International Journal of Chronic Obstructive Pulmonary Disease Dovepress

Open Access Full Text Article

ORIGINAL RESEARCH

Screening the Best Risk Model and Susceptibility SNPs for Chronic Obstructive Pulmonary Disease (COPD) Based on Machine Learning Algorithms

Zehua Yang*, Yamei Zheng*, Lei Zhang, Jie Zhao, Wenya Xu, Haihong Wu, Tian Xie, Yipeng Ding

Department of Respiratory and Critical Care Medicine, Hainan Affiliated Hospital of Hainan Medical University, Hainan General Hospital, Haikou, Hainan, 570311, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yipeng Ding; Tian Xie, Department of Respiratory and Critical Care Medicine, Hainan Affiliated Hospital of Hainan Medical University, Hainan General Hospital, 19 Xiuhua Road, Xiuying District, Haikou, Hainan, 570311, People's Republic of China, Tel +86-18976335858, Email yipengding2024@163.com; hpphxietian@163.com

Background and Purpose: Chronic obstructive pulmonary disease (COPD) is a common and progressive disease that is influenced by both genetic and environmental factors, and genetic factors are important determinants of COPD. This study focuses on screening the best predictive models for assessing COPD-associated SNPs and then using the best models to predict potential risk factors for COPD.

Methods: Healthy subjects (n=290) and COPD patients (n=233) were included in this study, the Agena MassARRAY platform was applied to genotype the subjects for SNPs. The selected sample loci were first screened by logistic regression analysis, based on which the key SNPs were further screened by LASSO regression, RFE algorithm and Random Forest algorithm, and the ROC curves were plotted to assess the discriminative performance of the models to screen the best prediction model. Finally, the best prediction model was used for the prediction of risk factors for COPD.

Results: One-way logistic regression analysis screened 44 candidate SNPs from 146 SNPs, on the basis of which 44 SNPs were screened or feature ranked using LASSO model, RFE-Caret, RFE-Lda, RFE-lr, RFE-nb, RFE-rf, RFE-treebag algorithms and random forest model, respectively, and obtained ROC curve values of 0.809, 0.769, 0.798, 0.743, 0.686, 0.766, 0.743, 0.719, respectively, so we selected the lasso model as the best model, and then constructed a column-line graph model for the 25 SNPs screened in it, and found that rs12479210 might be the potential risk factors for COPD.

Conclusion: The LASSO model is the best predictive model for COPD and rs12479210 may be a potential risk locus for COPD.

Keywords: COPD, LASSO, machine learning, predictive model, SNP

Introduction

Chronic obstructive pulmonary disease (COPD) has become a public health challenge due to its high prevalence worldwide and the associated disability, morbidity, mortality and socioeconomic burden.[1–3] Rehman et al reported a prevalence of COPD of 3.4–13.4%[3] in Europe and the United States and 3.5–19.1% in Asia due to urbanisation, industrial pollution, tanneries and high household use of biofuels.[4,5] The number of COPD deaths in China exceeded 900,000 in 2013 and COPD is now the third leading cause of death in China. Typical symptoms of COPD include dyspnoea, chronic cough and sputum production, and spirometry is considered the gold standard in the diagnosis of COPD,[6] however, early-stage COPD often goes undetected, resulting in patients with early-stage COPD being under-diagnosed and under-treated. Therefore, there is a need to develop a reliable early warning method for COPD. This will lead to early intervention and treatment of COPD.

A single nucleotide polymorphism (SNP) is a type of DNA polymorphism that refers to a change in a single nucleotide that result in different DNA sequences that, after transcription and translation, result in functional differences

Received: 17 May 2024
Accepted: 9 October 2024
Published: 5 November 2024

in the final expression of the protein.[7] SNPs are the most common genetic variation in the human genome and the most common form of DNA sequence variation that reflects individual differences. On average, there are about 1 SNPs per 1000 bases, and only a fraction of these specific SNPs are associated with disease.[8] They are known as the third generation of genetic markers because of their widespread use, large numbers, stable genetic properties and ease of automated batch detection. Currently, with the development of SNPs detection technology, it is widely used in the study of common and complex diseases, medical diagnosis, drug development and the exploration of disease susceptibility genes.[9]

In a genome-wide association study (GWAS), one study analysed a large cohort of patients and found that as many as 3 ~ 1 million SNPs in cases were COPD disease-associated loci.[10] In 2017, Wain LV et al found that 95 loci in FEVI, FVC and FEV1/FVC were associated with COPD risk, and enrichment analysis showed that these loci were associated with lung development, elastic fibres and epigenetic regulatory pathways.[11] In 2019, a study found 82 loci associated with COPD, with a total of 156 risk genes located in these loci.[12] Recently, Shrine N et al identified 257 loci associated with lung function phenotypes, of which 107 were identified as risk genes for COPD.[13] Currently, for the set of SNPs genes that are significantly associated with COPD susceptibility, it is crucial that gene targeting and identification of individual disease-causing variants is carried out in subsequent studies.[14]

Least Absolute Shrinkage and Selection Operator (LASSO) method is a statistical approach that integrates feature selection with regularization, which improves the predictive power of models by applying a penalty to the magnitude of the coefficients, thus reducing the complexity and preventing overfitting.[15] Recursive Feature Elimination (RFE) is an efficient machine learning technique suitable for both classification and regression. It works by determining the optimal dividing hyperplane in the feature space to distinguish between classes or to minimize errors in fitting the regression function.[16] Random Forest algorithm is a form of ensemble learning that operates by generating an ensemble of decision trees, and it enhances predictive accuracy and reliability by considering the majority vote among the trees for classification tasks or by averaging their predictions in the case of regression.[17] Cross-validation with Random Forest, LASSO and RFE algorithms was performed to mitigate the risk of overfitting.

Therefore, in this study, we used a variety of statistical algorithms to construct models by one-way logistic regression analysis, LASSO regression, RFE Algorithm and Random Forest with feature selection and screening of key SNPs, plotted column-line plots based on the SNPs screened by the best model, and assessed the discriminative power of the model in the original dataset using calibration curves and receiver operating characteristic (ROC) curves. To our knowledge, this is the first study to investigate the contribution of SNPs to COPD risk using LASSO regression, the RFE algorithm and random forests.

Materials and Methods

Study Population

A total of 233 people with COPD and 290 healthy controls were included in the study for a case control study. Based on the Global Initiative for Chronic Obstructive Lung Disease criteria, individuals were diagnosed with COPD with the ratio of forced expiratory volume in 1 second (FEV1) /forced vital capacity (FVC) < 70% and FEV1<80% predicted. COPD patients with a history of serious illnesses such as bronchial asthma, tuberculosis and lung cancer were not included in this study. The control group consisted of healthy people without pulmonary dysfunction, lung-related diseases, other chronic diseases and disorders, and severe endocrine, metabolic and nutritional disorders, who underwent a health check-up at the same hospital during the same period. Clinical characteristics of the subjects, including smoking, body mass index (BMI), complications, wheezing, gasping, chest pain and respiratory infections, were collected from medical records and questionnaires. The study protocol was approved by the Ethics Committee of Hainan Provincial People's Hospital in accordance with the Declaration of Helsinki. All subjects signed an informed consent form.

Selection of SNPs

We identified SNPs associated with COPD based on the literature in PubMed (https://pubmed.ncbi.nlm.nih.gov/) and our case-control study of COPD. We then screened SNPs located on these genes from the Chinese Han Beijing (CHB) dataset of the Thousand Genomes Project (https://www.internationalgenome.org/) and the Ensembl website (http://www.ensembl.org), considering only the minimum allele frequency (MAF) ≥ 0.05 for SNPs. Haploview v4.2 software (Broad Institute of MIT and Harvard) was used to predict marker SNPs for each gene.

