Development and evaluation of a polygenic risk score for lung cancer in never-smoking women: a large-scale prospective Chinese cohort study

Xiaoxia Wei[1†], Dianjianyi Sun[2,3†], Jiaxin Gao[1†], Jing Zhang[1†], Meng Zhu[1,4], Canqing Yu[2,3], Zhimin Ma[1], Yating Fu[1], Chen Ji[1], Pei Pei[3], Ling Yang[5,6], Iona Y. Millwood[5,6], Robin G.Walters[5,6], Yiping Chen[5,6], Huaidong Du[5,6], Guangfu Jin[1,4], Zhengming Chen[6], Zhibin Hu[1,7], Liming Li[2,3], Hongbing Shen[1,4,8], Jun Lv[2,3 ‡], Hongxia Ma[1,4,8‡]

1 Department of Epidemiology and Biostatistics, International Joint Research Center on Environment and Human Health, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China

2 Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China

3 Peking University Center for Public Health and Epidemic Preparedness & Response, Beijing, China

4 Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Nanjing Medical University, Nanjing, China

5 Medical Research Council Population Health Research Unit at the University of Oxford, Oxford, United Kingdom

6 Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, United Kingdom

7 State Key Laboratory of Reproductive Medicine, Center for Global Health, Nanjing Medical University, Nanjing, China

8 Research Units of Cohort Study on Cardiovascular Diseases and Cancers, Chinese Academy of Medical Sciences, Beijing 100730, China.

† Joint first authors, contributed equally.

‡ Joint last authors, contributed equally.

Correspondence to:

Hongxia Ma, Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Rd., Nanjing 211166, China. Email: hongxiama@njmu.edu.cn; or Jun Lv, Department of Epidemiology and Biostatistics, Peking

University Health Science Center, 38 Xueyuan Road, Beijing 100191, China, Phone: 010-82801620, Email: lvjun@bjmu.edu.cn.

Abbreviations

AUC: area under the curve

BMI: body mass index

CIs: confidence intervals

CKB: China Kadoorie Biobank

COPD: chronic obstructive pulmonary disease

C+T: clumping and thresholding

DHS: DNase I hypersensitivity site

FLCCA: Female Lung Cancer Consortium in Asia

GSA: Global Screening Array

GWASs: Genome-wide association studies

HRs: hazard ratios

LUAD: lung adenocarcinoma

LUSC: lung squamous cell carcinoma

NRI: net reclassification improvement

NSCLC: non-small cell lung cancer

ORs: odds ratios

PCA: principal component analysis

PRS: polygenic risk score

Q-Q: quantile-quantile

SEs: standard errors

Novelty and Impact Statement

We newly developed PRS for lung cancer risk in never-smoking women, and further validated

the performance of PRS in an independent cohort study. The GWAS-derived PRS-21 may provide complementary information to current lung cancer screening guidelines to better identify high-risk groups, especially in never-smoking women.

Abstract

The proportion of lung cancer in never smokers is rising, especially among Asian women, but there is no effective early detection tool. Here, we developed a polygenic risk score (PRS), which may help to identify the population with higher risk of lung cancer in never-smoking women. We first performed a large GWAS meta-analysis (8595 cases and 8275 controls) to systematically identify the susceptibility loci for lung cancer in never-smoking Asian women, and then generated a PRS using GWAS datasets. Furthermore, we evaluated the utility and effectiveness of PRS in an independent Chinese prospective cohort comprising 55,266 individuals. The GWAS meta-analysis identified 8 known loci and 1 new locus (5q11.2) at the genome-wide statistical significance level of $P < 5 \times 10^{-8}$. Based on the summary statistics of GWAS, we derived a polygenic risk score including 21 variants (PRS-21) for lung cancer in never-smoking women. Furthermore, PRS-21 had a hazard ratio per SD of 1.29 (95% CI = 1.18-1.41) in the prospective cohort. Compared with participants who had a low genetic risk, those with an intermediate (HR=1.32, 95% CI: 1.00-1.72) and high (HR=2.09, 95% CI: 1.56-2.80) genetic risk had a significantly higher risk of incident lung cancer. The addition of PRS-21 to the conventional risk model yielded a modest significant improvement in AUC (0.697 to 0.711) and net reclassification improvement (24.2%). The GWAS-derived PRS-21 significantly improves the risk stratification and prediction accuracy for incident lung cancer in never-smoking Asian women, demonstrating the potential for identification of high-risk individuals and early screening.

