BMC Pulmonary Medicine

RESEARCH

Open Access

Integrating genetic and clinical data to predict lung cancer in patients with chronic obstructive pulmonary disease

Zhan Gu[1], Yonghui Wu[1], Fengzhi Yu[1], Jijia Sun[2] and Lixin Wang[1*]

Abstract

Background Chronic obstructive pulmonary disease (COPD) is closely linked to lung cancer (LC) development. The aim of this study is to identify the genetic and clinical risk factors for LC risk in COPD, according to which the prediction model for LC in COPD was constructed.

Methods This is a case-control study in which patientis with COPD + LC as the case group, patientis with only COPD as the control group, and patientis with only LC as the second control group. A panel of clinical variables including demographic, environmental and lifestyle factors were collected. A total of 20 single nucleotide polymorphisms (SNPs) were genotyped. The univariate analysis, candidate gene study and multivariate analysis were applied to identify the independent risk factors, as well as the prediction model was constructed. The ROC analysis was used to evaluate the predictive ability of the model.

Results A total of 503 patients were finally enrolled in this study, with 188 patients for COPD + LC group, 162 patients for COPD group and 153 patients for LC group. The univariate analysis of clincial data showed compared with the patients with COPD, the patients with COPD + LC tended to have significantly lower BMI, higher smoking pack-years, and higher prevalence of emphysema. The results of the candidate gene study showed the rs1489759 in *HHIP* and rs56113850 in *CYP2A6* demonstrated significant differences between COPD and COPD + LC groups. By using multivariate logistic regression analysis, four variables including BMI, pack-years, emphysema and rs56113850 were identified as independent risk factors for LC in COPD and the prediction model integrating genetic and clinical data was constructed. The AUC of the prediction model for LC in COPD reached 0.712, and the AUC of the model for predicting LC in serious COPD reached up to 0.836.

Conclusion The rs56113850 (risk allele C) in *CYP2A6*, decrease in BMI, increase in pack-years and emphysema presence were independent risk factors for LC in COPD. Integrating genetic and clinical data for predicting LC in COPD demonstrated favorable predictive performance.

Keywords Lung cancer, Chronic obstructive pulmonary disease, Single nucleotide polymorphisms, Risk factors, Prediction model

*Correspondence:
Lixin Wang
1701005@tongji.edu.cn
[1]Department of Integrative Medicine, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China
[2]Department of Mathematics and Physics, Shanghai University of Traditional Chinese Medicine, Shanghai, China

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Lung cancer (LC) is one of the most common cancers worldwide with high prevalence and poor prognosis. As the leading cause of cancer deaths, LC resulted in 1.8 million deaths globally in 2020 [1]. Chronic obstructive pulmonary disease (COPD) is a respiratory illness that is characterized by chronic inflammation and irreversible airway obstruction. The incidence of COPD continues to rise and COPD ranks the third leading cause of mortality globally [2, 3]. LC is a common comorbid disease among patients with COPD, and the coexistence of COPD and LC is now becoming a serious public health concern. COPD is closely linked to LC and is recognized as a main risk factor for LC development, independently of tobacco exposure [4, 5]. Compared to people with normal lung function, patients with COPD are 4–10 times more likely to develop LC [6, 7]. Consequently, early identification and treatment of LC in patients with COPD is urgently neeeded.

COPD and LC are both complex diseases that result from the combined effects of genetic susceptibility and environmental factors, involved multiple pathophysiological mechanisms especially for chronic inflammation. In the past years, increasing numbers of single nucleotide polymorphisms (SNPs) might associated with both COPD and LC have been reported by genome-wide association studies (GWAS) and candidate gene studies [8–11]. Several of these loci and candidate genes may be involved in the LC development in COPD, including hedgehog interacting protein (*HHIP*), cholinergic receptor, neuronal nicotinic, α-polypeptide 3 (*CHRNA3*), *CHRNA5*, glycophorin A (*GYPA*), cytochrome P450 2A6 (*CYP2A6*), *CYP1A1*, matrix metalloproteinase-1 (*MMP-1*) among others. Besides genetic predisposition, demographic and environmental factors like advanced age, weight loss, smoking exposure, air pollution, dust exposure, a history of previous lung disease, and underlying inflammatory processes all potentially contribute to an increased susceptibility to developing LC in patients with COPD [12]. Nevertheless, few studies have comprehensively evaluated the synergistic effect of multiple risk factors including genetic and clinical variables on increasing the risk of LC in COPD.

The aim of this study is: (1) to evaluate a panel of clinical variables including demographic, environmental and lifestyle factors among three groups of patients (COPD without LC [COPD group], coexistence of COPD and LC [COPD+LC group], and LC without COPD [LC group]); (2) to perform a candidate gene study to investigate the susceptible gene polymorphisms for LC in COPD, focusing on the comparison between COPD and COPD+LC groups; and (3) to select the risk factors and construct a prediction model for LC in patients with COPD and serious COPD.

Methods

Study design and study subjects

This is a case-control study in which COPD+LC group as the case group, COPD group as the control group, and LC group as the second control group. All study subjects with informed consent were enrolled from January 2020 to December 2022 at Shanghai Pulmonary Hospital in Shanghai, China. This study was performed in accordance with the guidelines of the Helsinki Declaration (as revised in 2013) and was approved by the Ethics Committee of Shanghai Pulmonary Hospital affiliated to Tongji University (K20-036Z).

COPD was defined following the Global Initiative for Obstructive Lung Disease (GOLD) recommendations as the presence of persistent respiratory symptoms and the ratio of a forced expiratory volume in first second (FEV1) and forced vital capacity (FVC) was less than 0.70 after a bronchodilator test [13]. The severity of COPD was classified according to the criteria of GOLD. The diagnosis of LC was made after suggestive radiological findings with pathologic confirmation through histological or cytological specimens [14]. Of note, the patients with LC (COPD+LC group and LC group) enrolled in this study were all newly diagnosed LC cases. In the enrollment of COPD+LC group, once the patients were newly diagnosed with LC, then the history of COPD and lung function were detected to ensure that the LC occurrence was secondary to COPD.

