Associations of genetic risk and smoking with incident COPD

Pei-Dong Zhang [1,2,5], Xi-Ru Zhang [1,5], Ao Zhang [3], Zhi-Hao Li [1], Dan Liu [1], Yu-Jie Zhang [1] and Chen Mao [1,4]

[1]Dept of Epidemiology, School of Public Health, Southern Medical University, Guangzhou, China. [2]The Laboratory for Precision Neurosurgery, Nanfang Hospital, Southern Medical University, Guangzhou, China. [3]State Key Laboratory of Molecular Neuroscience and Center of Systems Biology and Human Health, Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China. [4]Dept of Laboratory Medicine, Microbiome Medicine Center, Zhujiang Hospital, Southern Medical University, Guangzhou, China. [5]Pei-Dong Zhang and Xi-Ru Zhang contributed equally to the work.

Corresponding author: Chen Mao (maochen9@smu.edu.cn)

Shareable abstract (@ERSpublications)
The polygenic risk score, constructed from 2.5 million variants, showed a significant association with incident COPD. Individuals with a high genetic risk may be more vulnerable to the lung-damaging effects of smoking and develop COPD. https://bit.ly/2T3lgub

Cite this article as: Zhang P-D, Zhang X-R, Zhang A, *et al*. Associations of genetic risk and smoking with incident COPD. *Eur Respir J* 2022; 59: 2101320 [DOI: 10.1183/13993003.01320-2021].

Received: 25 Feb 2021
Accepted: 14 June 2021

Abstract

Background Genetic factors and smoking contribute to chronic obstructive pulmonary disease (COPD), but whether a combined polygenic risk score (PRS) is associated with incident COPD and whether it has a synergistic effect on smoking remains unclear. We aimed to investigate the association of the PRS with COPD and explore whether smoking behaviours could modify such association.

Methods Multivariable Cox proportional hazards models were used to estimate hazard ratios (HRs) and 95% confidence intervals for the association of the PRS and smoking with COPD.

Results The study included 439 255 participants (mean age 56.5 years; 53.9% female), with a median follow-up of 9.0 years. PRS_{lasso} containing 2.5 million variants showed better discrimination and a stronger association for incident COPD than PRS_{279} containing 279 genome-wide significance variants. Compared with low genetic risk, the HRs of medium and high genetic risk were 1.39 (95% CI 1.31–1.48) and 2.40 (95% CI 2.24–2.56), respectively. The HR of high genetic risk and current smoking was 11.62 (95% CI 10.31–13.10) times that of low genetic risk and never smoking. There were significant interactions between PRS_{lasso} and smoking status for incident COPD ($p_{interaction}<0.001$). From low genetic risk to high genetic risk, the HRs of current smoking increased from 4.32 (95% CI 3.69–5.06) to 6.89 (95% CI 6.21–7.64) and the population-attributable risks of smoking increased from 42.7% to 61.1%.

Conclusions The PRS constructed from millions of variants below genome-wide significance showed significant associations with incident COPD. Participants with a high genetic risk may be more susceptible to developing COPD when exposed to smoking.

Introduction

Chronic obstructive pulmonary disease (COPD) is the most prevalent chronic respiratory disease, and is characterised as a progressive and not fully reversible airflow limitation. In 2017, 3.20 million deaths were attributed to COPD worldwide, 23% more than the number of deaths in 1990 [1]. Cigarette smoking is the major risk factor for COPD, mainly causing chronic inflammatory responses and oxidative stress [2]. Plenty of evidence has shown that avoiding smoking can reduce the risk of COPD [3]. However, a striking proportion of 25–45% of COPD cases occur in never-smokers and only 25% of continuous smokers will develop incident COPD [4, 5]. These findings highlight the potential importance of other risk factors, one of which is the genetic structure.

Over the past decade, genome-wide association studies (GWASs) have successfully identified multiple genetic variants associated with the risk of COPD and COPD-related phenotypes [6–8], including *FAM13A*, *HHIP*, *RIN3*, *CHRNA3/5* and *IREB2* [9–13]. However, each common variant's impact consistently and significantly associated with the risk of COPD is modest, indicating that individual genetic variation only explains a small fraction of COPD susceptibility. Aggregating multiple single

nucleotide polymorphisms (SNPs) with small effects to generate a composite polygenic risk score (PRS) may elucidate the genetic risk of complex diseases. Recently, based on case–control studies, a PRS that contains millions of SNPs that have not reached genome-wide significance has been verified to predict a higher risk for developing various diseases, including COPD [8, 14]. However, it is still unclear whether the PRS is associated with new-onset COPD events, and whether smoking and genetic factors have synergistic effects on COPD is still controversial.

The purpose of this study was to evaluate whether the PRS constructed from 2.5 million variants below genome-wide significance is associated with the risk of incident COPD in a large population-based cohort and whether there are differences in the effects of smoking among different genetic risks.

Methods

Study design

The UK Biobank study recruited more than 500 000 participants aged 40–69 years from the general population at 22 assessment centres throughout the UK between 2006 and 2010 [15]. Participants provided information on health-related aspects through extensive baseline questionnaires, verbal interviews and physical measurements. Self-reported ethnicities were categorised as Mixed, Black/African, East Asian, White/European, South Asian and Unknown. Participants were excluded if they withdrew from the study, their genotype data did not meet the quality control conditions, they had a relatedness of second degree or higher, or they had a history of COPD (figure 1). The UK Biobank received ethical approval from the Research Ethics Committee (REC 11/NW/0382) and participants provided written informed consent. Any additional ethical approval was adjudged unnecessary for the present study.