Keywords: lung cancer in never-smoking women, polygenic risk score, predictive markers

Introduction

Lung cancer is the second most common cancer worldwide, with 2.20 million newly diagnosed cases and 1.80 million deaths estimated in 2020 (*1*). Although tobacco smoking is known as the leading risk factor for lung cancer, the proportion of never-smokers with lung cancer has been increasing over time, especially in Asian countries (*2-4*). Epidemiologic studies indicate that lung cancer in never-smokers is mostly female, adenocarcinoma histologic subtype and East Asian ethnicity (*5*). Additionally, some risk factors have been

investigated to explain the occurrence of lung cancer in never-smokers, such as passive smoking, occupational exposure, indoor and outdoor pollution as well as genetic susceptibility (*6, 7*). Due to the distinct etiological factors between lung cancer in never-smokers and smokers, the challenges associated with prevention are heightened, thereby emphasizing only screening never-smokers identified as at high enough risk to warrant screening.

Polygenic risk score (PRS) is a useful tool to summarize an individual's genetic risk for a complex disease, obtained by aggregating and quantifying the effect of common variants associated with the disease (*8, 9*). In recent years, PRS has been widely used in cancer risk prediction and is capable of identifying a larger fraction of the population with risk equivalent to rare monogenic mutations (*10, 11*). For instance, PRS has been shown to effectively stratify breast cancer risk in women and even refine risk estimates for *BRCA1* and *BRCA2* mutation carriers (*12*). Genome-wide association studies (GWASs) have identified a number of genetic variants associated with the risk of lung cancer in never-smoking women (*13-16*), which offer an opportunity to generate a PRS to identify the subgroups of the population at the highest risk for this disease. A previous study derived a PRS based on 10 significant SNPs from lung cancer risk in never-smoking women in the Asian population (*17*); however, this PRS-10 was derived from a limited number of SNPs, and its efficiency and significance were not validated by prospective studies.

In this study, we performed a large genome-wide association meta-analysis with a total of 8595 cases and 8275 controls to systematically identify genetic factors for the risk of lung cancer in never-smoking Asian women, and then constructed and compared the risk assessment effect of a set of PRSs. Furthermore, we selected PRS-21 to examine its prediction value in a large prospective cohort of the Chinese population, a subset of women from the China Kadoorie Biobank (CKB) and also evaluated the performance of the conventional risk model when integrating PRS.

Methods

Study design and participants

This study was performed in two stages, with a meta-analysis and a prospective cohort study design (**Supplementary Figure 1**). In the first phase, we employed a case-control study design and performed a meta-analysis of data from four independent GWAS datasets, including a total of 8,595 cases and 8,275 controls from NJMU GWAS project (*18*) in never-smoking women in Asia (513 cases and 962 controls), NJMU Global Screening Array (GSA)

Project(*19*) (3,657 cases and 3,271 controls), NJMU OncoArray project (*19*) (296 cases and 301 controls), and Female Lung Cancer Consortium in Asia (FLCCA) GWAS (*14*) (4,129 cases and 3,741 controls). All participants were restricted to never-smoking female lung cancer cases and never-smoking female controls, and comprehensive details for individual studies have been previously reported (*14, 19*). Using the estimated effect sizes of the summary statistics from this GWAS meta-analysis, we derived a set of PRSs for lung cancer in never-smoking women.

In the second stage of the study, we further evaluated the prediction performance of PRSs in an independent cohort study of CKB. Details of the design and methods of the CKB have been described previously (*20*). Briefly, 512,715 adults aged 30-79 years were recruited during 2004-2008 across ten geographically diverse (five urban and five rural) regions in China. At baseline, all participants completed a laptop-based questionnaire on demographic characteristics, lifestyle factors, biomedical information, and provided physical measurements and a blood sample. Among those, 55,266 never-smoking female individuals with both genotypic and phenotypic data remained in the present study. The outcome of our study was lung cancer (ICD-10: C34). Participants were assessed for lung cancer risk between baseline and date of diagnosis, death, loss to follow-up, or study end date (December 31, 2016), whichever came first.