The inclusion criteria included physician-diagnosed COPD or LC, over the age of 40, Chinese Han subjects who were not relatred to each other, and complete data measurements. Subjects who had a history of asthma, tuberculosis and interstitial lung diseases that may affect lung function, or had a history of other malignancies or cardiovascular and cerebrovascular diseases were all excluded (Fig. 1).

Clinical data collection

All clinical variables including demographic, environmental and lifestyle factors were collected and recorded into the self-developed "iLUNG TCM" health management software system (Registration ID: 2022SR0999226). The collected clinical data included gender, age, body mass index (BMI), dust exposure, alcohol use, smoking status, physical activity, lung function and tumor characteristics. The dust exposure referred to the occupational exposure to any of the four dusts of silica, cement, coal and asbestos. The smoking status were categorized into three groups: non-smoker, former smoker, and active smoker. The smoking pack-years were calculated as well. The physical activity included exercise and pulmonary rehabilitation. The lung function was detected by masterlab spirometers (Jaeger, Wuezburg, Germany) by technicians specialized in respiratory functional tests.

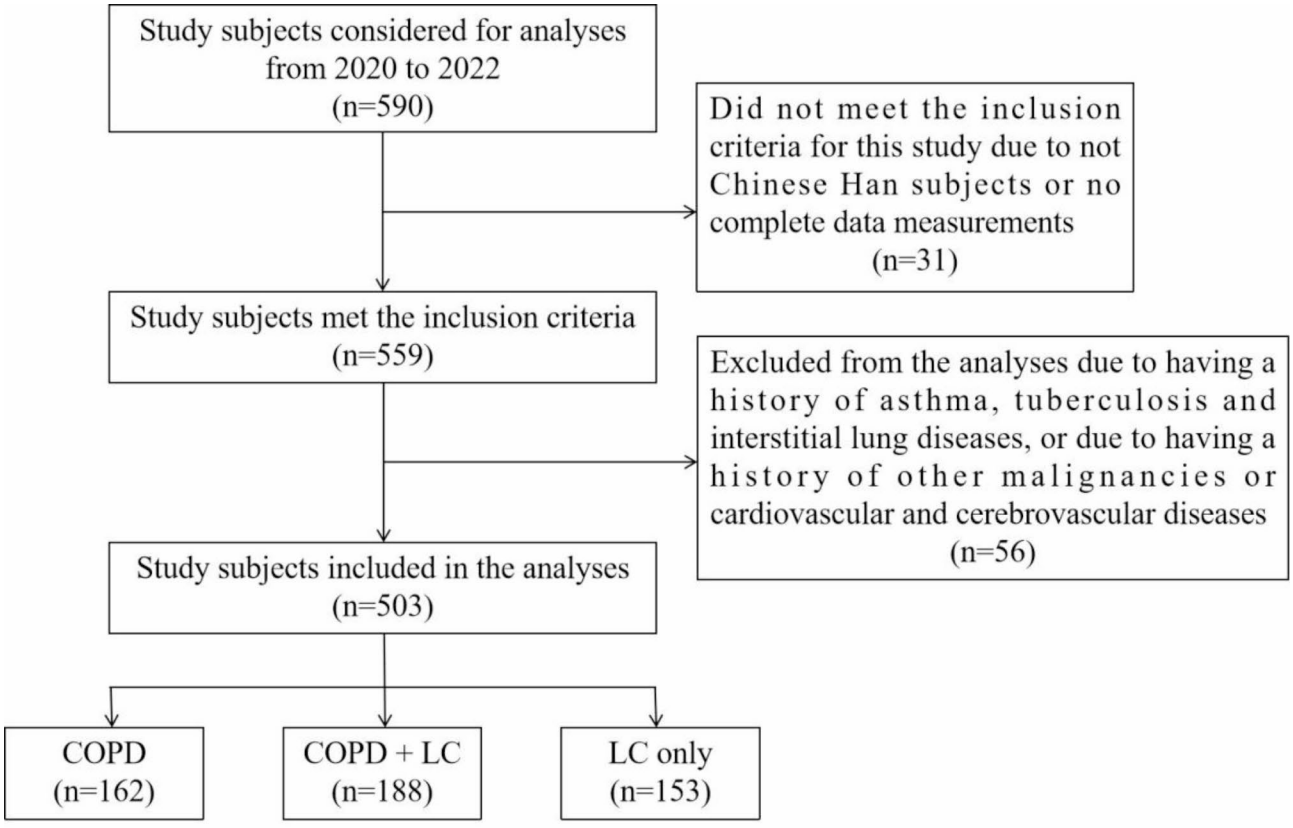

Fig. 1 Flowchart of the study

Emphysema was determined through computed tomography (CT) assessment by professional radiologists. The tumor characteristics included pathological types and the stage at diagnosis according to the TNM eight edition of LC.

SNPs selection and genotyping

The SNPs might involved in the LC development in COPD were selected from previous published GWAS or candidate gene studies [8–11, 15–19], and several SNPs associated with activation of inflammatory pathways were added [20–22]. A total of 20 SNPs were finally selected in this study and the basic information are listed in Table 1 according to the dbSNP database (https://www.ncbi.nlm.nih.gov/snp) and the 1000 Genomes Project (https://www.internationalgenome.org/).

Genomic DNA was extracted from all subjects from 1.5 mL of peripheral blood samples collected in EDTA collection tubes. DNA was extracted using a commercial extraction kit (Tiangen biotechnology, Beijing, China) for genotyping. All 20 SNPs were successfully genotyped by matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometry using the MassARRAY® analyzer platform (Sequenom, San Diego, CA) by Shanghai Fenglin Clinical Laboratory. Detailed information regarding the primers is available (Supplementary material 1).