Polygenic risk score

Based on the same GWAS, two sets of PRSs comprised of genetic variants related to lung function that showed the ability to predict COPD prevalence were included in this study [7, 8]. PRS_{279} was created following an additive model for 279 genome-wide significance variants (supplementary table E1) [7]. The number of risk alleles was summed after multiplication with the effect size between the SNPs and forced expiratory volume in 1 s (FEV_1)/forced vital capacity (FVC). PRS_{lasso} was created using a weighted sum of the two PRSs for FEV_1 and FEV_1/FVC [8], which were penalised using lasso regression to simplify the final model. A total of 2.5 million SNPs were calculated using lassosum version 0.4.5 [16]. The detailed

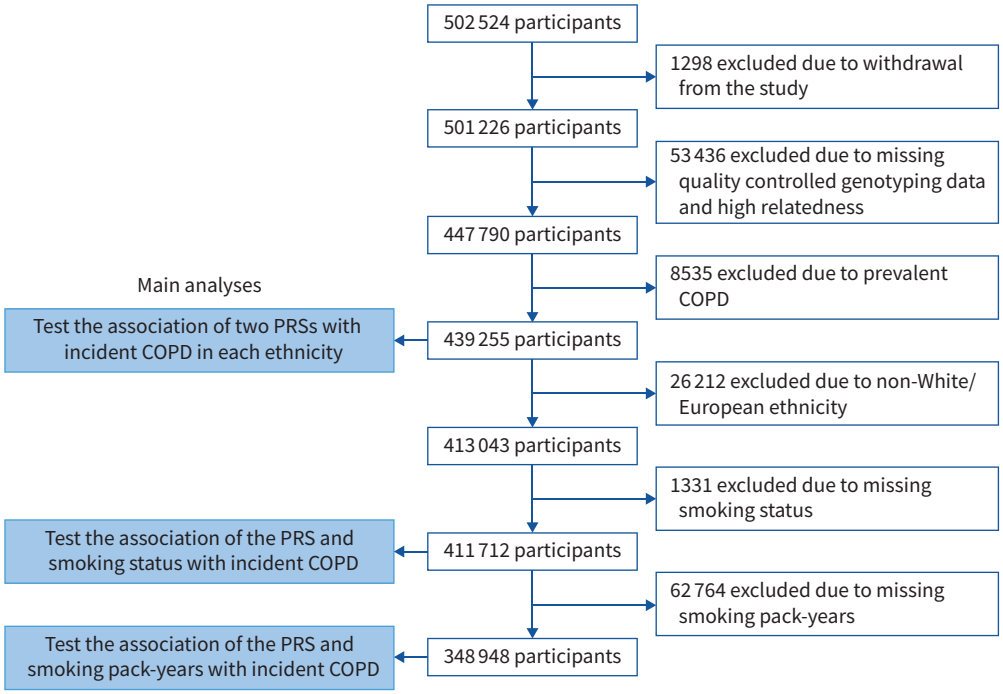


FIGURE 1 Flowchart of participant enrolment. COPD: chronic obstructive pulmonary disease; PRS: polygenic risk score.

derivation of PRS_{lasso} is shown in the supplementary material. Then, the two PRSs were categorised into low (lowest quintile), intermediate (quintiles 2–4) and high (highest quintile) risk.

Smoking status and pack-years

Touchscreen questionnaires collected information on smoking status and pack-years at baseline. Detailed definitions of smoking status and pack-years of smoking are provided in supplementary table E2. All participants were categorised as never, former and current smoking according to their smoking status, or as no (0), light (0.1–19.9), intermediate (20–39.9) and heavy (⩾40) smoking according to pack-years of smoking.

COPD and lung function

Trained healthcare technicians and nurses at UK Biobank assessment centres used a Pneumotrac 6800 spirometer (Vitalograph, Maids Moreton, UK) to perform lung function tests. Each participant performed two to three tests, and was analysed using the maximum acceptable value of FVC and FEV_1 [17]. Participants with incident COPD were identified as having a diagnosis in hospital admission electronic health records or death register data, or lung function test FEV_1/FVC ratio below the Global Lung Function Initiative (GLI) 2012 reference values [18] for the lower limit of normal post the date of baseline assessment [19]. We calculated the follow-up time from the date of attendance until the date of first diagnosis, date of death or 25 February 2018 for Wales and England, or 28 February 2017 for Scotland, whichever occurred first (supplementary table E3).

Statistical analyses

The associations between two sets of PRSs and incident COPD across strata of ethnicity were assessed by the area under the curve from receiver operating characteristic (ROC) models and the Harrell C-statistic from Cox proportional hazards models. We assessed whether Cox proportional hazards models and correlated ROC curves were significantly different using ANOVA and the Delong test.

The characteristics of the participants were summarised across incident COPD status as number (percentage) for categorical variables, mean with standard deviation for normally distributed variables and median (interquartile range (IQR)) for skewed variables. The association between genetic risk categories, smoking categories, and the combination of genetic and smoking categories (nine categories with low genetic risk and never smoking as a reference; 12 categories with low genetic risk and no smoking pack-years as a reference) and incident COPD were explored using multivariable Cox proportional hazards models. The covariates included in this study were recognised risk factors for COPD [20], which were unevenly distributed among the exposure groups. All models were adjusted for age, sex, education, socioeconomic status (household income and Townsend deprivation index [21]), body mass index, physical activity, healthy diet, alcohol consumption, passive smoking, occupational exposure, third-degree relatedness of individuals in the sample, genotyping chip and the first 20 principal components of ancestry (supplementary table E2). Detailed information on the number of missing covariates is shown in supplementary table E4. We used multiple imputations by chained equations to impute missing covariate values with the mice package version 3.12.0 [22]. The assumption for proportional hazards was evaluated by tests based on Schoenfeld residuals [23]; violation of this assumption was not observed in our analyses. Moreover, interactions between PRS and smoking status or pack-years were tested by adding the cross-product term to Cox models.

The associations between genetic risk and smoking pack-years and incident COPD were evaluated on a continuous scale with restricted cubic spline curves based on multivariable Cox proportional hazards models. To balance best fit and over-fitting in the main splines for incidence, the number of knots, between three and five, was chosen as the lowest value for the Akaike Information Criterion, but if within two of each other for different knots, the lowest number of knots was chosen. The sensitivity and subgroup analysis methods are shown in the supplementary material.

The population-attributable fraction (PAF), an estimate of the proportion of events that would have been prevented if all individuals would have been in a lower smoking category, was calculated [24]. Analyses were undertaken using R version 3.6.1 (R Center for Statistical Computing, Vienna, Austria). A p-value <0.05 (two-sided) was considered significant. Because we tested the joint association of genetic risk and smoking to maximise the likelihood of reporting true findings, we conservatively corrected for multiple testing using Bonferroni correction and set a significance level of 0.05/11=0.0045.