Genotyping and quality control

Multiple genotyping arrays for genome-wide genotyping were used in the four GWAS and CKB study samples (**Supplementary Table 1**). Detailed descriptions of the quality control procedures on genotyping data for each GWAS have been described previously (*14, 19*). Briefly, we performed quality control at both sample-level and variant-level. Qualified genotypes were phased with SHAPEIT v2 (*21, 22*) and imputation was performed with IMPUTE v2 (*23*) by using the reference of the 1000 Genomes Project Phase III database (October 2014 release).

Polygenic risk score

Based on the summary statistics from the meta-analysis of GWAS, we derived the PRS for lung cancer in never-smoking women, including two components: (i) derived PRSs using a clumping and thresholding (C+T) approach, with a range of p-value (5×10^{-8}, 5×10^{-7}, 5×10^{-6}, 5×10^{-5}, and 5×10^{-4}) and linkage disequilibrium-driven clumping procedure in PLINK version 1.90b (--clump), retaining the SNP with the smallest p-value excluding variants with $r^2 > 0.1$ in a 1000-kb window (**Supplementary Table 2-6**). The PRS with the highest area under the curve (AUC) and the lowest *P* value in the cohort study was selected as the best-performing

PRS. The best-performing PRS was obtained at a *P*-value threshold of 5×10^{-7}, including 16 variants (PRS-16). (ii) additional SNPs that were previously reported in lung cancer GWAS and replicated in our current study at *P* value <0.00058 (after multiple comparisons at 0.05/86, 86 were the sum of 16 SNPs included in best-performing PRS and 71 reported lung cancer susceptibility SNPs(*19*), including one duplicate SNP). When correlation exists (r^2 >0.1), 5 reported lung cancer susceptibility SNPs (rs2293607, rs2293607, rs4236709, rs10429489, rs1200399, and rs77468143) representing independent loci with the strongest statistical significance were retained. The final PRS (PRS-21) was constructed by combining above both components. The list of known lung cancer-related loci and the SNPs included in PRS-21 were shown in **Supplementary Table 7** and **Supplementary Table 8,** respectively.

Polygenic risk scores were generated by multiplying the dosage of the effect allele for each SNP by its respective weight (log OR). In addition, we compared the predictive performance of our PRS with two known PRSs from Asian populations: PRS-10 for lung cancer in never-smoking women (*17*) and PRS-19 for non-small cell lung cancer (NSCLC) (*19*).

Statistical analysis

For each GWAS dataset, a single-variant association analysis was performed using the SNPTEST software (V.2.54) based on a log-additive model (*24*). Per-allele odds ratios (ORs) and standard errors (SEs) from each GWAS dataset were combined by a fixed-effects meta-analysis using the METAL software (*25*). The genome-wide significance threshold was set at $P < 5.0 \times 10^{-8}$. For the CKB cohort study, we performed Cox proportional hazards regression models to assess the association between PRS and risk of lung cancer in never-smoking women to estimate hazard ratios (HRs) and 95% confidence intervals (CIs) with adjustment for age, education, body mass index (BMI), chronic obstructive pulmonary disease (COPD), personal history of cancer, family history of cancer, and the first ten principal components of ancestry determined through principal component analysis. The assumption of proportional hazards was assessed by testing the significance of Schoenfeld residuals.

A potential non-linear association between the PRS and the risk of lung cancer in never-smoking women was assessed by fitting a restricted cubic spline (*26*). The participants were categorized into low (quintile 1), intermediate (quintiles 2–4), and high (quintile 5) genetic risk according to quintiles of the PRS, and those with the lowest PRS as the reference group. Effect modification of the PRS by the epidemiologic risk factors was evaluated by fitting additional interaction terms in the model. We also estimated the 5-year and lifetime absolute

risks (up to 80 years of age) of developing lung cancer in never-smoking women based on Cox proportional hazards model.

In addition, we evaluated the incremental predictive ability of PRS for improving the conventional risk model on incident lung cancer in never-smoking women. At first, we selected risk factors based on prior evidence to construct a conventional risk model (*27-29*), which included age, education, BMI, COPD history, personal history of cancer, and family history of cancer. Then, AUC and net reclassification improvement (NRI) were used to estimate the discriminatory accuracy when PRS was added to the conventional risk model.

To evaluate the robustness of our results, we carried out several further sensitivity analyses: (1) excluding incident cases occurring during the first year of follow-up; (2) reclassifying genetic risk levels based on quartiles (bottom, 2-3, and top quartiles defined as low, intermediate, and high genetic risk, respectively) or tertiles (corresponding to low, intermediate and high genetic risk) of PRSs; (3) using a 10-fold cross-validation approach to avoid over-fitting of the risk model.