Statistical analysis

All statistical analyses and graphs were performed by SPSS version 23.0 (IBM, Armonk, USA), SHEsis platform (http://analysis.bio-x.cn/myAnalysis.php) (Bio-X Institutes, Shanghai, China) [23] and MedCalc version 22.001 (MedCalc, Ostend, Belgium). A descriptive analysis of all clinical variables among three groups of patients was carried out. Normally distributed data were expressed as mean±standard deviation (SD) and compared using Student's t tests. Non-normally distributed data were expressed as median (first quartile and third quartile (Q1, Q3)) and compared using Mann-Whitney U tests. Categorical data were expressed as count (percentage) and compared using Pearson's χ^2 tests. The genotype frequency distribution of SNPs was assessed using Pearson's χ^2 tests. The genetic association analysis was using logistic regression to calculate odds ratios (ORs) and 95% confidence intervals (CIs). Univariate logistic regression and multivariate logistic regression analyses were applied to identify the independent risk factors for LC in COPD, according to which the prediction model for LC in COPD was constructed. To evaluate the predictive ability and

Table 1 Basic information of selected SNPs

SNPs	Gene	Allele	Location	Function
rs7689420	*HHIP*	C/T	4:144647200	intron
rs1489759	*HHIP*	A/G	4:144553321	intron
rs10519717	*HHIP*	T/C	4:144559188	intron
rs6495309	*CHRNA3*	C/T	15:78622903	upstream
rs12910984	*CHRNA3*	A/G	15:78599285	intron
rs2202507	*GYPA*	T/G	4:144336529	intron
rs56113850	*CYP2A6*	T/C	19:40847202	intron
rs7326277	*FLT1*	T/C	13:28302077	3′UTR
rs2072493	*TLR5*	T/C	1:223111257	missense
rs1356888	*NRXN1*	C/T	2:50288880	intron
rs9224	*FAM13A*	G/A	4:88728508	3′UTR
rs4796793	*STAT3*	C/G	17:42390192	upstream
rs3744483	*STAT3*	T/C	17:42314420	3′UTR
rs1053005	*STAT3*	T/C	17:42313892	3′UTR
rs2293152	*STAT3*	C/G	17:42329511	intron
rs696	*NFKBIA*	C/T	14:35401887	3′UTR
rs8904	*NFKBIA*	G/A	14:35402011	3′UTR
rs2273650	*NFKBIA*	C/T	14:35401592	3′UTR
rs1799964	*TNF*	T/C	6:31574531	upstream
rs1800630	*TNF*	C/A	6:31574699	upstream

SNPs: single nucleotide polymorphisms; HHIP: hedgehog interacting protein; CHRNA3: cholinergic receptor, neuronal nicotinic, α-polypeptide 3; GYPA: glycophorin A; CYP2A6: cytochrome P450 2A6; FLT1: vascular endothelial growth factor receptor 1; TLR5: toll like receptor 5; NRXN1: neurexin 1; FAM13A: family with sequence similarity 13 member A; STAT3: signal transducer and activator of transcription 3; NFKBIA: nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor alpha; TNF: tumor necrosis factor

robustness of the model, receiver operating characteristic curve (ROC) analysis were used and the areas under ROC (AUCs) were calculated. For all of the above analyses, *P* value < 0.05 was considered statistically significant.

Results

Baseline information and clinical data of the study subjects

As shown in Fig. 1, a total of 503 patients were finally enrolled in this study, with 188 patients for COPD + LC group, 162 patients for COPD group and 153 patients for LC group. Baseline information and clinical data among three groups of patients are presented in Table 2. Compared with the patients with COPD, the patients with COPD + LC tended to have significantly lower BMI, higher smoking pack-years, and higher prevalence of emphysema ($P<0.05$). There were no differences in gender, age, dust exposure, alcohol use, physical activity, and respiratory functional parameters between COPD and COPD + LC groups. Compared with the patients with LC, the patients with COPD + LC had significantly lower non-smokers, higher smoking pack-years, and higher prevalence of emphysema ($P<0.001$).

The most frequent tumor pathological type in COPD + LC group was squamous (48.94%), followed by adenocarcinoma (34.04%), small cell LC (13.30%) and other types (3.72%). The prevalence of squamous was higher and the prevalence of adenocarcinoma was lower in COPD + LC group than that in LC group ($P<0.05$). There were no differences in stage at diagnosis.

Genotype frequency distribution and genetic association analysis

The genotyping call rates for each SNP reached 100% and the minor allele frequency (MAF) of each SNP were greater than 0.05. The genotype frequency distribution of each SNP between COPD and COPD + LC groups are given in Table 3. Of all 20 SNPs, the rs1489759 in *HHIP* and rs56113850 in *CYP2A6* demonstrated significant differences in genotype frequency distribution ($P<0.05$).

The rs1489759 and rs56113850 were further analyzed. As presented in Table 4, the genetic association analysis showed rs1489759 (risk allele A) increased the risk of LC in COPD in recessive and additive models ($P<0.05$), and rs56113850 (risk allele C) increased the risk of LC in COPD in dominant, recessive and additive models ($P<0.05$). Additionally, rs56113850 increased the disease risk in COPD + LC group compared with LC group ($P<0.001$), whereas there was no significant deviation of rs1489759 between COPD + LC and LC groups.

Risk factors and prediction model

Based on the results of the univariate analysis and candidate gene study, BMI, smoking pack-years, emphysema, rs1489759 and rs56113850 were included in the next multivariate logistic regression analysis to explore the risk factors for LC in COPD. Notably, rs1489759 and rs56113850 were analyzed using additive models. Ultimately, four variables including BMI, smoking pack-years, emphysema and rs56113850 were identified as independent risk factors for LC in COPD and the prediction model integrating genetic and clinical data was constructed. The ORs and 95%CIs of risk factors were shown in Table 5.