TABLE 1 Characteristics of White/European participants

Characteristics	Incident COPD (n=9577)	No incident COPD (n=402 135)
Mean±SD age, years	60.3±6.8	56.6±8.0
Sex		
Female	4443 (46.4)	217 769 (54.2)
Male	5134 (53.6)	184 366 (45.8)
Smoking status		
Never	2356 (24.6)	222 136 (55.2)
Former	4115 (43.0)	141 547 (35.2)
Current	3106 (32.4)	38 452 (9.6)
Smoking pack-years		
No (0)	2420 (28.9)	223 007 (65.5)
Light (0.1–19.9)	1359 (16.2)	63 624 (18.7)
Intermediate (20–39.9)	2266 (27.1)	38 363 (11.3)
Heavy (⩾40)	2332 (27.8)	15 577 (4.6)
Body mass index, kg·m^{-2}		
Mean±SD	28.1±5.6	27.4±4.7
<18.5	101 (1.1)	1907 (0.5)
18.5–24.9	2840 (29.7)	132 410 (32.9)
25–29.9	3680 (38.4)	172 087 (42.8)
⩾30	2956 (30.9)	95 731 (23.8)
Regular physical activity		
Yes	5083 (53.1)	235 141 (58.5)
No	4494 (46.9)	166 994 (41.5)
Diet (DASH score)		
Mean±SD	21.0±4.4	22.2±4.0
Ideal (26–35)	1675 (17.5)	92 142 (22.9)
Intermediate (19–25)	5293 (55.3)	245 715 (61.1)
Poor (7–18)	2609 (27.2)	64 278 (16.0)
Drinking status		
Never	359 (3.7)	12 557 (3.1)
Former	643 (6.7)	13 223 (3.3)
Current	8575 (89.5)	376 355 (93.6)
Alcohol consumption, g per day		
None (0)	3868 (40.4)	201 197 (50.0)
Light-to-moderate (male: 0–28; female: 0–14)	2865 (29.9)	90 783 (22.6)
Excessive (male: >28; female: >14)	2844 (29.7)	110 155 (27.4)
Passive smoking		
No	7000 (73.1)	319 610 (79.5)
Yes	2577 (26.9)	82 525 (20.5)
Occupational exposure		
Rarely/never	7992 (83.4)	315 068 (78.3)
Sometimes	885 (9.2)	54 432 (13.5)
Often	700 (7.3)	32 635 (8.1)
Median (IQR) TDI	−1.1 (−3.1–2.2)	−2.3 (−3.7–0.1)
Household income, GBP		
<18 000	3976 (41.5)	87 267 (21.7)
18 000–30 999	2690 (28.1)	103 125 (25.6)
31 000–51 999	1756 (18.3)	106 669 (26.5)
52 000–100 000	964 (10.1)	83 036 (20.6)
>100 000	191 (2.0)	22 038 (5.5)
Education		
Lower qualification	6255 (65.3)	205 215 (51.0)
Higher qualification	3322 (34.7)	196 920 (49.0)
Genetic risk category (PRS_{lasso})		
Low (lowest quintile)	1217 (12.7)	81 126 (20.2)
Intermediate (quintiles 2–4)	5270 (55.0)	241 756 (60.1)
High (highest quintile)	3090 (32.3)	79 253 (19.7)

Data are presented as n (%), unless otherwise stated. COPD: chronic obstructive pulmonary disease; DASH: Dietary Approaches to Stop Hypertension; IQR: interquartile range; TDI: Townsend deprivation index; PRS: polygenic risk score. All categorical variables globally significantly different between groups at p<0.001.

Results

Participant's characteristics

The process of enrolling participants in this study is shown in figure 1. The overall study population included 439 255 participants (mean±SD age 56.5±8.0 years), of which 236 795 (53.9%) individuals were female (supplementary table E5). PRS_{lasso} showed better discrimination for incident COPD (PRS_{lasso} C-statistic 0.60; PRS_{279} C-statistic 0.53; p<0.001) (supplementary table E6), while PRS_{lasso} also showed a greater risk of incident COPD for each quintile score increase among the White/European participants (PRS_{lasso} hazard ratio (HR) 1.24, 95% CI 1.23–1.26; PRS_{279} HR 1.07, 95% CI 1.05–1.08) (supplementary figure E1). However, there was no significant association between the two sets of PRSs and incident COPD in other ethnic groups, except for Black/African individuals, with HR 1.26 (95% CI 1.00–1.58) per quintile PRS_{lasso} increase. We subsequently analysed only individuals of White/European ethnicity for improving statistical power and predictive strength, and used PRS_{lasso} as the main genetic risk assessment method (figure 1).

Table 1 presents the baseline characteristics of eventually included participants. Of the 411 712 White/European individuals (mean±SD age 56.8±8.0 years), 222 212 (54.0%) were female. There were 145 662 (35.4%) former smokers and 41 558 (10.1%) current smokers, among which 40 629 (11.6%) individuals had intermediate smoking exposure (20–39.9 pack-years) and 17 909 (5.1%) individuals had heavy smoking exposure (⩾40 pack-years). Over 3 625 259 person-years of follow-up (median (IQR) length of follow-up 9.0 (8.3–9.5) years), there were 9577 cases of incident COPD. The characteristics of COPD cases determined by the two sources are shown in supplementary table E7. Participants who developed incident COPD were slightly older, more likely to be male and obese, had more smoking exposure, and had higher genetic risk scores.

Associations of genetic risk with incident COPD

For the increased genetic risk groups, the incidence and HR of COPD gradually increased. The high genetic risk group HR was 2.40 (95% CI 2.24–2.56) compared with the low genetic risk group. After additional adjustment for smoking status or pack-years, the HRs of the high genetic risk group were 2.37 (95% CI 2.21–2.53) and 2.27 (95% CI 2.12–2.44), respectively (table 2). The association between PRS_{lasso} on a continuous scale and risk of incident COPD was nonlinear ($p_{nonlinear}$<0.001); a high PRS_{lasso} presented a very high risk (figure 2). When genetic risk deciles were used instead of categories, the same trend of results was observed (supplementary table E9). Supplementary figure E3a shows the cumulative risk of incident COPD in each genetic risk group during follow-up.