All p-values were two-sided and $p<0.05$ was considered statistically significant. All statistical analyses were performed with R software, version 3.5.1.

Results

Characteristics of the subjects

The GWAS meta-analysis included 8,595 never-smoking female lung cancer cases and 8,275 never-smoking controls from four independent GWAS datasets. The characteristics of the participants are presented in **Supplementary Table 9**, in which 86.46% of cases were classified as lung adenocarcinoma (LUAD) and 11.34% as lung squamous cell carcinoma (LUSC). For the PRS application, a total of 55,266 never-smoking females with complete genotypes and phenotypes from the CKB cohort were included in the analysis. During a median follow-up of 10.53 years (IQR 9.58-11.38), 466 incident lung cancer cases occurred (**Supplementary Table 10**).

GWAS meta-analysis

A quantile-quantile (Q-Q) showed minimal evidence of systematic inflation from population stratification for the meta-analysis (λ = 1.02, **Supplementary Figure 2**). The principal component analysis (PCA) revealed good clustering of all case and control samples within the same cluster (**Supplementary Figure 3**). The GWAS meta-analysis identified 9 risk loci for lung cancer in never-smoking women at the genome-wide statistical significance threshold ($P < 5 \times 10^{-8}$) (**Table 1 and Supplementary Figure 4**). Of those, we confirmed the

association of 8 previously reported loci (3q28, 5p15.33, 6p21.32, 6p21.32, 6p21.1, 6q22.1, 10q25.2, and 11q23.3) and identified one novel risk loci at 5q11.2 (lead variant: rs1498606, $P = 4.79 \times 10^{-8}$). The lead SNP rs1498606 is located in the intronic region of the *PDE4D* gene on chromosome 5q11.2 (**Figure 1 A). Figure 1 B** depicted rs1498606 associated with the risk of lung cancer in never-smoking women in each GWAS study.

To annotate the underlying biological significance of novel loci in 5q11.2, we evaluated the potential functions of the lead variant rs1498606 and their correlated variants ($r^2 \geq 0.60$) and found that rs6450500 (with rs1498606 $r^2 = 1.0$) is more likely to be a functional SNP that overlapped with DNase I hypersensitivity site (DHS) or promoter/enhancer histone marks (eg, H3K4me1, H3K4me3, and H3K27ac) in lung tissues and transcription factor binding sites (S**upplementary table 11 and supplementary figure 5**).

PRS construction and evaluation

We generated a set of PRSs using effect size derived from GWAS meta-analysis at different significance thresholds. Then, we evaluated the PRSs in the CKB cohort and found that all five GWAS-derived PRSs were significantly associated with the risk of lung cancer in never-smoking women (**Table 2)**. The best-performing PRS was obtained at a *P* value threshold of 5×10^{-7} based on the C+T approach, resulting in a 16-SNP PRS (PRS-16), which showed the strongest association ($HR_{per\ 1\ SD}$=1.28, 95% CI: 1.17-1.40; AUC=0.573). When the 16-SNP PRS was supplemented with 71 reported lung cancer susceptibility SNPs, a new PRS with 21 SNPs (PRS-21) was further derived and showed improved prediction accuracy ($HR_{per\ 1\ SD}$=1.29, 95% CI: 1.18-1.41; AUC=0.576), which was ultimately used to define genetic risk for all analyses. Moreover, the PRS-21 had better performance prediction than the two published PRSs.

The polygenic risk score derived from 21 SNPs could significantly predict risk of lung cancer among never-smoking women in the CKB cohort. Incident lung cancer cases in never-smoking women had a higher PRS-21 than those without incident lung cancer (**Figure 2 A**). A linear and positive association between the PRS-21 and the risk of incident lung cancer in never-smoking women using restricted cubic spline regression was found ($P_{non\text{-}linear}$=0.8762; **Figure 2 B**). We also observed a dose-response relationship between PRS quintiles and the risk of lung cancer in never-smoking women ($P_{trend}= 5.90 \times 10^{-8}$) (**Figure 2 C and Supplementary Table 12**). Additionally, we assessed the effect of the PRS by categorizing the PRS distribution into low (bottom quintile), intermediate (quintiles 2-4) and high (top quintile) genetic risk groups. Compared with participants who had a low genetic risk, those