Predictive ability evaluation

ROC analysis for the model to predict LC in COPD is presented in Fig. 2. The AUC of the prediction model integrating genetic and clinical data reached 0.712 (95%CI: 0.661–0.759, $P<0.001$), with the sensitivity was 48.94% and specificity was 84.57%. Additionally, the Hosmer-Lemeshow test demonstrated that the model was a good fit (*P* value: 0.516 > 0.05). If the prediction model only including clinical data, the AUC was only 0.695 (95%CI: 0.644–0.743, $P<0.001$).

For serious COPD patients (FEV1% < 50%), the prediction model including BMI, smoking pack-years, emphysema and rs56113850 was also constructed and ROC analysis is presented in Fig. 3. The AUC of the model for predicting LC in serious COPD reached up to 0.836

Table 2 Baseline information and clinical data of the study subjects

Variables	COPD (*n* = 162)	LC only (*n* = 153)	COPD + LC (*n* = 188)	P_1	P_2
Baseline characteristics					
Male / female, n	148/14	127/26	167/21	0.432	0.121
Age(year), Mean ± SD	69.04 ± 7.33	67.51 ± 7.18	68.30 ± 7.26	0.345	0.314
BMI(kg/m^2), Median (Q1, Q3)	24.49 (22.43, 26.52)	23.05 (20.76, 25.07)	23.02 (21.06, 24.98)	< 0.001	0.910
Dust exposure, n (%)	22 (13.58)	19 (12.42)	17 (9.04)	0.179	0.313
Alcohol use, n (%)	88 (54.32)	73 (47.71)	92 (48.94)	0.315	0.822
Smoking status					
Non-smoker, n (%)	16 (9.88)	36 (23.53)	12 (6.38)	0.230	< 0.001
Former smoker, n (%)	86 (53.09)	65 (42.48)	94 (50.00)	0.565	0.166
Active smoker, n (%)	60 (37.04)	52 (33.99)	82 (43.62)	0.211	0.070
Pack-years, Mean ± SD	35.64 ± 15.66	32.97 ± 20.97	40.93 ± 16.30	0.002	< 0.001
Physical activity					
Exercise, n (%)	42 (25.93)	36 (23.53)	40 (21.28)	0.306	0.619
Pulmonary rehabilitation, n (%)	16 (9.88)	10 (6.54)	11 (5.85)	0.159	0.794
Lung function					
FEV1%, Mean ± SD	66.40 ± 18.37	99.54 ± 8.49	65.04 ± 18.44	0.491	< 0.001
FEV1 / FVC%, Median (Q1, Q3)	63.10 (56.06, 67.43)	83.94 (78.79, 88.27)	63.53 (55.41, 68.13)	0.667	< 0.001
GOLD I-II / III-IV, n	135/27	/	150/38	0.395	/
Emphysema, n (%)	82 (50.62)	24 (15.69)	131 (69.68)	< 0.001	< 0.001
Tumor characteristics					
Adenocarcinoma, n (%)	/	71 (46.41)	64 (34.04)	/	0.020
Squamous, n (%)	/	45 (29.41)	92 (48.94)	/	< 0.001
SCLC, n (%)	/	28 (18.30)	25 (13.30)	/	0.205
Other types, n (%)	/	9 (5.88)	7 (3.72)	/	0.348
TNM stage I-II / III-IV, n	/	37/116	43/145	/	0.776

P_1: COPD + LC group vs. COPD group; P_2: COPD + LC group vs. LC group

COPD: chronic obstructive pulmonary disease; LC: lung cancer; SD: standard deviation; BMI: body mass index; FEV1: forced expiratory volume in first second; FVC: forced vital capacity; GOLD: Global Initiative for Obstructive Lung Disease; SCLC: small cell lung cancer

Table 3 Genotype frequency distribution of each SNP

SNPs	Genotype	COPD	COPD + LC	*P*
rs7689420	CC/CT/TT	55/74/33	64/92/32	0.698
rs1489759	AA/AG/GG	58/91/13	94/81/13	0.027
rs10519717	TT/TC/CC	81/65/16	97/74/17	0.942
rs6495309	CC/CT/TT	57/71/34	56/94/38	0.469
rs12910984	AA/AG/GG	49/89/24	53/98/37	0.487
rs2202507	TT/TG/GG	48/86/28	57/91/40	0.576
rs56113850	TT/TC/CC	63/63/36	39/87/62	< 0.001
rs7326277	TT/TC/CC	74/69/19	81/82/25	0.850
rs2072493	TT/TC/CC	85/66/11	107/67/14	0.618
rs1356888	CC/CT/TT	118/34/10	140/41/7	0.567
rs9224	GG/GA/AA	110/41/11	129/51/8	0.564
rs4796793	CC/CG/GG	71/76/15	80/82/26	0.408
rs3744483	TT/TC/CC	73/80/9	81/90/17	0.462
rs1053005	TT/TC/CC	67/71/24	79/93/16	0.163
rs2293152	CC/CG/GG	42/85/35	62/96/30	0.225
rs696	CC/CT/TT	57/75/30	59/97/32	0.610
rs8904	GG/GA/AA	54/76/32	58/98/32	0.607
rs2273650	CC/CT/TT	84/66/12	95/83/10	0.647
rs1799964	TT/TC/CC	101/50/11	103/76/9	0.159
rs1800630	CC/CA/AA	95/54/13	98/80/10	0.168

SNPs: single nucleotide polymorphisms; COPD: chronic obstructive pulmonary disease; LC: lung cancer

(95%CI: 0.723–0.916, $P < 0.001$), with the sensitivity was 86.84% and specificity was 77.78%.