Associations of smoking with incident COPD

As the smoking status changed and smoking pack-years increased, the incidence and HR of COPD also increased. The HRs of the current and heavy smoking groups were 5.97 (95% CI 5.64–6.32) and 8.32 (95% CI 7.81–8.86), respectively, compared with the never smoking group. After additional adjustment for PRS_{lasso}, the HRs were slightly lower than before among the smoking groups (table 3). The association between smoking pack-years on a continuous scale and risk of incident COPD was nonlinear ($p_{nonlinear}$<0.001); heavy smoking presents a very high risk (supplementary figure E2). When the number of smoking pack-years was further subdivided into more categories, the same results were observed (supplementary table E10). Compared with former smokers, the associations between smoking pack-years and COPD risk were stronger among current smokers (supplementary table E11). The cumulative risk of incident COPD in each smoking status and pack-year group during follow-up is shown in supplementary figure E3b and c.

Associations of smoking and genetic risk with incident COPD

Combining genetic risk and smoking to group the entire cohort, the risk of incident COPD still increased with smoking and genetic risk (figure 3). Compared with the low genetic risk and never smoking group, the high genetic risk and current smoking group HR was 11.62 (95% CI 10.31–13.10). A similar pattern was observed among the genetic risk and smoking pack-year groups, and the highest risk was observed among individuals with high genetic risk and heavy smoking exposure (HR 14.85, 95% CI 13.09–16.84).

Moreover, we observed significant interactions between the PRS_{lasso} categories and smoking status or pack-years (both $p_{interaction}$<0.001) (table 4), and the interactions between PRS_{279} and smoking were not significant (smoking status: $p_{interaction}$=0.116; smoking pack-years: $p_{interaction}$=0.334) (supplementary figure E4). When stratified by current smokers and former smokers, the interactions between genetic risk and smoking pack-years for incident COPD remained significant in both groups (both $p_{interaction}$<0.001) (supplementary table E12). The impact of smoking was substantially more pronounced in those with an elevated genetic risk. In the low, intermediate and high genetic risk groups, the HRs of current smoking

https://doi.org/10.1183/13993003.01320-2021

TABLE 2 Risk of incident chronic obstructive pulmonary disease (COPD) according to genetic risk

Genetic risk	Total participants, n	COPD cases, n (%)	Person-years	IR[#]	Model 1[¶]			Model 2[+]			Model 3[§]		
					HR (95% CI)	p-value	p_{trend}-value	HR (95% CI)	p-value	p_{trend}-value	HR (95% CI)	p-value	p_{trend}-value
Low	82 343	1217 (1.48)	727 481	1.67	1.00 (reference)		<0.001	1.00 (reference)		<0.001	1.00 (reference)		<0.001
Intermediate	247 026	5270 (2.13)	2 176 250	2.42	1.39 (1.31–1.48)	<0.001		1.38 (1.30–1.47)	<0.001		1.35 (1.27–1.45)	<0.001	
High	82 343	3090 (3.75)	721 528	4.28	2.40 (2.24–2.56)	<0.001		2.37 (2.21–2.53)	<0.001		2.27 (2.12–2.44)	<0.001	

IR: incidence rate; HR: hazard ratio. [#]: per 1000 person-years. [¶]: Model 1: Cox proportional hazards regression adjusted for age, sex, education, Townsend deprivation index, income, body mass index, diet, physical activity, alcohol consumption, occupational exposure, passive smoking, relatedness, genotyping chip and the first 20 principal components of ancestry (p_{trend}-value calculated treating the polygenic risk score as a continuous variable); [+]: Model 2: Cox proportional hazards regression adjusted for Model 1 and smoking status categories (p_{trend}-value calculated treating the genetic risk score as a continuous variable); [§]: Model 3: Cox proportional hazards regression adjusted for Model 1 and smoking pack-years categories (p_{trend}-value calculated treating the genetic risk score as a continuous variable).

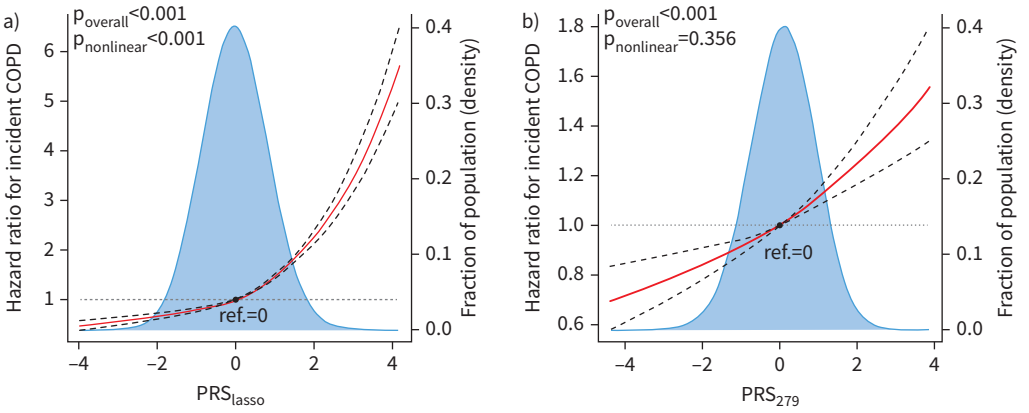


FIGURE 2 Multivariable adjusted hazard ratios for incident chronic obstructive pulmonary disease (COPD) according to a) PRS_{lasso} and b) PRS_{279} on a continuous scale. Solid red lines are multivariable adjusted hazard ratios with dashed black lines showing 95% confidence intervals derived from restricted cubic spline regressions with three knots. Reference (ref.) lines for no association are indicated by the grey dotted horizontal lines at HR=1. Blue shaded curves show the fraction of the population with different exposures. PRS: polygenic risk score.

were 4.32 (95% CI 3.69–5.06), 5.92 (95% CI 5.48–6.39) and 6.89 (95% CI 6.21–7.64), respectively, compared with never smoking (table 4). Supplementary figure E5 shows the cumulative risk of incident COPD in each smoking status and pack-year group among each genetic risk category during follow-up.

The same pattern of associations was observed in a series of sensitivity analyses with additional adjustment for asthma, chronic pulmonary infections and residential air pollution, excluding participants related to others and participants who developed outcomes within the first 2 years of follow-up (supplementary tables E13 and E14). Subgroup analyses were performed by age and sex (supplementary tables E15 and E16). The risk of incident COPD increased with elevated smoking and the genetic risk was more evident in elderly individuals aged >60 years ($p_{interaction}$<0.001).