with an intermediate (HR=1.32, 95% CI: 1.00-1.72) and high (HR=2.09, 95% CI: 1.56-2.80) genetic risk had a significantly greater risk of incident lung cancer (**Figure 2 D and Supplementary Table 13**). These associations did not materially change when excluding incident cases that occurred in the first year of follow-up (**Supplementary table 14**). Similarly, reclassification of genetic risk by dividing PRS into quartiles or tertiles also showed a positive association with the risk of lung cancer in never-smoking women (**Supplementary table 15**). Positive associations were also observed in stratified analyses including age, education, BMI, COPD history, personal history of cancer, and family history of cancer, and there was no evidence of interaction (**Supplementary table 16**).

Estimated 5-year and lifetime absolute risks of lung cancer in never-smoking women

Estimated 5-year and lifetime absolute risks for different PRS categories among never-smoking women in the CKB cohort are shown in **Figure 3**. The 5-year absolute risk for never-smoking women in the highest 1% of the PRS reached 2.48% (**Figure 3A**). Furthermore, we assessed how the PRS affected lifetime trajectories of lung cancer in never-smoking women at risk. The cumulative risks by age of 80 for lung cancer events were 2.26% for individuals with low genetic risk, 2.95% for individuals with intermediate genetic risk, and 4.71% among those with high genetic risk, respectively (**Figure 3B**).

Incremental value of PRS in risk prediction

We further evaluated the incremental predictive ability of PRS-21 for lung cancer in never-smoking women. **Supplementary Table 17** lists the model parameters in the CKB cohort. The addition of the PRS to this baseline model containing conventional risk factors showed a statistically significant improvement in discrimination, with the AUC improved from 0.697 to 0.711 (difference, 1.4%; $P<0.001$) (**Supplementary Figure 6**). Similar results were observed in analyses using the 10-fold cross-validation approach **(Supplementary Table 18)**. Furthermore, there was a significant reclassification improvement, with an overall improvement in the net risk stratification of 24.2% (13.7%-34.8%) (**Supplementary Table 19**).

Discussion

In this study, leveraging large GWAS data sets of women with Asian ancestry, we identified a novel risk locus and newly developed PRS for lung cancer risk in never-smoking women, and further validated the performance of PRS in an independent prospective cohort study. Our results suggested that GWAS-derived PRS-21 is associated with the risk of lung cancer in never-smoking women with a dose-response relationship and is powerful to identify high-risk

individuals. Furthermore, adding PRS to the conventional risk model can provide a modest improvement in the prediction of lung cancer in never-smoking women.

Most of the existing screening criteria for lung cancer, both the NLST (*27*) and USPSTF (*30*), focus on heavy smokers, while the proportion of non-smoking lung cancer is increasing, and there has been no clear guideline for screening criteria. The PRS can identify individuals with relatively high risk and provide valuable information for individual risk assessments to guide individualized screening strategies. A previous study derived a PRS for lung cancer in never-smoking Asian women based on 10 SNP and examined the interaction between the PRS and household coal use(*17*); however, the performance of PRS lacked validation in an independent cohort study and thus its prediction value could not be examined. Besides, our previous study used 19 SNPs to construct a PRS for NSCLC and showed predictive effects in the CKB cohort comprising 95,408 individuals (*19*). In the present study, we generated a PRS with 21 SNPS based on a large-sample, multi-center GWAS dataset of women with Asian ancestry, and proved that the GWAS-derived PRS significantly predicted the risk of lung cancer in never-smoking women in an independent nationwide prospective cohort study, and had better performance prediction than the two known PRSs. Moreover, the PRS-21 displayed the ability to substantially stratify lung cancer in never-smoking women risk trajectories within each genetic risk category. Participants in the top 20% of PRS-21 compared with those in the bottom 20% had a 2.09-fold risk of developing lung cancer. Furthermore, our study revealed that individuals within the top 1% of the PRS exhibit a five-year lung cancer risk of 2.48%, a significant enough value to warrant the implementation of lung cancer screening. (*27*). Consequently, the utilization of genetic risk score may aid in formulating a tailored screening strategy for never-smoking women, taking into account their unique risk factors. These findings suggest that PRS-21 may provide complementary information to current lung cancer screening guidelines to better identify high-risk groups, especially in never-smoking women.