Discussion

The association between COPD and LC has garnered significant attention from researchers and clinicians in recent years. At a rate of 0.8–1.7% of patients with COPD develop LC per year and LC accounts for 33% of all COPD-related deaths [24]. There is a growing interest in identifying LC risk in patients with COPD to realize effectively management. In this study, we evaluated the association of both genetic susceptibility and clinical variables including demographic, environmental and lifestyle factors with LC in COPD. By using univariate analysis, candidate gene study and multivariate analysis, BMI, smoking pack-years, emphysema and rs56113850 in *CYP2A6* were ultimately identified as independent risk factors and applied to predict LC in patients with COPD and serious COPD. Moreover, besides the patients with coexistence of COPD and LC, two control groups (patients with only COPD and patients with only LC) were set in this study to comprehensively assess the association of some variables with COPD + LC. To the best of

Table 4 Genetic association analysis of rs1489759 and rs56113850

SNPs	Genotype	COPD + LC vs. COPD			COPD + LC vs. LC only		
		OR	95%CI	*P*	OR	95%CI	*P*
rs1489759	GA-AA vs. GG	1.174	0.528–2.612	0.693	1.146	0.507–2.589	0.744
	AA vs. GG-GA	1.793	1.166–2.756	0.008	0.843	0.550–1.293	0.435
	AA vs. GA vs. GG	1.503	1.063–2.125	0.021	0.919	0.653–1.292	0.627
rs56113850	TC-CC vs. TT	2.431	1.514–3.903	< 0.001	3.871	2.409–6.218	< 0.001
	CC vs. TT-TC	1.722	1.067–2.781	0.026	3.470	1.965–6.128	< 0.001
	CC vs. TC vs. TT	1.677	1.258–2.236	< 0.001	2.602	1.893–3.577	< 0.001

SNPs: single nucleotide polymorphisms; COPD: chronic obstructive pulmonary disease; LC: lung cancer; OR: odds ratio; CI: confidence interval

Table 5 Multivariate analysis to determine the independent risk factors for LC in COPD

Variables	OR	95%CI	*P*
BMI	0.861	0.798–0.930	< 0.001
Pack-years	1.022	1.007–1.037	0.004
Emphysema	1.612	1.010–2.574	0.045
rs56113850	1.684	1.240–2.286	0.001
Intercept	/	/	0.035

LC: lung cancer; COPD: chronic obstructive pulmonary disease; OR: odds ratio; CI: confidence interval; BMI: body mass index

our knowledge, this is the first study to integrate genetic and clinical data to predict LC in patients with COPD.

The chromosomal region 19q13.2 contains the primary nicotine metabolizing gene, *CYP2A6* [25]. The *CYP2A6* gene is a highly polymorphic enzyme that metabolizes nicotine to cotinine, then cotinine to trans-3'-hydroxycotinine (3HC), and the nicotine metabolite ratio (3HC/cotinine) means the efficacy of nicotine metabolism through *CYP2A6*. It is demonstrated that the genetic variants in *CYP2A6* are associated with nicotine metabolism, smoking behavior, smoking cessation, tobacco-related LC risk [26]. As slower nicotine metabolism, *CYP2A6* activity variation may influence LC risk via tobacco exposure and procarcinogen activation [27]. The rs56113850, the sentinel associated SNP in *CYP2A6*, was reported the effect allele C was associated with increased nicotine metabolism activity [28, 29]. A GWAS summary statistics analysis showed the results that the effect allele C of rs56113850 was associated with an increased risk of heavier smoking, COPD and LC [16]. A recent single-variant and Mendelian randomization analysis acknowledged the causal pathway connecting rs56113850 in *CYP2A6*, cigarette consumption, and LC susceptibility in smokers [30]. These results all supported our findings. In the present study, rs56113850 (risk allele C) was firstly discovered increased the risk of LC in COPD in Chinese Han population after adjustment for confounding factors. Compared with COPD group or LC group, rs56113850 all increased the disease risk in COPD+LC group. Our results suggested rs56113850 in *CYP2A6* might be a potential genetic biomarker of LC in COPD.

The rs1489759 in *HHIP* was found associated with LC in COPD in unadjusted models in this study, but this association was no longer significant after adjustment. The *HHIP* gene locates on chromosome 4q31 locus and acts as a negative regulator of Hedgehog signaling by binding to Hedgehog protein. Hedgehog signaling pathway is involved in epithelial-mesenchymal transition, mediates cigarette induced oncogenic transformation of bronchial epithelial cells and is essential for cellular proliferation of many LC cell lines. In addition to this, the expression of *HHIP* may cause changes in lung repair mechanisms and then lead to COPD [15, 31]. It has also been reported the rs1489759 might associated with both COPD and LC in several studies [15, 32]. Of note, some of the SNPs identified associated with LC in COPD in other studies were not significant in this study. The differences in the results with other findings may be due to the ethnic and regional distributional differences of the samples.

Regarding clinical data, our results showed decrease in BMI, increase in smoking pack-years and emphysema presence were independent risk factors for LC in COPD. BMI is a measurement of body fat based on height and weight. Mounting evidences have suggested that BMI is inversely associated with the risk of LC [33, 34]. A recent pooled analysis of 10 prospective cohort studies indicated that leanness (BMI < 18.5 kg/m^2) was associated with a higher risk of LC and every 5 kg/m^2 increase in BMI was associated with a 21% lower risk of LC [35]. Our study focused on patients with COPD and found BMI was inversely associated with LC in COPD. It is well known that COPD and LC are two of the most important smoking-related diseases. Smoking plays a pivotal role in the development of both diseases. Our findings showed that pack-years in patients with COPD+LC were much higher than that in patients with only COPD, indicating the underlying mechanisms of nicotine-induced carcinogenesis. Emphysema is a type of lung damage related to smoking. Previous studies have demonstrated that the degree of emphysematous lesions was associated with LC development in COPD [24, 36]. Similarly, our results found the prevalence of emphysema was much higher in COPD+LC group than that in COPD group.