Population-attributable fractions

If all individuals do not smoke, 54.9% (95% CI 53.1–56.6%, based on smoking status) to 55.3% (95% CI 53.4–57.1%, based on smoking pack-years) new-onset COPD events might be prevented during follow-up. Another reality was that if all current smokers quit smoking, new-onset events might be reduced by 20.2%

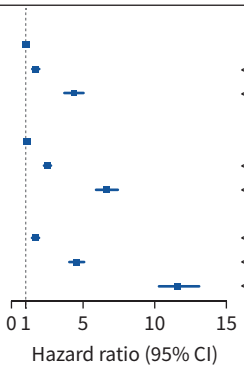
TABLE 3 Risk of incident chronic obstructive pulmonary disease (COPD) according to smoking categories

Smoking	Total participants, n	COPD cases, n (%)	Person-years	IR[#]	Model 1[¶]			Model 2[+]		
					HR (95% CI)	p-value	p_{trend}-value	HR (95% CI)	p-value	p_{trend}-value
Smoking status										
Never	224 492	2356 (1.05)	1 996 181	1.18	1.00 (reference)		<0.001	1.00 (reference)		<0.001
Former	145 662	4115 (2.83)	1 274 040	3.23	2.26 (2.15–2.38)	<0.001		2.27 (2.16–2.39)	<0.001	
Current	41 558	3106 (7.47)	355 038	8.75	5.97 (5.64–6.32)	<0.001		5.93 (5.60–6.28)	<0.001	
Smoking pack-years										
No (0)	225 427	2420 (1.07)	2 004 025	1.21	1.00 (reference)		<0.001	1.00 (reference)		<0.001
Light (0.1–19.9)	64 983	1359 (2.09)	572 502	2.37	1.98 (1.85–2.11)	<0.001		1.99 (1.86–2.12)	<0.001	
Intermediate (20–39.9)	40 629	2266 (5.58)	349 951	6.48	4.32 (4.07–4.58)	<0.001		4.29 (4.04–4.56)	<0.001	
Heavy (⩾40)	17 909	2332 (13.02)	145 833	15.99	8.32 (7.81–8.86)	<0.001		8.19 (7.69–8.73)	<0.001	

IR: incidence rate; HR: hazard ratio. [#]: per 1000 person-years; [¶]: Model 1: Cox proportional hazards regression adjusted for age, sex, education, Townsend deprivation index, income, body mass index, diet, physical activity, alcohol consumption, occupational exposure, passive smoking, relatedness, genotyping chip and the first 20 principal components of ancestry (p_{trend}-value calculated treating each smoking category as a continuous variable); [+]: Model 2: Cox proportional hazards regression adjusted for Model 1 and polygenic risk score (p_{trend}-value calculated treating each smoking category as a continuous variable).

a) Genetic risk (PRS_{lasso}) and smoking status

Subgroup	Total participants, n	COPD cases, n	Hazard ratio (95% CI)	p-value
Low genetic risk				
Never smoking	44506	377	1.00 (reference)	
Former smoking	30265	509	1.69 (1.48–1.93)	<0.001
Current smoking	7572	331	4.36 (3.75–5.05)	<0.001
Intermediate genetic risk				
Never smoking	134862	1321	1.13 (1.01–1.26)	0.041
Former smoking	87219	2257	2.51 (2.25–2.80)	<0.001
Current smoking	24945	1692	6.60 (5.89–7.40)	<0.001
High genetic risk				
Never smoking	45124	658	1.67 (1.47–1.90)	<0.001
Former smoking	28178	1349	4.53 (4.04–5.08)	<0.001
Current smoking	9041	1083	11.62 (10.31–13.10)	<0.001



b) Genetic risk (PRS_{lasso}) and smoking pack-years

Subgroup	Total participants, n	COPD cases, n	Hazard ratio (95% CI)	p-value
Low genetic risk				
No (0) smoking	44676	386	1.00 (reference)	
Light (0.1–19.9) smoking	13577	156	1.34 (1.11–1.62)	0.002
Intermediate (20–39.9) smoking	7802	261	3.19 (2.72–3.73)	<0.001
Heavy (≥40) smoking	3040	260	6.61 (5.63–7.75)	<0.001
Intermediate genetic risk				
No (0) smoking	135420	1352	1.12 (1.00–1.26)	0.044
Light (0.1–19.9) smoking	38907	730	2.15 (1.90–2.43)	<0.001
Intermediate (20–39.9) smoking	24222	1214	4.71 (4.20–5.29)	<0.001
Heavy (≥40) smoking	10743	1304	9.35 (8.32–10.51)	<0.001
High genetic risk				
No (0) smoking	45331	682	1.68 (1.48–1.90)	<0.001
Light (0.1–19.9) smoking	12499	473	4.22 (3.69–4.83)	<0.001
Intermediate (20–39.9) smoking	8605	791	8.52 (7.54–9.64)	<0.001
Heavy (≥40) smoking	4126	768	14.85 (13.09–16.84)	<0.001

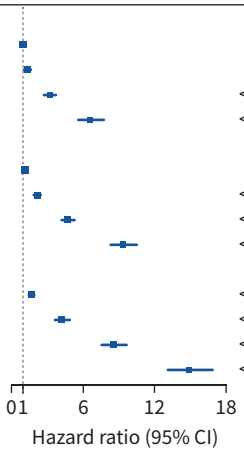


FIGURE 3 Risk of incident chronic obstructive pulmonary disease (COPD) according to a) genetic risk (PRS_{lasso}) and smoking status or b) genetic risk (PRS_{lasso}) and smoking pack-years. Grey dashed vertical lines indicate the reference value of HR=1. PRS: polygenic risk score.

(95% CI 19.9–20.5%). In addition, if smoking pack-years were reduced by one or two levels, 40.7% (95% CI 39.6–41.8%) or 53.1% (95% CI 51.4–54.7%) of new incident cases might be prevented. Further analyses stratified by genetic risk category showed that 42.7% (95% CI 37.1–48.0%), 54.1% (95% CI 51.7–56.4%) and 61.1% (95% CI 58.2–63.9%) of incident COPD cases were attributed to smoking among the low, intermediate and high genetic risk populations (table 5).