Genomic DNA Extraction and SNPs Genotyping

Peripheral blood samples were collected from all subjects and genome extraction kits were purchased from Xi'an Gold & Magnesium Co. Amplification primers were designed using the MassARRAY Assay Design software and genotyping was performed using the MassARRAY platform (Agena, San Diego, CA, USA). The generated assay data was analysed using AgenaTyper v4.0 software, which requires a call rate of ≥95% for candidate SNPs.

Definition of Data Characteristics

The total study population in this study was 523 individuals, with the minimum allele being the risk allele in the healthy control population, and 0, 1 and 2 denoting the number of risk alleles carried by an individual, being 2 carried by AA, 1 carried by AB and 0 carried by BB (the minimal allele was A). In addition, we specified the number of COPD patients and healthy controls as the dependent variable and the number of SNPs carrying risk alleles in each sample as the explanatory variable. These data were finally screened as data features for machine learning by one-way logistic regression and LASSO model, RFE-Caret, RFE-Lda, RFE-lr, RFE-nb, RFE-rf, RFE-treebag algorithms and random forest model.

Annotation Analysis of SNPs

Expression quantitative trait locus (eQTL) analysis can identify possible causative genes within COPD susceptibility loci.[18] Motifs are a class of gene loci that can influence gene expression, and most of these loci are SNPs. In this study, we used the online tool HaploReg v4.1 (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php) to perform functional annotation analyses of the screened SNPs, including eQTL analysis, motif change regulation analysis and SNPs mapping.

Data Analysis

In this study, we used R v4.2.1 to perform batch one-way logistic regression analysis on 146 SNPs loci from 523 samples, and screened the SNPs obtained by screening in LASSO regression, RFE algorithm and randomforest algorithm, respectively, to construct the models associated with COPD risk, and plotted ROC curves to evaluate the model classification performance was selected, and the model with the best performance was selected to construct the column line graph of SNPs loci associated with COPD risk. The Hosmer-Lemeshow test was used to assess the goodness of fit of the column line plots and visualised by calibration curves. SPSS 22.0 statistical software was used and comparisons of normally distributed measures were analysed by ANOVA, with measures expressed as mean ± s of x and non-normally distributed measures expressed as median (interquartile range) using the rank sum test. Count data were analysed using the χ^2 test. Logistic regression analysis was performed using the Wald test with $p<0.05$ as a statistically significant difference.

LASSO regression using the R package "glmnet" and 10-fold cross-validation using the "cv" function. Use the Glmnet package to obtain the most appropriate penalty factor λ. The importance of each SNPs was assessed using the R package "randomforest" and the Lda, lr, nbFuncs, rf and treebag parameters in "caret", followed by plotting the ROC curves using the functions in the "pROC" package and performing the Hosmer-Lemeshow test using the R package "ResourceSelection", where a significant p-value indicates a poorly fitted model.

Results

Genotyping results of SNPs

Based on the screening criteria, 146 SNPs from 43 genes were screened and genotyped among 233 COPD patients and 290 healthy controls using the Agena MassARRAY technique, and all SNPs met the typing success rate of ≥95% and Hardy-Weinberg equilibrium p>0.05 after chi-square test. The information corresponding to 146 of these SNPs and 43 genes is shown in Table S1. The results of 146 SNPs genotyping were displayed in Table 1.

One-Way Logistic Regression Analyses

The results of univariate analysis showed that among the successfully typed loci, 44 SNPs had statistically significant effects on the risk of COPD ($p< 0.05$) (Table 2).

LASSO Regression Analysis

Based on the results of the 10-fold cross-validation, we obtained the value of λ at the minimum of the mean square error (MSE) (lambda.min) and the value of λ one standard error away from the minimum of the MSE (lambda.1se), with the corresponding number of SNPs varying with the value of the penalty coefficient λ (Figure 1A). In this study, we chose λ=0.033, which had the highest penalty value, as the optimal λ. Figure 1B shows a total of 25 significant SNPs observed at λ=0.033, of which 13 SNPs were positively correlated with the risk of COPD, namely rs12479210 (β=0. 411), rs1420101 (β=0.0000572), rs9320913 (β=0.128), rs4646437 (β=0.0611), rs298207 (β=0.0207), rs16907751 (β=0.377), rs759648 (β=0.126), rs2420915 (β=0.0520), rs78750958 (β=0. 0520), rs1484215 (β=0.0846), rs3024622 (β=0.165), rs1038376 (β=0.511) and rs2853676 (β=0.209), and 12 SNPs were negatively correlated with the risk of COPD, namely rs13097407 (β=−0.152), rs352140 (β=−0.0769), rs911186 (β=−1. 42), rs2505059 (β=−0.141), rs10245353 (β=−0.128), rs4719841 (β=−0.231), rs13271489 (β=−0.294), rs7934083 (β=−0.441), rs9525927 (β=−0.197), rs3093193 (β=−0.233), rs3093110 (β=−0.250), rs4803420 (β=−0.115) (Table 3). The area under the curve (AUC) of the ROC curve was 0.809, an indication that the model had good classification results (Figure 1C).

RFE Algorithm

Based on the RFE algorithm analysis, a total of 38 significant SNPs were screened in the caret model, 42 significant SNPs in the Lda model, 42 significant SNPs in the lr model, 4 significant SNPs in the nb model, 42 significant SNPs in the rf model, and 44 significant SNPs in the treebag model (Table 4). In addition, the AUC of the ROC curve of the caret model is 0.769, the AUC of the ROC curve of the Lda model is 0.798, the AUC of the ROC curve of the lr model is 0.743, the AUC of the ROC curve of the nb model is 0.686, the AUC of the ROC curve of the rf model is 0.766, the AUC of the ROC curve of the treebag model is 0.743, and all these AUC values have AUC values of 0.769. 686, the AUC of the ROC curve of the rf model is 0.766, the AUC of the ROC curve of the treebag model is 0.743, and all these AUC values have AUC > 0.5, so all six models are considered to have good classification performance (Figure 2).

Random Forest(RF) Assessment

To assess the significance of the contribution of SNPs obtained from genotyping to COPD risk, we made a random forest decision based on the characteristics of the sample data described above. In the random forest model, the relative importance of a variable is the total reduction in node impurity when that variable is equally distributed across all trees, and node impurity is defined by the Gini coefficient. Therefore, we ranked the variables according to the size of the average decreasing Gini coefficient of the Random Forest output and ranked the 44 SNPs in order of importance from largest to smallest (Table 5). The AUC of the ROC curve is 0.719, which is an indication that the model has a good classification performance (Figure 3).

Identification and Validation of Personalized Predictive Models

Based on the above AUC values, the performance of these eight classifiers was evaluated, and Lasso (0.809) > lda (0.798) > caret (0.769) > rf (1.766) > lr (0.743) = treebag (0.743) > RF (0.719) > nb (0.686), the 25 SNPs screened by the