Those clinical risk factors discovered in this study are similar to the COPD-LUCSS prediction tool for LC in

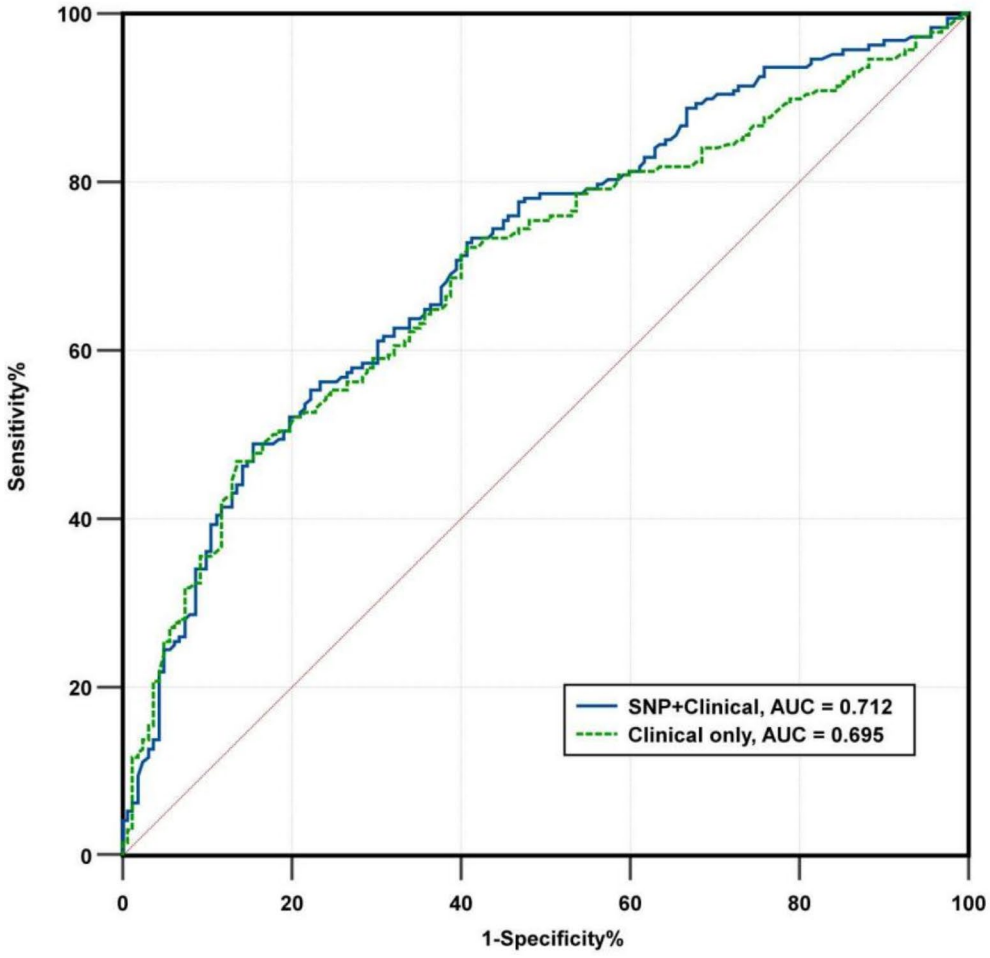

Fig. 2 ROC of the model to predict LC in COPD

COPD, which included age greater than 60 years, BMI less than 25 kg/m^2, pack-years greater than 60, and the presence of emphysema [37]. Given the prominent role of genetic factors, we carried out the candidate gene study, and then identified the rs56113850 in *CYP2A6* as an independent genetic risk factor for LC in COPD, and ultimately developed a prediction model integrating genetic and clinical data. The AUC of this prediction model was modest and less than 30% of patients using this model would be misclassified. Moreover, the predictive ability for LC risk of this model would significantly increase for patients with serious COPD and only less than 20% of patients using this model would be misclassified.

There are, however, several limitations of this study. The subjects in this study were all Chinese Han ethnicity and the results of the candidate gene study may not be applicable to other ethnic groups. The predictive ability of the model ought to be validated against other cohorts of patients from other settings. Furthermore, a prospective cohort study with a larger sample size, and more tag SNPs covered the genes, and more comprehensive clinical information (e.g., duration of COPD, family history of LC, history of acute exacerbations of COPD, and diffusing capacity for carbon monoxide) is needed.

Conclusion

In summary, our study identified the rs56113850 (risk allele C) in *CYP2A6*, decrease in BMI, increase in smoking pack-years and emphysema presence as independent risk factors for LC in COPD. Based on those risk factors, the prediction model integrating genetic and clinical data was constructed for predicting LC risk in COPD. The model demonstrated favorable predictive performance, and could facilitate the early identification and management of the patients with COPD at high risk of LC.