Discussion

In this large population-based prospective cohort study, the PRS constructed from more than 2.5 million variations, which did not reach genome-wide significance, showed better predictive accuracy for incident COPD. After stratifying more than 411000 White/European participants with the PRS, we found that smoking was more strongly associated with incident COPD in the high genetic risk group. The population-attributable risks of smoking increased from 42.7% to 61.1% from low genetic risk to high genetic risk.

Moll *et al.* [8] developed PRS_{lasso} based on the GWAS for FEV_1 and FEV_1/FVC ratio, and showed that it could significantly improve the predictive power of COPD in a case–control study. This PRS construction method containing millions of variants below genome-wide significance was also applied for other diseases [14]. However, the current study was based on a prospective study design and our results suggest that the PRS was a significant predictor of new-onset COPD cases. The significant nonlinear association between PRS_{lasso} and COPD risks allowed the discovery of individuals with extremely high genetic risk.

TABLE 4 Risk of incident chronic obstructive pulmonary disease (COPD) according to smoking category within each genetic risk category

Subgroup	Total participants, n	COPD cases, n (%)	Person-years	IR[#]	HR (95% CI)[¶]	p-value	p_{trend}-value	$p_{interaction}$-value
Genetic risk and smoking status								
Low genetic risk								<0.001
Never smoking	44 506	377 (0.85)	395 923	0.95	1.00 (reference)		<0.001	
Former smoking	30 265	509 (1.68)	265 814	1.91	1.72 (1.50–1.97)	<0.001		
Current smoking	7572	331 (4.37)	65 744	5.03	4.32 (3.69–5.06)	<0.001		
Intermediate genetic risk								
Never smoking	134 862	1321 (0.98)	1 199 115	1.10	1.00 (reference)		<0.001	
Former smoking	87 219	2257 (2.59)	763 457	2.96	2.22 (2.08–2.38)	<0.001		
Current smoking	24 945	1692 (6.78)	213 678	7.92	5.92 (5.48–6.39)	<0.001		
High genetic risk								
Never smoking	45 124	658 (1.46)	401 144	1.64	1.00 (reference)		<0.001	
Former smoking	28 178	1349 (4.79)	244 769	5.51	2.70 (2.45–2.97)	<0.001		
Current smoking	9041	1083 (11.98)	75 616	14.32	6.89 (6.21–7.64)	<0.001		
Genetic risk and smoking pack-years								
Low genetic risk								<0.001
No (0) smoking	44 676	386 (0.86)	397 378	0.97	1.00 (reference)		<0.001	
Light (0.1–19.9) smoking	13 577	156 (1.15)	119 956	1.30	1.36 (1.13–1.64)	0.001		
Intermediate (20–39.9) smoking	7802	261 (3.35)	67 823	3.85	3.28 (2.78–3.87)	<0.001		
Heavy (⩾40) smoking	3040	260 (8.55)	25 260	10.29	6.79 (5.69–8.10)	<0.001		
Intermediate genetic risk								
No (0) smoking	135 420	1352 (1.00)	1 203 814	1.12	1.00 (reference)		<0.001	
Light (0.1–19.9) smoking	38 907	730 (1.88)	342 898	2.13	1.93 (1.76–2.11)	<0.001		
Intermediate (20–39.9) smoking	24 222	1214 (5.01)	208 991	5.81	4.25 (3.92–4.61)	<0.001		
Heavy (⩾40) smoking	10 743	1304 (12.14)	87 883	14.84	8.46 (7.77–9.21)	<0.001		
High genetic risk								
No (0) smoking	45 331	682 (1.50)	402 833	1.69	1.00 (reference)		<0.001	
Light (0.1–19.9) smoking	12 499	473 (3.78)	109 647	4.31	2.47 (2.20–2.78)	<0.001		
Intermediate (20–39.9) smoking	8605	791 (9.19)	73 136	10.82	5.00 (4.49–5.56)	<0.001		
Heavy (⩾40) smoking	4126	768 (18.61)	32 690	23.49	8.56 (7.63–9.60)	<0.001		

IR: incidence rate; HR: hazard ratio. [#]: per 1000 person-years; [¶]: Cox proportional hazards regression adjusted for age, sex, education, Townsend deprivation index, income, body mass index, diet, physical activity, alcohol consumption, occupational exposure, passive smoking, relatedness, genotyping chip and the first 20 principal components of ancestry (p_{trend}-value calculated treating each smoking category as a continuous variable).

Meanwhile, our results also demonstrated that the PRS constructed based on cross-sectional data showed associations not only with peak lung function but also with accelerated lung function decline [25, 26], which is the main feature of COPD. Therefore, this study provided evidence from a large-sample prospective cohort study for the application of the PRS. Moreover, due to the inclusion of below genome-wide significance variants, the study also showed that the new method based on regularised regression might significantly improve the predictive performance of the PRS.

TABLE 5 Population-attributable fraction (PAF) per smoking group

Hypothesis	Whole population	Low genetic risk	Intermediate genetic risk	High genetic risk
Smoking status				
Former and current smoking to never smoking	54.9 (53.1–56.6)	42.7 (37.1–48.0)	54.1 (51.7–56.4)	61.1 (58.2–63.9)
Current smoking to former smoking	20.2 (19.9–20.5)	16.7 (15.7–17.7)	19.9 (19.4–20.3)	21.0 (20.6–21.5)
Smoking pack-years				
Ever smoking to no smoking (0 pack-years)	55.3 (53.4–57.1)	43.8 (38.0–49.5)	54.6 (52.0–57.0)	60.9 (57.8–63.8)
Reduce smoking pack-years by two levels	53.1 (51.4–54.7)	43.0 (37.7–48.1)	52.5 (50.3–54.7)	57.4 (54.7–60.0)
Reduce smoking pack-years by one level	40.7 (39.6–41.8)	34.7 (30.9–38.2)	40.6 (39.1–42.1)	42.0 (40.2–43.6)

Data are presented as PAF % (95% CI).

The predictive power for incident COPD of the PRS has not been well verified in other ethnic groups except for White/European and Black/African participants. The main reason for this may be that the sample size of other ethnic groups included in the UK Biobank study is limited and there were not enough outcome events recorded during the follow-up, which may lead to inadequate statistical power. In addition, most GWASs are currently conducted on the European ancestry population, and directly generalising the weights and risk loci to other races/ethnicities may attenuate its predictive power accuracy [27]. Thus, using appropriate methods to develop additional PRSs in multi-ethnic populations is critical to implement precision medicine and prevention in global health [28].