Table 1 The Results of 146 SNPs Genotyping Using the MassARRAY Platform

SNP ID	Genotype	Group		SNP ID	Genotype	Group		SNP ID	Genotype	Group		SNP ID	Genotype	Group	
		Control	Case			Control	Case			Control	Case			Control	Case
rs2295359	A A	34	30	rs85	C C	32	32	rs10036748	C C	10	3	rs1801275	G G	8	6
	G G	124	102		C T	139	105		CT	76	62		A G	86	71
	A G	132	101		T T	117	96		T T	204	168		A A	196	156
rs7517847	G G	50	48	rs10245353	A A	49	32	rs10069690	T T	7	4	rs5744174	G G	14	6
	T T	104	65		C A	142	99		CT	55	66		G A	102	104
	G T	134	120		C C	99	101		C C	226	163		A A	174	123
rs2201841	A A	15	15	rs2290263	A G	19	32	rs10439478	C C	39	29	rs352140	T T	47	22
	G A	111	103		A A	271	201		CA	107	91		C C	102	110
	G G	159	115						A A	144	113		C T	141	101
rs12743974	G G	17	17	rs4719841	A A	49	17	rs1056629	C C	51	43	rs3804099	C C	37	26
	G A	115	100		G G	94	107		T T	102	87		C T	128	94
	A A	158	116		A G	147	109		CT	136	103		T T	125	113
rs10889677	C C	17	15	rs7780562	A A	29	13	rs1056654	A A	51	43	rs3804100	C C	28	22
	C A	111	102		C A	122	97		G G	103	87		C T	126	79
	A A	160	115		C C	139	119		GA	136	103		T T	136	132
rs2201584	A A	26	30	rs483916	C A	12	2	rs1056675	C C	45	33	rs5743705	C C	0	1
	G A	119	102		A A	278	231		T T	93	83		C T	43	22
	G G	143	101						TC	147	115		T T	247	210
rs10489626	G C	26	28	rs13271489	C T	20	4	rs10936599	T T	60	44	rs6430491	A A	54	23
	C C	262	205		T T	270	228		C C	79	61		G G	95	90
									CT	151	127		G A	141	120
rs6659932	A A	1	0	rs6994670	G G	23	8	rs11125529	A A	3	9	rs2593704	G G	22	12
	C A	21	24		A G	114	82		CA	72	55		C G	118	71
	C C	268	208		A A	152	143		C C	215	169		C C	146	150
rs1874791	A A	9	18	rs298207	A A	7	15	rs11191865	A A	45	33	rs911186	G G	1	0
	G A	99	79		G A	90	76		G G	100	91		G A	37	2
	G G	180	136		G G	190	138		AG	141	109		A A	252	231

(*Continued*)

https://doi.org/10.2147/COPD.S478634

https://doi.org/10.2147/COPD.S478634
DovePress

Table 1 (Continued).

SNP ID	Genotype	Group		SNP ID	Genotype	Group		SNP ID	Genotype	Group		SNP ID	Genotype	Group	
		Control	Case			Control	Case			Control	Case			Control	Case
rs6679356	C C	1	0	rs6473227	A A	63	51	rs11859599	C C	3	6	rs9320913	A A	40	51
	T C	21	28		C C	85	65		GC	57	38		C C	105	53
	T T	267	205		C A	142	117		G G	230	189		C A	145	124
rs3790567	A A	15	23	rs16907751	T T	6	16	rs11896604	G G	13	20	rs28681535	T T	53	47
	A G	118	94		T C	82	71		CG	82	63		G G	88	75
	G G	157	116		C C	202	142		C C	195	150		T G	147	109
rs6689306	A A	54	61	rs759648	C C	8	21	rs12615793	A A	3	9	rs3093203	A A	11	17
	G G	79	65		C A	74	76		AG	72	56		A G	111	98
	A G	155	107		A A	208	136		G G	215	168		G G	166	104
rs4537545	T T	50	44	rs2608029	C C	5	4	rs12621038	T T	34	35	rs3093193	G G	28	9
	C C	95	96		G C	78	64		C C	122	104		G C	124	80
	T C	144	93		G G	206	165		TC	132	94		C C	138	144
rs4845625	T T	56	66	rs13280095	C C	4	2	rs12765878	C C	45	33	rs12459936	T T	51	57
	C C	77	64		C A	61	41		T T	99	91		C C	89	67
	C T	155	103		A A	224	190		CT	146	109		T C	149	108
rs4129267	T T	52	46	rs2420915	A A	34	36	rs1682111	A A	14	10	rs3093144	T T	7	5
	C C	94	93		G G	128	79		TA	109	87		T C	87	49
	T C	142	94		G A	128	118		T T	167	136		C C	196	179
rs2228145	C C	52	46	rs1907240	G G	41	45	rs17045754	C C	6	9	rs3093110	G G	5	0
	A A	97	93		A A	108	78		GC	82	63		G A	67	29
	C A	140	94		G A	141	108		G G	202	161		A A	217	204
rs72823641	T A	15	5	rs2257129	T T	40	39	rs2075786	G G	13	8	rs2099361	C C	22	16
	T T	275	228		C C	106	76		AG	78	62		C A	129	91
					T C	141	117		A A	197	159		A A	134	126
rs12479210	T T	31	68	rs7934083	C C	16	3	rs2188971	T T	27	25	rs4803418	G G	55	43
	C C	138	55		C T	80	41		C C	125	99		C C	64	69
	C T	121	110		T T	194	187		TC	138	107		G C	170	121
rs3771180	T T	5	0	rs78750958	A A	16	17	rs2188972	G G	67	54	rs2505059	G G	34	14
	G T	46	25		A G	101	107		A A	76	68		A A	108	113
	G G	238	208		G G	173	109		GA	147	111		A G	147	104

Dovepress

rs1420101	T T C C T C	31 138 121	68 55 110	rs9533803	C C T T T C	68 67 155	55 67 111	rs2242652	A A GA G G	8 58 223	4 67 162	rs12979270	C C C A A A	17 122 140	21 106 104
rs3771175	A A A T T T	4 43 243	0 25 208	rs9525927	A A G G G A	59 79 152	36 97 100	rs2297441	A A G G AG	57 97 135	38 81 114	rs843706	C C A A C A	50 106 132	29 98 106
rs10208293	A A A G G G	9 69 208	2 32 199	rs7981875	G G G A A A	3 59 226	3 53 169	rs2320615	A A AG G G	10 82 198	10 69 154	rs843711	C C T T C T	50 106 131	29 98 105
rs10197862	G G G A A A	5 46 237	0 25 208	rs9527345	T T C T C C	28 109 153	17 94 122	rs2853676	T T CT C C	6 58 226	7 62 164	rs843720	G G G T T T	23 117 150	10 95 128
rs1861245	T T T C C C	25 49 216	11 23 199	rs2252932	G G G A A A	7 86 197	5 63 164	rs2853677	G G A A AG	34 138 118	23 85 125	rs843748	A A G A G G	9 91 184	12 75 143
rs9807989	C C T C T T	9 57 222	2 32 199	rs2997119	A A G G G A	68 89 13	57 56 116	rs2967361	T T GT G G	8 86 195	18 57 158	rs843752	G G T T G T	69 94 125	52 71 110
rs13015714	G G T T G T	11 50 228	36 79 118	rs2280274	T T T A A A	2 55 233	1 33 199	rs35073794	AG G G	5 285	3 230	rs9325507	T T C C TC	45 99 146	33 91 109
rs2287037	T T C C C T	37 128 123	61 59 113	rs4388726	A A A G G G	1 31 258	0 23 210	rs3751862	C C CA A A	0 37 253	6 33 193	rs9420907	CA A A	21 269	11 222
rs2058622	A A G G G A	83 69 135	39 78 116	rs11207535	G G G A A A	3 57 230	1 44 188	rs3792792	CT T T	6 284	7 226	rs4803420	T T T G G G	12 109 167	6 61 166
rs3771166	A A A G G G	9 60 219	2 33 198	rs10889159	T T T C C C	3 69 215	1 49 182	rs3814220	G G A A GA	45 99 146	33 91 109	rs1038376	T T T A A A	4 53 233	11 84 138

(*Continued*)

Table 1 (Continued).

