A. and W. Qing-Guo. *XGBoost Model for Chronic Kidney Disease Diagnosis.* IEEE/ACM Trans Comput Biol Bioinform, 2019.
35. Yu, D., Z. Liu, C. Su, et al. *Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier.* Thorac Cancer, 2020. **11**(1): p. 95-102.
36. Liu, L., Y. Yu, Z. Fei, et al. *An interpretable boosting model to predict side effects of analgesics for osteoarthritis.* BMC Syst Biol, 2018. **12**(Suppl 6): p. 105.
37. Ji, X., W. Tong, Z. Liu, et al. *Five-Feature Model for Developing the Classifier for Synergistic vs. Antagonistic Drug Combinations Built by XGBoost.* Front Genet, 2019. **10**: p. 600.
38. Ding, W., G. Chen, and T. Shi. *Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis.* Epigenetics, 2019. **14**(1): p. 67-80.

39. Fu, B., P. Liu, J. Lin, et al. *Predicting Invasive Disease-Free Survival for Early-stage Breast Cancer Patients Using Follow-up Clinical Data.* IEEE Trans Biomed Eng, 2018.

Tables

Table 1 The primer sequences of the 14 GWAS risk loci

SNP	GWAS *P*-value	Region	Functional class	Reported gene(s)	Forward	Reverse
rs6967330	3 x10-14	7q22.3	missense_variant	*CDHR3*	CAGGAGAACTCCAGCTGGTAAC	GCTTGTGTCTTCGTTGTTGG
rs928413	9 x10-13	9p24.1	upstream_gene_variant	*IL33*	GGATGAGGGGCTAAGATTCC	ATACAGATTCCCTGCCTTGG
rs4833095	5 x10-12	4p14	missense_variant	*TLR1*	CAGCTGGAGGATCCTAATGAA	TTCTGGCGAAACTTCAAACA
rs7212938	4 x10-10	17q21.1	missense_variant	*GSDMA*	CACTCAGTGTGGCTCCCAAG	AAGCACCATTTCCATGCACT
rs4143832	1 x10-10 (EA)	5q31.1	regulatory_region_variant	*IL5*	TCTGTGTGTGGCTCATCAGG	TGTGAATGGGACTCAGTGGA
rs404860	4 x10-23	6p21.32	non_coding_transcript_exon_variant	*NOTCH4*	GCACCTCTGTGGGCTATGAT	CCAGATACGGAGGCAAATCT
rs3117098	5 x10-12	6p21.32	non_coding_transcript_exon_variant	*BTNL2*	TTGCACAACATCTAAAGGTTAACTA	GGCCTTCACAAGCACAGACT
rs7775228	5 x10-15	6p21.32	regulatory_region_variant	*HLA-DQB1*	TTCTTGCAGGAGGAAAGGAA	TGCAAAGCCCCTTTATCATT
rs2305480	1 x10-7	17q21.1	missense_variant	*GSDMB*	GCTGCTCATCTCCCACACTT	AAGTCCAGAATGGCTTTTGC
rs3894194	5 x10-9	17q21.1	missense_variant	*GSDMA*	GAGGCTGTCAAGTGGTGTCA	CCCACCTCCACAGAGACAAT
rs9268516	1 x10-8	6p21.32	upstream_gene_variant	*HLA-DRA, BTNL2*	GTTGAAGGCAGAGCCAAATC	ACAACTGAACCCAGCCAAAG
rs3771180	2 x10-15	2q12.1	5_prime_UTR_variant	*IL1RL1*	TGGCCAAATCTATGACTTGTT	TCCTCTCAAGGGATTACTCAATG

rs1837253	1 x10-14	5q22.1	upstream_gene_variant	*TSLP*	AGGGCTACCCCTTGACTCAC	CCAACCAGGATTTGCAAGAA
rs7588010	4 x10-9	2p12	regulatory_region_variant	*TACR1, FAM176A*	TCATTTTCCCCATAGAAGCA	GGGCTTTAGCCTGAGATCATT

Table 2 GWAS risk loci genotype and clinical data of cases and controls

Risk features	Cases (n=123)	Controls (n=100)	$\chi 2$	*P*
rs6967330(GG/GA/AA)	109/14/0	83/15/2	3.217	0.200
rs928413(AA/AG/GG)	99/24/0	79/19/2	2.483	0.289
rs4833095(CC/CT/TT)	35/62/26	20/57/23	2.135	0.344
rs7212938(TT/GT/GG)	39/64/20	28/47/25	2.621	0.270
rs4143832(CC/AC/AA)	96/25/2	79/20/1	0.170	0.919
rs404860(TT/CT/CC)	59/50/14	47/39/14	0.350	0.840
rs3117098(TT/TC/CC)	48/56/19	55/39/6	7.991	0.018
rs7775228(TT/CT/CC)	84/33/6	64/26/10	2.184	0.336
rs2305480(CC/CT/TT)	67/45/11	45/43/12	2.060	0.357
rs3894194(TT/CT/CC)	36/56/31	29/51/20	0.998	0.607
rs9268516(CC/CT/TT)	97/24/2	77/21/2	0.128	0.938
rs3771180(CC/AC/AA)	107/16/0	92/8/0	1.440	0.230
rs1837253(TT/CT/CC)	31/69/23	48/41/11	12.785	0.002
rs7588010(CC/AC/AA)	67/49/7	52/37/11	2.104	0.349
BMI (light weight /normal weight /overweight)	15/81/21	21/68/11		0.037
Exposure to allergens (yes)	20 (16.3%)	14 (14.0%)	0.218	0.641
Exposure to tobacco smoke (yes)	42 (34.1%)	24 (24.0%)	2.725	0.099
Family history	46(37.4%)	0(0.0%)	47.118	<0.000