Smoking can directly induce tissue damage through oxidative stress or indirectly induce an inflammatory response, resulting in irreversible airway remodelling. Whether smoking and genetic susceptibility interact with the occurrence and development of COPD and decreased lung function has long been a concern. This study is the first in a large population-based prospective study to confirm that smoking has a more significant impact on new-onset COPD in a high genetic risk population. Both the attributable risks and the differences in smoking hazard ratios increased from low genetic risk to high genetic risk. Similar results have been reported: individuals with a high genetic risk for low FEV_1/FVC (a PRS constructed with 26 SNPs) are more susceptible to the deleterious effects of smoking [29]. However, studies of the UK Biobank suggested that smoking and genetic effects generally act independently. Wain *et al.* [30] included 50 008 individuals with extreme FEV_1 and smoking behaviours for a GWAS. The results showed shared genetic causes of low FEV_1 between heavy smokers and never-smokers, and smoking only likely interacted with a small proportion of the genetic effects. Shrine *et al.* [7] combined 279 FEV_1/FVC-related variants to construct a PRS. Based on the cross-sectional study results, no significant smoking–PRS interaction for FEV_1/FVC was observed and only a weak interaction for COPD was observed. A GWAS of 5070 participants found five FEV_1/FVC-related variants, but no interaction between them and smoking was found [31]. In this study, the significant interaction may be attributed to the novel PRS estimation method, which combines the two lung function phenotypes of FEV_1 and FEV_1/FVC ratio and includes 2.5 million variants. Meanwhile, the interaction may also benefit from identifying extremely high genetic risk populations brought about by the significant nonlinear association between PRS_{lasso} and incident COPD risk. It provided a more accurate description of the genetic characteristics of lung function. The results also suggested that similar PRS estimation methods may be an effective tool for discovering interactions between genetics and environmental factors.

Previous studies showed that the PRS was associated with several computed tomography imaging phenotypes, including quantitative emphysema and airway wall thickness measures, and that it was associated with reduced lung growth patterns in children with asthma [8]. These characteristics may be why an individual with high genetic risk has more severe damage or poor repair and is more likely to develop COPD after smoking exposure. Moreover, multiple genes, including *CHRNA3* and *CHRNA5* at 15q25, were strongly associated with lung function, COPD and smoking behaviours [30, 32–35]. A higher PRS may lead to more undocumented smoking exposures, including deeper inhalation depth and tobacco selection with a higher nicotine content. It is challenging to attempt a rigorous biological mechanism for millions of SNPs, but it is worth noting that a model called Omnigenic may serve as its theoretical basis [36]. This model refers to the existence of a considerable number ("omni-" refers to "all") of genes that may contribute to disease risk, among which peripheral genes (numerous, pleiotropy and regulatory effects) play a synergistic role by influencing core genes (rare, specificity, interpretable biological roles) through a regulatory network [37]. Although in the current study it was difficult to solve the model's concerns about regulatory networks and rare mutations, there was a consensus on the involvement of below genome-wide significance variants. Therefore, we speculate that this may be an essential strategy for genetic risk assessment.

Considering that more than half of new-onset COPD cases are still attributed to smoking, all populations, especially those who have a high genetic risk, are recommended to strengthen interventions to protect lung function, including smoking cessation at an earlier age to bring about more benefits. PRS-informed intervention may be crucial. After becoming aware of their genetic risk, the high-risk population may actively choose a healthier lifestyle, confirmed in a study of α_1-antitrypsin deficiency [38]. To date, lung function testing is not included in routine COPD screening [39, 40]. As the cost continues to decrease, genome-wide genotyping that only needs to be performed once in a lifetime can provide an evaluation of various phenotypes and may become an additional solution. This study suggests that the PRS was associated with incident COPD in nonsmokers. Therefore, the PRS also provides additional information independent of traditional factors for these populations and individuals may be more active in reducing exposure to other risk factors after receiving this information. Whether PRS-informed early disease screening and intervention can improve COPD underdiagnosis and reduce the overall burden of severe COPD deserves more in-depth research.

Strengths and limitations

Our study has several significant strengths, including the prospective population-based study design, large sample size and detailed information on related covariates.

Some limitations should also be considered. First, smoking behaviours were self-reported and lacked information on tobacco type, inhalation depth and smoking space, and these may cause recall bias and decreased accuracy. In addition, smoking was not randomly assigned and behaviour at baseline may be affected by lung function or other unmeasured variables. Second, rare variants that may have enormous functions were not included in this study, leading to genetic risk estimation inaccuracy. Third, PRS_{lasso} used in this study was developed from the GWAS conducted in the UK Biobank (n=321 047) and SpiroMeta (n=79 055). The samples from the UK Biobank occupied 80% of the study population. This may lead to an over-fitting issue in the current analysis. Fourth, although we introduced lung function testing results during follow-up to compensate for the lack of just relying on a hospital diagnosis to define incident COPD, the UK Biobank has not yet repeated the lung function measurement of all participants and there may still be an underdiagnosis problem. Fifth, incident COPD cases were collected through hospital records and death registries, and some mild cases or cases only receiving primary care could be missed. Failure to obtain information about the severity of the disease may obscure potential dose–response relationships. Sixth, smoking exposure may change during follow-up, leading to deviations in the accurate exposure of risk factors.

Conclusions

The PRS, constructed from 2.5 million variants below genome-wide significance, showed a significant association with incident COPD. The study also found that participants with a high genetic risk may be more susceptible to developing COPD if exposed to smoking.

Acknowledgements: We are grateful to UK Biobank participants. This research has been conducted using the UK Biobank resource (www.ukbiobank.ac.uk) under application number 43795. We are grateful to the International COPD Genetics Consortium (www.copdconsortium.org) for providing the polygenic risk score calculation code.

Author contributions: C. Mao had full access to all the data in the study, and took responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: P-D. Zhang and X-R. Zhang. Acquisition, analysis or interpretation of data: all authors. Drafting of the manuscript: P-D. Zhang and A. Zhang. Critical revision of the manuscript for important intellectual content: all authors. Statistical analysis: P-D. Zhang and Z-H. Li. Obtained funding: C. Mao. Administrative, technical or material support: D. Liu and Y-J. Zhang. Study supervision: C. Mao.