SNP ID	Genotype	Group		SNP ID	Genotype	Group		SNP ID	Genotype	Group		SNP ID	Genotype	Group	
		Control	Case			Control	Case			Control	Case			Control	Case
rs6543124	A A	4	0	rs1155002	T T	15	17	rs4809324	C C	19	5	rs2239347	A A	88	60
	A T	57	31		C C	155	89		TC	89	80		C A	147	116
	T T	229	202		T C	119	127		T T	182	146		C C	51	57
rs3804795	C C	12	14	rs3735451	C C	18	16	rs6010620	G G	46	23	rs3024622	G G	52	59
	T C	106	81		C T	132	101		A A	113	103		C C	107	58
	T T	171	138		T T	139	116		GA	135	107		C G	131	116
rs2290610	C C	45	46	rs4646440	A A	10	12	rs6010621	G G	37	23	rs1484215	T T	8	14
	T T	102	76		A G	113	80		GT	129	102		T C	89	88
	C T	142	111		G G	164	141		T T	123	108		C C	193	131
rs13097407	G G	2	1	rs35564277	C C	2	4	rs6089953	G G	39	29	rs12912592	T T	1	0
	G A	40	14		C T	36	30		A A	112	101		T G	48	29
	A A	248	218		T T	250	199		AG	139	103		G G	238	204
rs334782	C C	14	9	rs4646437	A A	1	8	rs6713088	G G	67	45	rs8105767	G G	44	45
	C T	111	73		A G	72	71		C C	90	67		A A	99	73
	T T	165	151		G G	216	153		CG	128	118		AG	147	114
rs3856850	G G	43	53	rs111853758	G G	5	0	rs7248488	A A	28	25	rs843645	G G	50	39
	A A	102	75		G T	50	30		C C	125	99		T T	118	86
	A G	145	105		T T	234	195		CA	137	109		GT	122	108
rs4787951	C C	26	31	rs4494250	A A	10	6	rs7708392	G G	10	3	rs8103163	A A	27	25
	C T	135	109		A G	102	72		GC	76	61		C C	125	99
	T T	125	93		G G	171	155		C C	204	169		CA	137	109
rs3785356	T T	25	29	rs75665761	A A	2	0								
	C C	84	71		A G	36	35								
	C T	177	133		G G	252	198								

Table 2 Univariate Logistic Regression Results (Only Significant SNPs are Included in the Table)

Number of Sites	SNP_ID	OR 95% CI	*p*-value	Number of Sites	Characteristics	OR 95% CI	*p*-value
1	rs1155002	1.63 (1.22–2.18)	0.001	23	rs4719841	0.59 (0.45–0.77)	*p*<0.001
2	rs12479210	2.34 (1.81–3.02)	*p*<0.001	24	rs4646437	1.62 (1.14–2.3)	0.007
3	rs3771180	0.54 (0.33–0.88)	0.013	25	rs483916	0.2 (0.04–0.9)	0.037
4	rs1420101	2.34 (1.81–3.02)	*p*<0.001	26	rs13271489	0.24 (0.08–0.7)	0.009
5	rs3771175	0.6 (0.37–0.97)	0.039	27	rs6994670	0.69 (0.52–0.93)	0.013
6	rs10208293	0.48 (0.32–0.72)	*p*<0.001	28	rs298207	1.37 (1.01–1.85)	0.045
7	rs10197862	0.54 (0.33–0.87)	0.012	29	rs16907751	1.48 (1.09–2.02)	0.012
8	rs1861245	0.63 (0.45–0.86)	0.004	30	rs759648	1.76 (1.31–2.37)	*p*<0.001
9	rs9807989	0.59 (0.39–0.88)	0.009	31	rs111853758	0.62 (0.39–0.97)	0.037
10	rs2287037	1.91 (1.49–2.45)	*p*<0.001	32	rs2420915	1.36 (1.05–1.75)	0.02
11	rs2058622	0.65 (0.51–0.83)	0.001	33	rs7934083	0.5 (0.35–0.71)	*p*<0.001
12	rs3771166	0.57 (0.38–0.86)	0.007	34	rs78750958	1.48 (1.11–1.96)	0.007
13	rs6543124	0.56 (0.36–0.88)	0.012	35	rs9525927	0.67 (0.52–0.86)	0.002
14	rs6430491	0.72 (0.56–0.94)	0.015	36	rs1484215	1.51 (1.12–2.05)	0.007
15	rs2593704	0.65 (0.49–0.87)	0.004	37	rs3024622	1.46 (1.14–1.86)	0.002
16	rs13097407	0.44 (0.25–0.8)	0.006	38	rs3093203	1.5 (1.12–2.01)	0.007
17	rs352140	0.66 (0.51–0.86)	0.002	39	rs3093193	0.59 (0.44–0.79)	*p*<0.001
18	rs911186	0.06 (0.01–0.25)	*p*<0.001	40	rs3093144	0.69 (0.48–0.97)	0.034
19	rs2505059	0.65 (0.49–0.85)	0.002	41	rs3093110	0.42 (0.27–0.67)	*p*<0.001
20	rs9320913	1.61 (1.24–2.09)	*p*<0.001	42	rs4803420	0.61 (0.44–0.84)	0.003
21	rs10245353	0.77 (0.6–0.99)	0.042	43	rs1038376	2.51 (1.77–3.57)	*p*<0.001
22	rs2290263	2.27 (1.25–4.12)	0.007	44	rs2853676	1.42 (1.01–2.01)	0.045

Note: p<0.05: indicates statistical significance.
Abbreviation: OR, Odds Ratio.

best LASSO model were selected as independent predictors of COPD risk. Based on HaploReg v4.1 database, the potential functions of these SNPs were displayed in Table S2. Nomogram for predictive models were constructed based on 25 SNPs screened by the best LASSO model (Figure 4A). Nomogram results showed that rs1038376 and rs12479210 polymorphic loci contributed most to the increased risk of developing COPD, whereas rs13097407, rs352140, rs911186, rs2505059, rs1024535, rs471984, rs1327148, rs7934083, rs952592, rs3093193, rs3093110 and rs4803420 risk alleles were the protective factors for COPD risk. Figure 4B shows the calibration curves for the Nomogram we constructed, and the actual curves are closer to the ideal curves, indicating that the model is well calibrated in the dataset.

Discussion

COPD is an irreversible and progressive disease, so there is an urgent need to diagnose COPD in its early stages.[19] A combination of genome-wide association studies and candidate gene analysis can help identify genetic variants that contribute to an individual's predisposition to COPD.[10] Although various types of risk prediction models have been developed in abundance in recent years, most are based on individual models or algorithms for prediction, eg Jin et al identified race SNPs by filtering through best linear unbiased prediction (BLUP) in a linear mixed model,[20] correlation

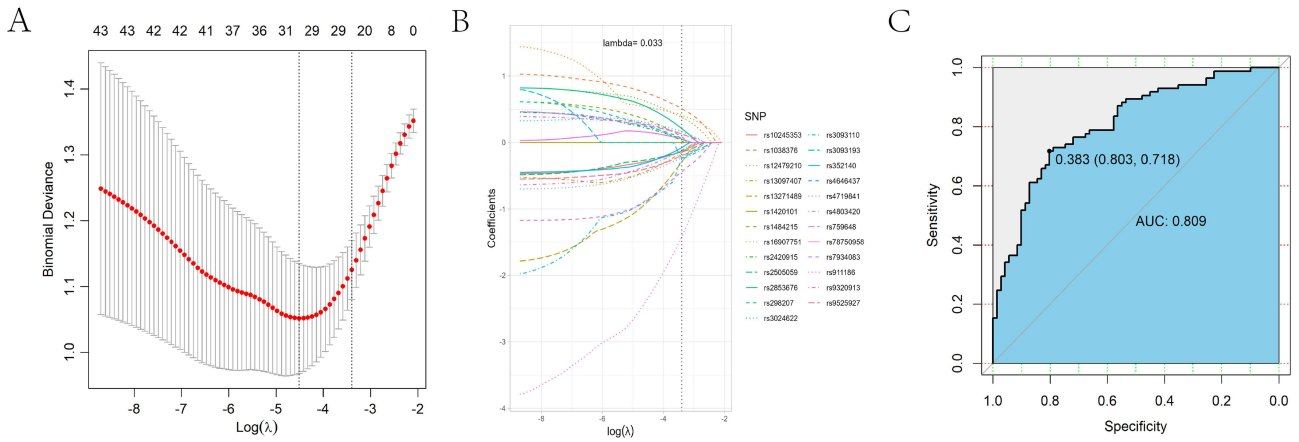

Figure 1 LASSO regression analysis. (**A**) 10-fold cross-validation of the results. The value in the middle of the two dotted lines is the range of the positive and negative standard deviations of log(λ). The dotted line on the left indicated the value of the harmonic parameter log(λ) when the error of the model is minimized. 25 variables were selected when log(λ) = 0.033. (**B**) LASSO coefficient profiles of 25 significant SNPs. A vertical line was drawn at the value chosen by 10-fold cross-validation. As the value of λ decreased, the degree of model compression increased and the function of the model to select important variables increased. (**C**) Receiver operating characteristic (ROC) curves of 25 SNPs in LASSO regression analysis. AUC = 0.809.