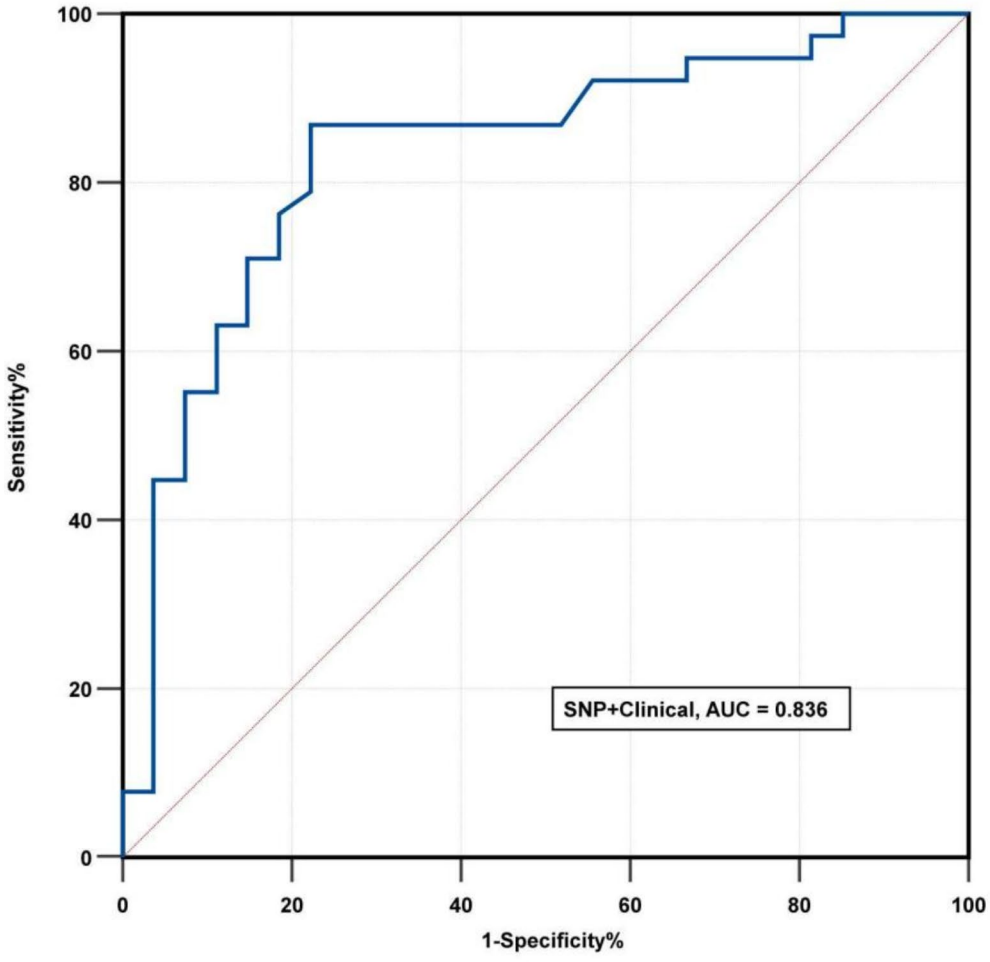

Fig. 3 ROC of the model to predict LC in serious COPD

Abbreviations

AUCs	Areas under receiver operating characteristic curve
BMI	Body mass index
CHRNA3	Cholinergic receptor, neuronal nicotinic, α-polypeptide 3
CHRNA5	Cholinergic receptor, neuronal nicotinic, α-polypeptide 5
CIs	Confidence intervals
COPD	Chronic obstructive pulmonary disease
CT	Computed tomography
CYP1A1	Cytochrome P450 1A1
CYP2A6	Cytochrome P450 2A6
FAM13A	Family with sequence similarity 13 member A
FEV1	Forced expiratory volume in first second
FLT1	Vascular endothelial growth factor receptor 1
FVC	Forced vital capacity
GOLD	Global Initiative for Obstructive Lung Disease
GWAS	Genome-wide association studies
GYPA	Glycophorin A
HHIP	Hedgehog interacting protein
LC	Lung cancer
MAF	Minor allele frequency
MALDI-TOF	Matrix-assisted laser desorption/ionization time of flight
MMP-1	Matrix metalloproteinase-1
NFKBIA	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor alpha
NRXN1	Neurexin 1
ORs	Odds ratios
ROC	Receiver operating characteristic curve
SCLC	Small cell lung cancer
SD	Standard deviation
SNPs	Single nucleotide polymorphisms
STAT3	Signal transducer and activator of transcription 3
TLR5	Toll like receptor 5
TNF	Tumor necrosis factor
3HC	3′-hydroxycotinine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12890-024-03444-5.

Supplementary Material 1: Detailed information regarding the primers of SNPs in this study

Acknowledgements

We thank all the partners and staff who helped us in the process of this study.

Author contributions

ZG and LW designed the study; ZG, YW and FY collected the data and completed the experiments; ZG and JS analyzed the data; ZG wrote the

manuscript; LW revised the manuscript content. All authors have read and approved the final submitted manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (82004115, 82374345) and the Shanghai Key Discipline of Traditional Chinese Medicine (shzyyzdxk-2024106).

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

Declarations

Ethics approval and consent to participate
This study was performed in accordance with the guidelines of the Helsinki Declaration (as revised in 2013) and was approved by the Ethics Committee of Shanghai Pulmonary Hospital affiliated to Tongji University (K20-036Z). Written informed consent was obtained from all subjects. This type of research was considered no need for submission of Clinical Trial Number.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 1 March 2024 / Accepted: 9 December 2024
Published online: 18 December 2024