Conflict of interest: None declared.

Support statement: This work was supported by the National Natural Science Foundation of China (81973109), the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2019), the Construction of High-level University of Guangdong (G820332010, G618339167 and G618339164), and the Guangzhou Science and Technology Project (202002030255). Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Li X, Cao X, Guo M, *et al.* Trends and risk factors of mortality and disability adjusted life years for chronic respiratory diseases from 1990 to 2017: systematic analysis for the Global Burden of Disease Study 2017. *BMJ* 2020; 368: m234.
2 Celli BR, Decramer M, Wedzicha JA, *et al.* An Official American Thoracic Society/European Respiratory Society Statement: research questions in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2015; 191: e4–e27.
3 Postma DS, Bush A, van den Berge M. Risk factors and early origins of chronic obstructive pulmonary disease. *Lancet* 2015; 385: 899–909.
4 Løkke A, Lange P, Scharling H, *et al.* Developing COPD: a 25 year follow up study of the general population. *Thorax* 2006; 61: 935–939.
5 Salvi SS, Barnes PJ. Chronic obstructive pulmonary disease in non-smokers. *Lancet* 2009; 374: 733–743.
6 Busch R, Hobbs BD, Zhou J, *et al.* Genetic association and risk scores in a chronic obstructive pulmonary disease meta-analysis of 16 707 subjects. *Am J Respir Cell Mol Biol* 2017; 57: 35–46.
7 Shrine N, Guyatt AL, Erzurumluoglu AM, *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet* 2019; 51: 481–493.

8 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: 696–708.
9 Cho MH, Boutaoui N, Klanderman BJ, *et al.* Variants in *FAM13A* are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010; 42: 200–202.
10 Silverman EK, Vestbo J, Agusti A, *et al.* Opportunities and challenges in the genetics of COPD 2010: an International COPD Genetics Conference report. *COPD* 2011; 8: 121–135.
11 Cho MH, Castaldi PJ, Wan ES, *et al.* A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet* 2012; 21: 947–957.
12 Cho MH, McDonald ML, Zhou X, *et al.* Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med* 2014; 2: 214–225.
13 Soler Artigas M, Wain LV, Miller S, *et al.* Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat Commun* 2015; 6: 8658.
14 Elliott J, Bodinier B, Bond TA, *et al.* Predictive accuracy of a polygenic risk score-enhanced prediction model *vs* a clinical risk score for coronary artery disease. *JAMA* 2020; 323: 636–645.
15 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12: e1001779.
16 Mak TSH, Porsch RM, Choi SW, *et al.* Polygenic scores *via* penalized regression on summary statistics. *Genet Epidemiol* 2017; 41: 469–480.
17 De Matteis S, Jarvis D, Hutchings S, *et al.* Occupations associated with COPD risk in the large population-based UK Biobank cohort study. *Occup Environ Med* 2016; 73: 378–384.
18 Quanjer PH, Stanojevic S, Cole TJ, *et al.* Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012; 40: 1324–1343.
19 Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management and Prevention of COPD. 2018. Available from: http://goldcopd.org/
20 Vogelmeier CF, Criner GJ, Martinez FJ, *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD Executive Summary. *Am J Respir Crit Care Med* 2017; 195: 557–582.
21 Townsend P. Deprivation. *J Soc Policy* 1987; 16: 125–146.
22 van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011; 45: 67.
23 Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982; 69: 239–241.
24 World Health Organization. Metrics: Population Attributable Fraction (PAF). www.who.int/healthinfo/global_burden_disease/metrics_paf/en Date last accessed: 22 August 2020.
25 John C, Soler Artigas M, Hui J, *et al.* Genetic variants affecting cross-sectional lung function in adults show little or no effect on longitudinal lung function decline. *Thorax* 2017; 72: 400–408.
26 Oelsner EC, Ortega VE, Smith BM, *et al.* A genetic risk score associated with chronic obstructive pulmonary disease susceptibility and lung structure on computed tomography. *Am J Respir Crit Care Med* 2019; 200: 721–731.
27 Peterson RE, Kuchenbaecker K, Walters RK, *et al.* Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* 2019; 179: 589–603.
28 Wojcik GL, Graff M, Nishimura KK, *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 2019; 570: 514–518.
29 Aschard H, Tobin MD, Hancock DB, *et al.* Evidence for large-scale gene-by-smoking interaction effects on pulmonary function. *Int J Epidemiol* 2017; 46: 894–904.
30 Wain LV, Shrine N, Miller S, *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* 2015; 3: 769–781.
31 van der Plaat DA, de Jong K, Lahousse L, *et al.* Genome-wide association study on the FEV_1/FVC ratio in never-smokers identifies *HHIP* and *FAM13A*. *J Allergy Clin Immunol* 2017; 139: 533–540.
32 Liu JZ, Tozzi F, Waterworth DM, *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; 42: 436–440.
33 Tobacco GC. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; 42: 441–447.
34 Thorgeirsson TE, Gudbjartsson DF, Surakka I, *et al.* Sequence variants at *CHRNB3–CHRNA6* and *CYP2A6* affect smoking behavior. *Nat Genet* 2010; 42: 448–453.
35 Wilk JB, Shrine NR, Loehr LR, *et al.* Genome-wide association studies identify *CHRNA5/3* and *HTR4* in the development of airflow obstruction. *Am J Respir Crit Care Med* 2012; 186: 622–632.
36 Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to Omnigenic. *Cell* 2017; 169: 1177–1186.
37 Wray NR, Wijmenga C, Sullivan PF, *et al.* Common disease is more complex than implied by the core gene Omnigenic model. *Cell* 2018; 173: 1573–1580.

38 Carpenter MJ, Strange C, Jones Y, *et al.* Does genetic testing result in behavioral health change? Changes in smoking behavior following testing for alpha-1 antitrypsin deficiency. *Ann Behav Med* 2007; 33: 22–28.
39 Qaseem A, Wilt TJ, Weinberger SE, *et al.* Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. *Ann Intern Med* 2011; 155: 179–191.
40 Siu AL, Bibbins-Domingo K, Grossman DC, *et al.* Screening for chronic obstructive pulmonary disease: US Preventive Services Task Force recommendation statement. *JAMA* 2016; 315: 1372–1377.