between IL95R SNPs and the risk of COPD as calculated by logistic regression analysis according to Zhou et al,[21] although the overall predictive ability of KNN, LR and XGboost models has been reported,[19] the most effective model for predicting genetic polymorphisms has not been reported in individual prediction models.

Previous studies assessed the heritability of COPD and related phenotypes in smokers among the non-Hispanic whites.[22] Matthew Moll constructed a polygenic risk score using a genome-wide association study of lung function for COPD from the UK Biobank and SpiroMeta.[23] A multi-ancestry genome-wide association analyses and systematic

Table 3 Significant SNPs and Their Coefficients After LASSO Regression

Number	SNP_ID	Coefficients
(Intercept)	(Intercept)	−0.460
1	rs12479210	0.411
2	rs1420101	0.0000572
3	rs13097407	−0.152
4	rs352140	−0.0769
5	rs911186	−1.42
6	rs2505059	−0.141
7	rs9320913	0.128
8	rs10245353	−0.128
9	rs4719841	−0.231
10	rs4646437	0.0611
11	rs13271489	−0.294
12	rs298207	0.0207
13	rs16907751	0.377

(*Continued*)

Table 3 (Continued).

Number	SNP_ID	Coefficients
14	rs759648	0.126
15	rs2420915	0.0520
16	rs7934083	−0.441
17	rs78750958	0.0516
18	rs9525927	−0.197
19	rs1484215	0.0846
20	rs3024622	0.165
21	rs3093193	−0.234
22	rs3093110	−0.250
23	rs4803420	−0.115
24	rs1038376	0.511
25	rs2853676	0.209

Table 4 SNPs Selected by Caret, Lda, Lr, Nb, Rf and Treebag Models Constructed Based on Recursive Feature Elimination (RFE) Algorithm

Number	SNP_ID (Caret)	SNP_ID (Lda)	SNP_ID (lr)	SNP_ID (nb)	SNP_ID (rf)	SNP_ID (Treebag)
1	rs1420101	rs12479210	rs1038376	rs12479210	rs1038376	rs12479210
2	rs12479210	rs1420101	rs4719841	rs1420101	rs3024622	rs1420101
3	rs1038376	rs2287037	rs2420915	rs1038376	rs1155002	rs9320913
4	rs2287037	rs1038376	rs911186	rs2287037	rs9525927	rs352140
5	rs9525927	rs7934083	rs1484215		rs10245353	rs4719841
6	rs4719841	rs4719841	rs1155002		rs1484215	rs2593704
7	rs4803420	rs9320913	rs9525927		rs1420101	rs1038376
8	rs3093110	rs1155002	rs4803420		rs911186	rs7934083
9	rs2853676	rs9525927	rs2853676		rs12479210	rs1155002
10	rs3024622	rs4803420	rs2290263		rs9320913	rs9525927
11	rs1484215	rs2058622	rs12479210		rs4803420	rs911186
12	rs3093203	rs10208293	rs352140		rs6430491	rs2287037
13	rs3093193	rs352140	rs16907751		rs4719841	rs1484215
14	rs2058622	rs78750958	rs3093203		rs352140	rs3024622
15	rs352140	rs4646437	rs7934083		rs2420915	rs10245353
16	rs2593704	rs911186	rs298207		rs7934083	rs6430491
17	rs9320913	rs3024622	rs2593704		rs16907751	rs2505059

(*Continued*)

Table 4 (Continued).

Number	SNP_ID (Caret)	SNP_ID (Lda)	SNP_ID (lr)	SNP_ID (nb)	SNP_ID (rf)	SNP_ID (Treebag)
18	rs2505059	rs759648	rs759648		rs78750958	rs759648
19	rs1155002	rs298207	rs10208293		rs3093203	rs3093193
20	rs78750958	rs3093193	rs9807989		rs298207	rs3093203
21	rs10208293	rs2593704	rs3771175		rs759648	rs2058622
22	rs16907751	rs1861245	rs6994670		rs2593704	rs298207
23	rs2420915	rs2505059	rs3024622		rs6994670	rs4646437
24	rs6994670	rs3093110	rs111853758		rs3093193	rs4803420
25	rs759648	rs6430491	rs9320913		rs2505059	rs78750958
26	rs1861245	rs3771166	rs10245353		rs3093110	rs2853676
27	rs7934083	rs9807989	rs2505059		rs4646437	rs111853758
28	rs911186	rs3093203	rs78750958		rs2058622	rs16907751
29	rs6430491	rs6543124	rs13097407		rs2290263	rs2420915
30	rs4646437	rs6994670	rs3093110		rs2287037	rs10208293
31	rs3771166	rs16907751	rs4646437		rs3093144	rs2290263
32	rs3093144	rs10197862	rs6543124		rs2853676	rs6994670
33	rs10245353	rs13097407	rs3093144		rs111853758	rs3093110
34	rs9807989	rs1484215	rs1861245		rs13097407	rs3093144
35	rs111853758	rs2420915	rs2287037		rs10208293	rs1861245
36	rs13097407	rs3771180	rs3771166		rs1861245	rs13097407
37	rs6543124	rs2853676	rs3093193		rs13271489	rs3771180
38	rs298207	rs2290263	rs2058622		rs3771166	rs13271489
39		rs3093144	rs6430491		rs3771180	rs3771175
40		rs3771175	rs10197862		rs10197862	rs10197862
41		rs13271489	rs3771180		rs9807989	rs9807989
42		rs10245353	rs13271489		rs6543124	rs483916
43						rs3771166
44						rs6543124

variant-to-gene mapping strategies implicate new genes and pathways influencing lung function and COPD risk.[24] Jingzhou Zhang reported that a polygenic risk score is associated with earlier age of diagnosis of COPD and retains predictive value when added to known early-life risk factors in 6647 non-Hispanic White (NHW) and 2464 African American (AA) participants.[25] Moreover, in 400,102 individuals of European ancestry, a new genetic signals for lung function highlight pathways and COPD associations across multiple ancestries.[13] Despite the advancements in COPD risk modeling, the majority of these studies have been centered on European populations. There are few studies on COPD risk models in Chinese Han population.