(yes)

Table 3 Performance comparison among the four different classifiers: gradient boosting (XGBoost), decision tree (DT), support vector machine (SVM) and random forest (RF)

	Classifiers	Accuracy	Precision	Recall	F1_score	AUC
GWAS risk loci only	XGBoost	0.601	0.634	0.678	0.649	0.643
	DT	0.494	0.538	0.566	0.547	0.484
	SVM	0.569	0.590	0.714	0.641	0.586
	RF	0.551	0.584	0.655	0.613	0.562
Clinical data only	XGBoost	0.696	0.867	0.538	0.653	0.714
	DT	0.684	0.865	0.514	0.633	0.694
	SVM	0.689	0.862	0.527	0.642	0.711
	RF	0.661	0.826	0.507	0.613	0.711
Four significant differences features	XGBoost	0.686	0.858	0.527	0.644	0.702
	DT	0.678	0.852	0.514	0.627	0.684
	SVM	0.682	0.845	0.527	0.641	0.690
	RF	0.682	0.845	0.527	0.641	0.692
Combined GWAS risk loci and clinical data	XGBoost	0.708	0.760	0.715	0.728	0.797
	DT	0.687	0.762	0.651	0.694	0.755
	SVM	0.678	0.792	0.608	0.667	0.788
	RF	0.678	0.738	0.660	0.691	0.756

Figures

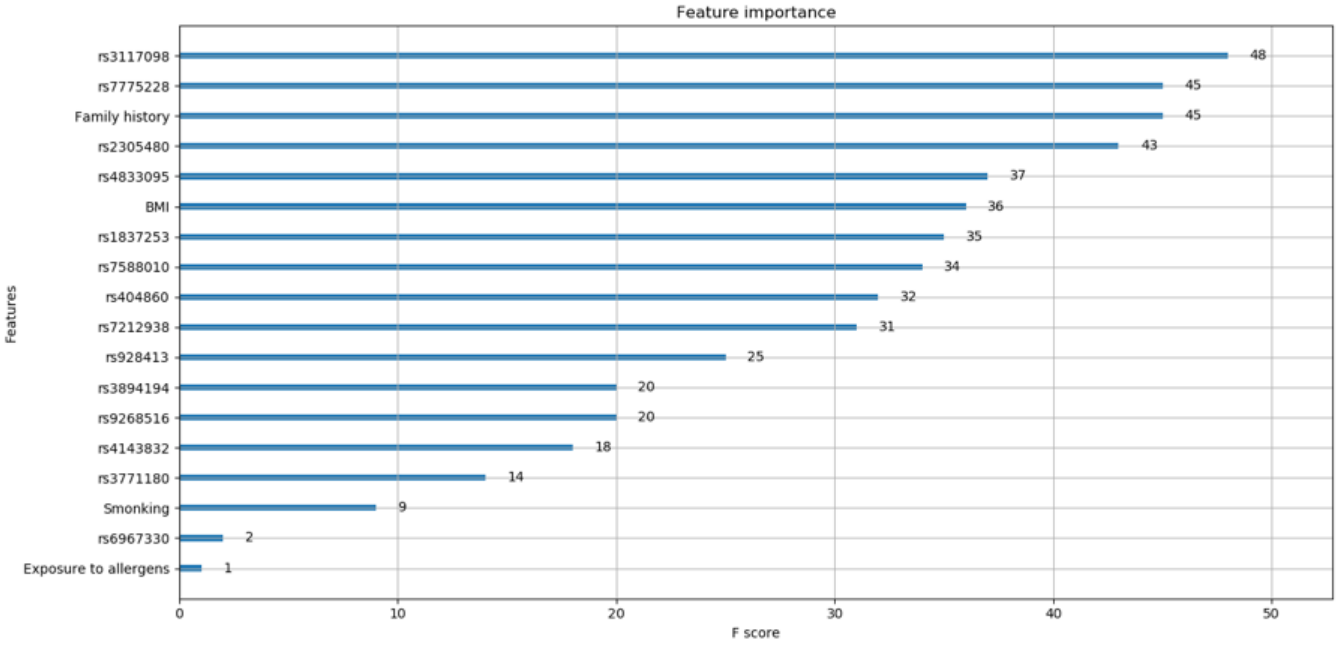

Figure 1

Feature importance plot of features