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Guan WJ, Zheng XY, Chung KF, Zhong NS. Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. Lancet. 2016;388(10054):1939–51.
3. Chen S, Kuhn M, Prettner K, Yu F, Yang T, Bärnighausen T, et al. The global economic burden of chronic obstructive pulmonary disease for 204 countries and territories in 2020-50: a health-augmented macroeconomic modelling study. Lancet Glob Health. 2023;11(8):e1183–93.
4. Park HY, Kang D, Shin SH, Yoo KH, Rhee CK, Suh GY, et al. Chronic obstructive pulmonary disease and lung cancer incidence in never smokers: a cohort study. Thorax. 2020;75(6):506–9.
5. Denholm R, Schüz J, Straif K, Stücker I, Jöckel KH, Brenner DR, et al. Is previous respiratory disease a risk factor for lung cancer? Am J Respir Crit Care Med. 2014;190(5):549–59.
6. Sekine Y, Katsura H, Koh E, Hiroshima K, Fujisawa T. Early detection of COPD is important for lung cancer surveillance. Eur Respir J. 2012;39(5):1230–40.
7. Young RP, Hopkins RJ, Christmas T, Black PN, Metcalf P, Gamble GD. COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. Eur Respir J. 2009;34(2):380–6.
8. Houghton AM. Mechanistic links between COPD and lung cancer. Nat Rev Cancer. 2013;13(4):233–45.
9. Li Y, Sheu CC, Ye Y, de Andrade M, Wang L, Chang SC, et al. Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. Lancet Oncol. 2010;11(4):321–30.
10. Yang IA, Holloway JW, Fong KM. Genetic susceptibility to lung cancer and co-morbidities. J Thorac Dis. 2013;5(Suppl 5):S454–62.
11. Xu J, Shang Y, Cai F, Zhang S, Xiao Z, Wang H, et al. Correlation between lung cancer and the HHIP polymorphisms of chronic obstructive pulmonary disease (COPD) in the Chinese Han population. Genes Immun. 2019;20(4):273–80.
12. Zhou C, Qin Y, Zhao W, Liang Z, Li M, Liu D, et al. International expert consensus on diagnosis and treatment of lung cancer complicated by chronic obstructive pulmonary disease. Transl Lung Cancer Res. 2023;12(8):1661–701.
13. Singh D, Agusti A, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the GOLD science committee report 2019. Eur Respir J. 2019;53(5):1900164.
14. Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The eighth edition lung cancer stage classification. Chest. 2017;151(1):193–203.
15. Young RP, Whittington CF, Hopkins RJ, Hay BA, Epton MJ, Black PN, et al. Chromosome 4q31 locus in COPD is also associated with lung cancer. Eur Respir J. 2010;36(6):1375–82.
16. Bray MJ, Chen LS, Fox L, Hancock DB, Culverhouse RC, Hartz SM, et al. Dissecting the genetic overlap of smoking behaviors, lung cancer, and chronic obstructive pulmonary disease: a focus on nicotinic receptors and nicotine metabolizing enzyme. Genet Epidemiol. 2020;44(7):748–58.
17. Wang H, Yang L, Deng J, Wang B, Yang X, Yang R, et al. Genetic variant in the 3'-untranslated region of VEGFR1 gene influences chronic obstructive pulmonary disease and lung cancer development in Chinese population. Mutagenesis. 2014;29(5):311–7.
18. Baranašić J, Šutić M, Catalano C, Drpa G, Huhn S, Majhen D, et al. TLR5 variants are associated with the risk for COPD and NSCLC development, better overall survival of the NSCLC patients and increased chemosensitivity in the H1299 cell line. Biomedicines. 2022;10(9):2240.
19. Yu Y, Mao L, Lu X, Yuan W, Chen Y, Jiang L, et al. Functional variant in 3'UTR of FAM13A is potentially associated with susceptibility and survival of lung squamous carcinoma. DNA Cell Biol. 2019;38(11):1269–77.
20. He F, Yang R, Li XY, Ye C, He BC, Lin T, et al. Single nucleotide polymorphisms of the NF-κB and STAT3 signaling pathway genes predict lung cancer prognosis in a Chinese Han population. Cancer Genet. 2015;208(6):310–8.
21. Gong WJ, Ma LY, Hu L, Lv YN, Huang H, Xu JQ, et al. STAT3 rs4796793 contributes to lung cancer risk and clinical outcomes of platinum-based chemotherapy. Int J Clin Oncol. 2019;24(5):476–84.
22. Córdoba-Lanús E, Baz-Dávila R, de-Torres JP, Rodríguez-Pérez MC, Maca-Meyer N, Varo N, et al. TNFA-863 polymorphism is associated with a reduced risk of chronic obstructive pulmonary disease: a replication study. BMC Med Genet. 2011;12:132.
23. Shi YY, He L. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. Cell Res. 2005;15(2):97–8.
24. Machida H, Inoue S, Shibata Y, Kimura T, Ota T, Ishibashi Y, et al. The incidence and risk analysis of lung cancer development in patients with chronic obstructive pulmonary disease: possible effectiveness of annual CT-screening. Int J Chron Obstruct Pulmon Dis. 2021;16:739–49.
25. Hancock DB, Markunas CA, Bierut LJ, Johnson EO. Human genetics of addiction: new insights and future directions. Curr Psychiatry Rep. 2018;20(2):8.
26. Tanner JA, Tyndale RF. Variation in CYP2A6 activity and personalized medicine. J Pers Med. 2017;7(4):18.
27. Gao Y, Miksys S, Palmour RM, Tyndale RF. The influence of tobacco smoke/nicotine on CYP2A expression in human and African green monkey lungs. Mol Pharmacol. 2020;98(6):658–68.
28. Buchwald J, Chenoweth MJ, Palviainen T, Zhu G, Benner C, Gordon S, et al. Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent. Mol Psychiatry. 2021;26(6):2212–23.
29. Patel YM, Park SL, Han Y, Wilkens LR, Bickeböller H, Rosenberger A, et al. Novel association of genetic markers affecting CYP2A6 activity and lung cancer risk. Cancer Res. 2016;76(19):5768–76.
30. Du M, Xin J, Zheng R, Yuan Q, Wang Z, Liu H, et al. CYP2A6 activity and cigarette consumption interact in smoking-related lung cancer susceptibility. Cancer Res. 2024;84(4):616–25.
31. Watkins DN, Berman DM, Burkholder SG, Wang B, Beachy PA, Baylin SB. Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer. Nature. 2003;422(6929):313–7.
32. Young RP, Hopkins RJ, Whittington CF, Hay BA, Epton MJ, Gamble GD. Individual and cumulative effects of GWAS susceptibility loci in lung cancer: associations after sub-phenotyping for COPD. PLoS ONE. 2011;6(2):e16476.
33. Chen J, Ke K, Liu Z, Yang L, Wang L, Zhou J, et al. Body mass index and cancer risk: an umbrella review of meta-analyses of observational studies. Nutr Cancer. 2023;75(4):1051–64.
34. Yu D, Zheng W, Johansson M, Lan Q, Park Y, White E, et al. Overall and central obesity and risk of lung cancer: a pooled analysis. J Natl Cancer Inst. 2018;110(8):831–42.
35. Kawai S, Lin Y, Tsuge H, Ito H, Matsuo K, Wada K, et al. Body mass index and lung cancer risk: pooled analysis of 10 prospective cohort studies in Japan. Cancer Sci. 2024;00:1–14.

36. Chubachi S, Takahashi S, Tsutsumi A, Kameyama N, Sasaki M, Naoki K, et al. Radiologic features of precancerous areas of the lungs in chronic obstructive pulmonary disease. Int J Chron Obstruct Pulmon Dis. 2017;12:1613–24.
37. de-Torres JP, Wilson DO, Sanchez-Salcedo P, Weissfeld JL, Berto J, Campo A, et al. Lung cancer in patients with chronic obstructive pulmonary disease. Development and validation of the COPD lung cancer screening score. Am J Respir Crit Care Med. 2015;191(3):285–91.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.