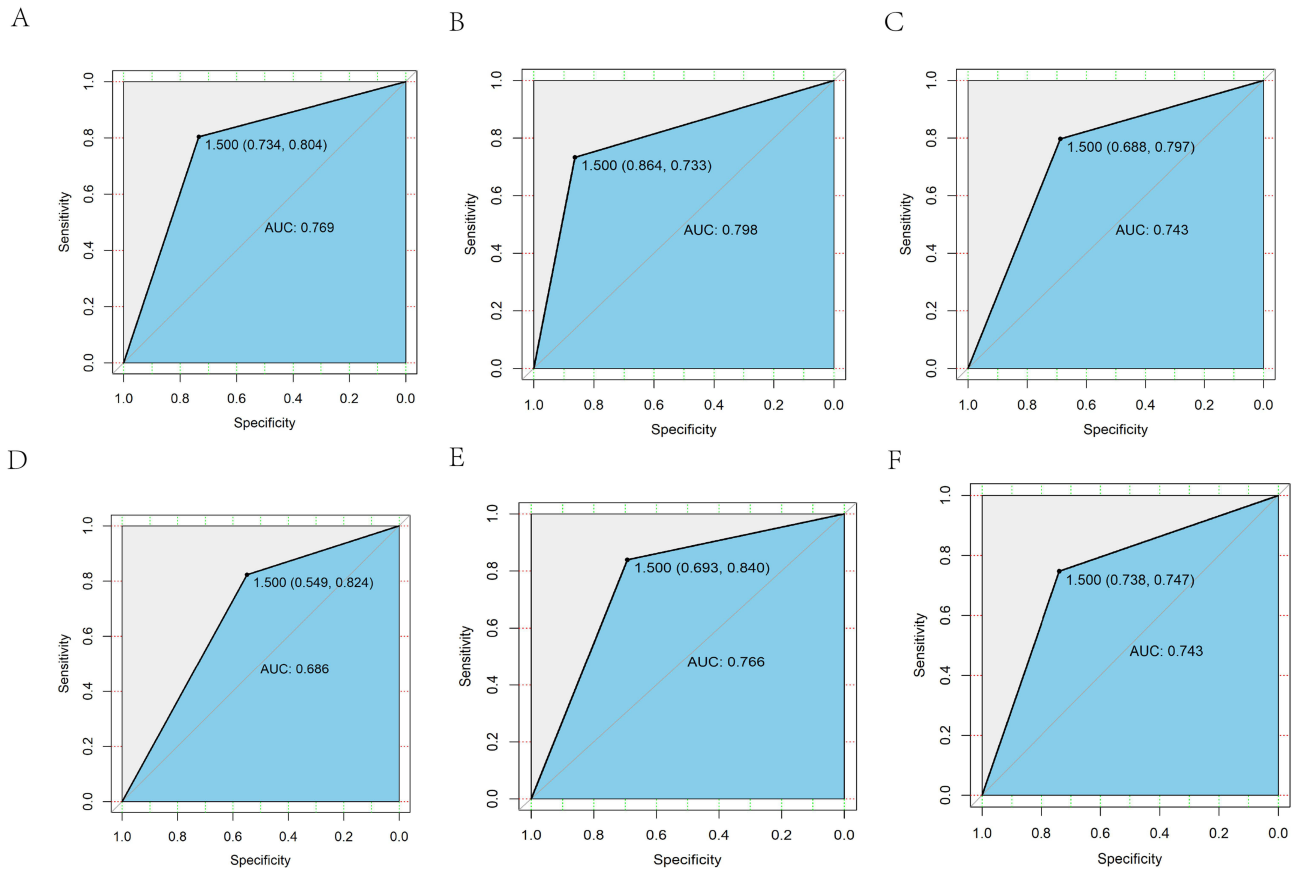

Figure 2 ROC curves for the six models of Recursive Feature Elimination (RFE). (**A**) ROC curves of 38 SNPs in caret model. AUC = 0.769. (**B**) ROC curves of 42 SNPs in Lda model. AUC = 0.798. (**C**) ROC curves of 42 SNPs in lr model. AUC = 0.734. (**D**) ROC curves of 4 SNPs in nb model. AUC = 0.686. (**E**) ROC curves of 42 SNPs in rf model. AUC = 0.766. (**F**) ROC curves of 44 SNPs in treebag model. AUC = 0.734.

In this study, we included SNPs that have been published as significant in association analyses for COPD. In total, we included 146 significant loci. On this basis, 233 patients diagnosed at Hainan Provincial People's Hospital and 290 healthy controls who underwent medical check-ups during the same period were screened using the Agena

Table 5 Random Forest Decision Results for 44 SNPs (MeanDecreaseGini Coefficients Represent the Importance of SNPs, Ranked from Most to Least)

Number	SNP_ID	Mean Decrease Gini	Number	SNP_ID	Mean Decrease Gini
1	rs1155002	7.50	23	rs6994670	4.58
2	rs352140	6.85	24	rs4803420	4.24
3	rs911186	6.44	25	rs3093203	4.15
4	rs1038376	6.26	26	rs2287037	4.12
5	rs3024622	6.03	27	rs4646437	4.05
6	rs9320913	5.83	28	rs2058622	3.73
7	rs10245353	5.69	29	rs2853676	3.36

(*Continued*)

Table 5 (Continued).

Number	SNP_ID	Mean Decrease Gini	Number	SNP_ID	Mean Decrease Gini
8	rs6430491	5.61	30	rs3093110	3.23
9	rs1420101	5.58	31	rs3093144	3.18
10	rs12479210	5.55	32	rs111853758	2.59
11	rs9525927	5.53	33	rs10208293	2.21
12	rs4719841	5.44	34	rs13097407	2.18
13	rs7934083	5.40	35	rs2290263	1.95
14	rs2420915	5.22	36	rs1861245	1.81
15	rs759648	5.10	37	rs3771166	1.38
16	rs78750958	5.02	38	rs9807989	1.26
17	rs2593704	5.01	39	rs10197862	1.20
18	rs298207	4.91	40	rs6543124	1.12
19	rs3093193	4.80	41	rs3771180	1.03
20	rs2505059	4.75	42	rs3771175	0.99
21	rs16907751	4.70	43	rs13271489	0.98
22	rs1484215	4.70	44	rs483916	0.64

Note: MeanDecreaseGini: The value indicates the relative importance of the variable from large to small, and is the total decrease in node impurity when splitting the variable averaged over all trees, with node impurity defined by the Gini coefficient.

MassARRAY technique in a case-control study method, and 44 SNPs were significantly associated with COPD susceptibility using one-way logistic regression analysis. The contribution of these 44 SNPs to the risk of COPD was then assessed using models constructed by LASSO, Caret, LDA, LR, NB, Rf and Treebag and the Random Forest model, comparing the classification performance of the different models and working to find a predictive model with higher performance.

LASSO is a regression analysis method that performs both variable selection and regularisation to improve the predictive accuracy and interpretability of statistical models.[26] An attractive feature for SNPs selection is the sparsity of the LASSO model and the shrinking of the regression coefficients, which can be effective in selecting SNPs that predict quantitative traits but are limited by certain conditions.[27] Jeremy Sabourin's study shows that the performance of LASSO-based RMA methods in distinguishing between multiple real signals and highly correlated SNPs can be continuously improved by randomising the penalty parameter.[28] In genomic studies, the ability to identify SNPs that affect a target trait is important for understanding the genetic basis of the trait.[29] Caret (Classification And REgression Training) is a powerful package for building, evaluating and comparing predictive models in the R language.[30] `Caret` provides a unified interface that makes it much easier to switch between algorithms.[31] On this basis, we used the RFE-Caret, RFE-Lda, RFE-lr, RFE-nb, RFE-rf and RFE-treebag algorithms to assess the risk of SNPs for COPD. In previous studies of SNPs, Caret has tuned models to select appropriate parameters to improve model accuracy. In the diabetes study, Quincy A Hathaway performed 10-fold cross-validation of the results using LDA, NB, Support Vector Machine (SVM) and Classification and Regression Tree (CART) models. The ultimate goal is to select the optimal model to determine the biomarkers of the disease.[32] Random forest models make predictions by constructing multiple decision trees and combining them together.[33] SNP data usually contains a large number of features, and Random Forest can

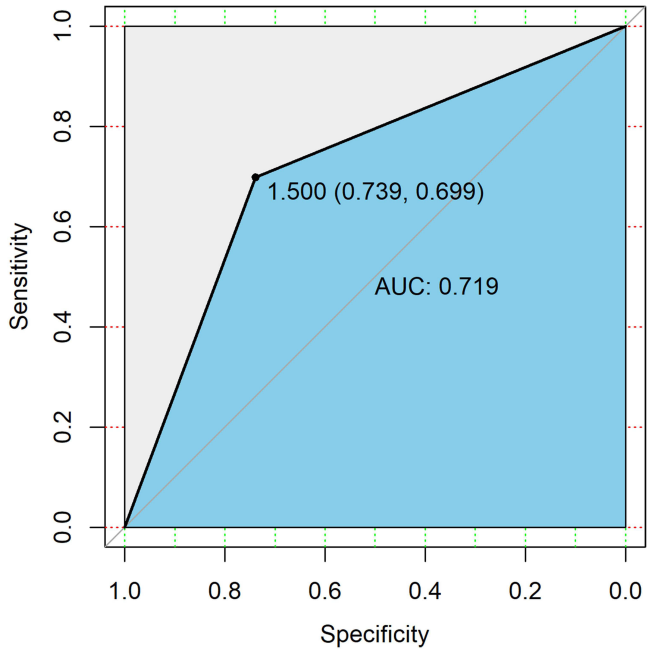

Figure 3 ROC curves of 44 SNPs for random forest model. AUC = 0.719.

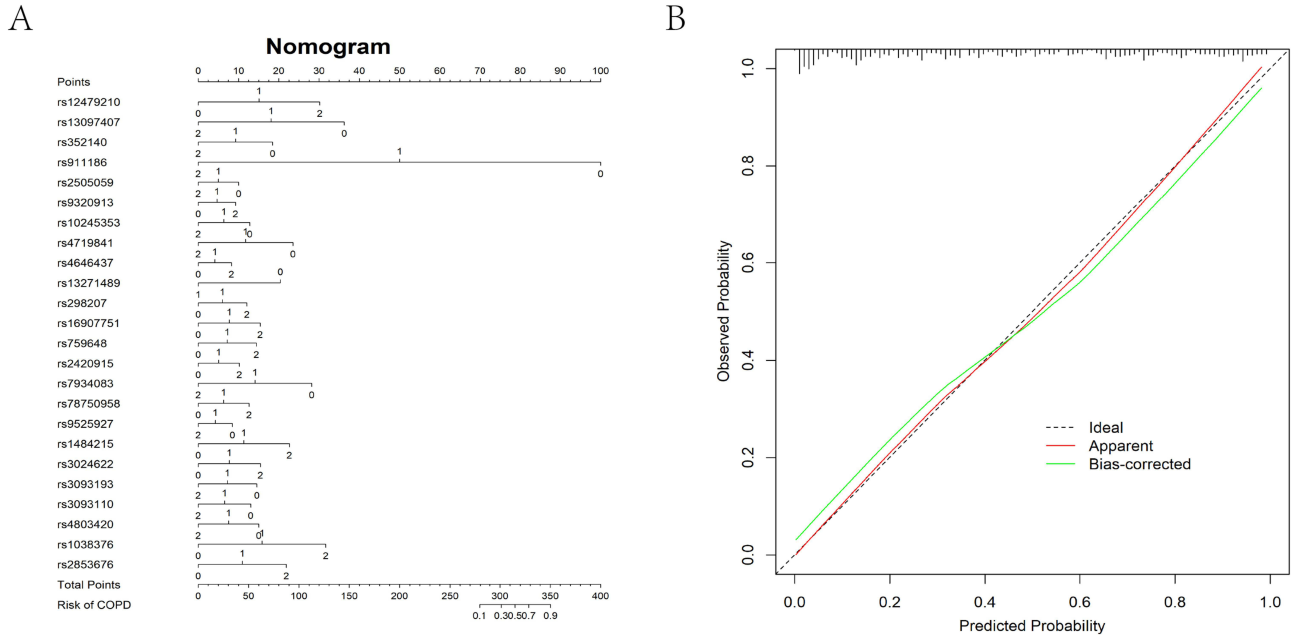

Figure 4 LASSO model Performance validation. (**A**) Nomogram Nomogram model predicting COPD risk. The nomogram is used by summing all points identified on the scale for each variable. (**B**) Curve of calibration for predicting COPD risk. The predicted Probability by the nomogram model is plotted on the x-axis, and the Observed Probability is plotted on the y-axis.

effectively deal with high-dimensional data to predict the most important SNPs in the dataset.[34] One study used random forest modelling to distinguish the ability of Parkinson's patients from controls.[35] The RF algorithm is trained on relevant data and the discriminative importance of individual SNPs is assessed by a technical construct known as graph depth.[36] As in our preliminary study, the predictive power of the tested SNPs was visualised and quantified using ROC curves and AUC, respectively.[37]

In addition to screening the best predictive models, we performed a column-line graphical model of the risk of incident COPD for the 25 independent predictors screened by the best model, lasso, and found that among the 25 high-risk SNPs, the rs1038376 and rs12479210 polymorphic loci contributed most to the increased risk of incident COPD. This result was crudely demonstrated in previous studies, where rs1038376 A/T and A/T-T/T/T were associated with an increased risk of COPD in co-dominant and dominant models, respectively, compared to the AA genotype.[38] Notably, rs12479210 was screened and strongly correlated with COPD in all of the above models, but no other study has yet clarified its association with COPD. Studies have shown that rs12479210, a candidate SNP for the IL-1RL1 gene, is significantly associated with lung cancer risk,[39] that IL-1RL1 is considered a targeted biomarker or target for pharmacological intervention in asthma,[40] and that people with COPD have a higher risk of lung cancer.[41] In conclusion, combining previous studies and our prediction results, we speculate that rs12479210 may be a potential risk locus for COPD.

However, our studies invariably have some limitations. On the one hand, although this was a case-control study, the study population was mostly from Hainan Province, China, so it would be cautious to generalise the conclusions or findings of this study to the general population. On the other hand, and we did not have external data to validate it, so we need to obtain more external data to further evaluate the nomogram constructed in this study.

Conclusions

In conclusion, based on the combination of single-factor analysis, LASSO regression, RFE algorithm and random forest model, 25 SNPs were screened to construct a simple prediction model with high predictive performance for COPD risk in the Chinese Han population.

Data Sharing Statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethical Approval

This study was conducted under the standards approved by the Ethics Committee of Hainan Provincial People's Hospital and was in accordance with the ethical principles of the World Medical Association Declaration of Helsinki for medical research involving humans. Informed consent was obtained from all individual participants included in this study.

Consent for Publication

Consent to publish statements must confirm that the details of any images, videos, recordings, etc can be published, and that the person(s) providing consent have been shown the article contents to be published.

Acknowledgments

We thank all members of our research team for their contributions to this study, as well as Hainan Provincial People's Hospital and all participants for their support to this study. We also thank National Natural Science Foundation of China for funding this study.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by Hainan Province Science and Technology Special Fund (No. ZDYF2024SHFZ094), the research project of Innovation Platform for Academicians of Hainan Province (YSPTZX202312), the Innovation

Platform for Academicians of Hainan Province and the Key Research and Development Program of National Natural Science Foundation of China (No. 81860015).

Disclosure

The authors declare no conflicts of interest in this work.

References

1. Global. regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the global burden of disease study 2015. *Lancet*. 2016;388(10053):1545–1602. doi:10.1016/S0140-6736(16)31678-6
2. Adeloye D, Chua S, Lee C, et al. Global and regional estimates of COPD prevalence: systematic review and meta-analysis. *J Global Health*. 2015;5 (2):020415. doi:10.7189/jogh.05.020415
3. Anees Ur R, Ahmad Hassali MA, Muhammad SA, et al. The economic burden of chronic obstructive pulmonary disease (COPD) in the USA, Europe, and Asia: results from a systematic review of the literature. *Expert Rev Pharmacoecon Outcomes Res*. 2020;20(6):661–672. doi:10.1080/14737167.2020.1678385
4. AS B, MA M, WM V, et al. International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *Lancet*. 2007; 370(9589):741–750.
5. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet*. 2007; 370(9589):765–773.
6. KF R, Hurd S, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med*. 2007;176(6):532–555. doi:10.1164/rccm.200703-456SO
7. Wang CD, Chen N, Huang L, et al. Impact of CYP1A1 polymorphisms on susceptibility to chronic obstructive pulmonary disease: a meta-analysis. *Biomed Res Int*. 2015;2015:942958. doi:10.1155/2015/942958
8. Humbert M, Montani D, Perros F, Dorfmüller P, Adnot S, Eddahibi S. Endothelial cell dysfunction and cross talk between endothelium and smooth muscle cells in pulmonary arterial hypertension. *Vasc Pharmacol*. 2008;49(4–6):113–118. doi:10.1016/j.vph.2008.06.003
9. Yuksel H, Yilmaz O, Karaman M, et al. Role of vascular endothelial growth factor antagonism on airway remodeling in asthma. *Annals Allergy Asthma Immunol*. 2013;110(3):150–155. doi:10.1016/j.anai.2012.12.015
10. Marciniak SJ, Lomas DA. Genetic susceptibility. *Clinics Chest Med*. 2014;35(1):29–38. doi:10.1016/j.ccm.2013.10.008
11. BD H, de Jong K, Lamontagne M, et al. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nature Genet*. 2017;49(3):426–432. doi:10.1038/ng.3752
12. Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nature Genet*. 2019;51(3):494–505. doi:10.1038/s41588-018-0342-2
13. Shrine N, AL G, AM E, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet*. 2019;51(3):481–493. doi:10.1038/s41588-018-0321-7
14. Wu MC, Kraft P, Epstein MP, et al.:Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010;86 (6):929–942. doi:10.1016/j.ajhg.2010.05.002
15. Omranian N, Eloundou-Mbebi JM, Mueller-Roeber B, Nikoloski Z. Gene regulatory network inference using fused LASSO on multiple data sets. *Sci Rep*. 2016;6:20533. doi:10.1038/srep20533
16. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*. 2018;15(1):41–51.
17. Yu X, Zeng Q. Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes. *Aquatic Toxicol*. 2022;251:106265. doi:10.1016/j.aquatox.2022.106265
18. Lamontagne M, JC B, Obeidat M, et al. Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. *Human Mol Gene*. 2018;27(10):1819–1829. doi:10.1093/hmg/ddy091
19. Ma X, Wu Y, Zhang L, et al. Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population. *J Transl Med*. 2020;18(1):146. doi:10.1186/s12967-020-02312-0
20. Gim J, An J, Sung J, Silverman EK, Cho MH, Won S. A between ethnicities comparison of chronic obstructive pulmonary disease genetic risk. *Front Genet*. 2020;11:329. doi:10.3389/fgene.2020.00329
21. Zhou Y, Chen J, Bai F, et al. Suggestive evidence of genetic association of IL23R polymorphisms with chronic obstructive pulmonary disease risk in the Chinese population. *J Gene Med*. 2023;25(5):e3479. doi:10.1002/jgm.3479
22. Zhou JJ, Cho MH, Castaldi PJ, Hersh CP, Silverman EK, Laird NM. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am J Respir Crit Care Med*. 2013;188(8):941–947. doi:10.1164/rccm.201302-0263OC
23. Moll M, Sakornsakolpat P, Shrine N, et al. Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. *Lancet Respir Med*. 2020;8(7):696–708. doi:10.1016/S2213-2600(20)30101-6
24. Shrine N, AG I, Chen J, et al. Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. *Nat Genet*. 2023;55(3):410–422. doi:10.1038/s41588-023-01314-0
25. Zhang J, Xu H, Qiao D, et al. A polygenic risk score and age of diagnosis of COPD. *Eur Respir J*. 2022;60(3):2101954. doi:10.1183/13993003.01954-2021
26. Kang J, Choi YJ, Kim IK, et al. LASSO-based machine learning algorithm for prediction of lymph node metastasis in T1 colorectal cancer. *Cancer Res Treat*. 2021;53(3):773–783. doi:10.4143/crt.2020.974
27. Feng ZZ, Yang X, Subedi S, McNicholas PD. The LASSO and sparse least square regression methods for SNP selection in predicting quantitative traits. *IEEE/ACM trans comput biol bioinfo*. 2012;9(2):629–636. doi:10.1109/TCBB.2011.139
28. Yang C, Wan X, Yang Q, Xue H, Yu W. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. *BMC Bioinf*. 2010;1(1 Suppl):S18. doi:10.1186/1471-2105-11-S1-S18

29. Wang H, Zhang Y, Chen L, et al. Identification of clinical prognostic features of esophageal cancer based on m6A regulators. *Front Immunol*. 2022;13:950365. doi:10.3389/fimmu.2022.950365
30. Beck MW. NeuralNetTools: visualization and analysis tools for neural networks. *J Stat Software*. 2018;85(11):1–20. doi:10.18637/jss.v085.i11
31. TM D, Dankers F, Valdes G, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. *Med Phys*. 2018;45(7):3449–3459. doi:10.1002/mp.12967
32. QA H, SM R, MV P, et al. Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. *Cardiovasc diabetol*. 2019;18(1):78. doi:10.1186/s12933-019-0879-0
33. Elbeltagi A, Pande CB, Kumar M, et al. Prediction of meteorological drought and standardized precipitation index based on the random forest (RF), random tree (RT), and Gaussian process regression (GPR) models. *Environ Sci Pollut Res Int*. 2023;30(15):43183–43202. doi:10.1007/s11356-023-25221-3
34. Botta V, Louppe G, Geurts P, Wehenkel L. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One*. 2014;9(4):e93379. doi:10.1371/journal.pone.0093379
35. Cibulka M, Brodnanova M, Grendar M, et al. Alzheimer's disease-associated SNP rs708727 in SLC41A1 may increase risk for parkinson's disease: report from enlarged Slovak study. *Int J Mol Sci*. 2022;23(3):1604. doi:10.3390/ijms23031604
36. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–1930. doi:10.1161/CIRCULATIONAHA.115.001593
37. Cibulka M, Brodnanova M, Grendar M, et al. SNPs rs11240569, rs708727, and rs823156 in SLC41A1 do not discriminate between Slovak patients with idiopathic parkinson's disease and healthy controls: statistics and machine-learning evidence. *Int J Mol Sci*. 2019;20(19):4688. doi:10.3390/ijms20194688
38. Ding Y, Li Q, Feng Q, et al. CYP2B6 genetic polymorphisms influence chronic obstructive pulmonary disease susceptibility in the Hainan population. *Int J Chronic Obstr*. 2019;14:2103–2115. doi:10.2147/COPD.S214961
39. Li Q, Zhang C, Cheng Y, et al. IL1RL1 polymorphisms rs12479210 and rs1420101 are associated with increased lung cancer risk in the Chinese Han population. *Front Genetics*. 2023;14:1183528. doi:10.3389/fgene.2023.1183528
40. Saikumar Jayalatha AK, Hesse L, Ketelaar ME, Koppelman GH, Nawijn MC. The central role of IL-33/IL-1RL1 pathway in asthma: from pathogenesis to intervention. *Pharmacol Ther*. 2021;225:107847. doi:10.1016/j.pharmthera.2021.107847
41. Forder A, Zhuang R, VGP S, et al. Mechanisms contributing to the comorbidity of COPD and lung cancer. *Int J Mol Sci*. 2023;24(3):2859. doi:10.3390/ijms24032859