The incorporation of PRS and nongenetic risk factors have been reported to improve the risk discrimination of diseases such as coronary heart disease (*8*), breast (*31*), gastric (*11*) and esophageal cancer (*32*). In a recent analysis based on the UK Biobank prospective cohort, adding the 128-SNP European PRS to the epidemiologic risk factors model yielded an increment of AUC of about 1.7% in never-smoking lung cancer events (*33*). In our analyses, we also demonstrated that PRS-21 was a significant predictor of lung cancer in never-smoking Asian women events, with a modest improvement in prediction accuracy when the

PRS was added to a conventional model (difference AUC, 1.4%; $P<0.001$). Further studies to validate a comprehensive model for individualized risk prediction based on the combined effects of genetic and conventional risk factors will be needed.

In addition, it should be noted that the new locus at 5q11.2 identified by our GWAS meta-analysis is associated with the risk of breast cancer (*34*) and epithelial ovarian cancer (*35*), but the first time involved in the risk of lung cancer in never-smoking women. The lead SNP rs1498606 is located in the intronic region of the *PDE4D* gene on chromosome 5q11.2, and functional annotation showed that rs6450500 (with rs1498606 $r^2 = 1.0$) at 5q11.2 localizes in a region with histone modification signatures, suggesting that the locus has transcriptional enhancer and promoter activity. *PDE4D*, a tumor-promoting factor (*36*), plays a role in cancer pathophysiology by altering tumor cell proliferation (*37, 38*), tumor cell differentiation (*39*), and loss of E-cadherin (*40*). However, there is currently a lack of research on the relationship between target genes and the pathogenesis of lung cancer in never-smoking women. Therefore, further studies to elucidate the underlying biological mechanisms are warranted.

Our study has several important strengths. We developed PRS based on a large-sample, multi-center GWAS database of lung cancer in never-smoking women, which provided robust effect size estimates for SNPs. Furthermore, the validation samples from the CKB cohort were utterly independent of the GWAS meta-analysis, which avoided overfitting of the PRS effect and confirmed the predictive value of the PRS. However, we also acknowledged several limitations of this study. First, missing genotype data were imputed in GWAS using different genotyping arrays, which may lead to partial deviation in estimates of SNP effects. Second, we only assessed the overall risk of lung cancer, but the risk estimates might differ by tumor stage or subtype; however, this information was not available in the CKB cohorts, making further analysis limited. Third, the construction and evaluation of the conventional risk model were constructed in the CKB cohort, which may lead to overfitting to some extent. However, this would not affect the discriminative power of assessing the addition of PRS to the conventional risk model.

In summary, we developed a GWAS-derived PRS and validated its effective prediction value in a large cohort for the risk of lung cancer in never-smoking women. These findings suggest that the PRS-21 may be useful for identifying high-risk individuals and early screening of lung cancer in never-smoking women.

Author Contributions

XW, DS, JG, JL and HM contributed to the study design and sample collection, and

supervised the whole project. XW, JZ, MZ, and HM contributed to the data interpretation, data analysis, and writing of the manuscript. CY, ZM, YF, CJ, PP, LY, IM, RW, YC, HD, GJ, ZC, ZH, LL, and HS contributed to the study design, sample collection, data interpretation of the present analysis. All of the authors reviewed or revised the manuscript. The work reported in the paper has been performed by the authors, unless clearly specified in the text.

Acknowledgments

We thank all the study participants and research staff for their contributions and commitment to the present study.

Funding

This work was supported by National Natural Science Foundation of China (81922061, 81973123, 81803306); Natural Science Foundation of Jiangsu Province (BK20180675); CAMS Innovation Fund for Medical Sciences (2019RU038); and National Science Foundation for Post-doctoral Scientists of China (2018M640466).

Conflict of interest

No competing interests to declare.

Ethics Statement

Written informed consent was obtained from all participants in this study. The institutional review boards of Nanjing Medical University approved the study protocol. The China Kadoorie Biobank was approved by the Ethical Review Committee of the Oxford Tropical Research Ethics Committee, University of Oxford, and the Chinese Center for Disease Control and Prevention.

Data Availability Statement

The GWAS summary statistics are available through the NHGRI-EBI GWAS Catalog under study accession number GCST90277434. The acquisition policies and procedures of China Kadoorie Biobank are available at www.ckbiobank.org. Further information is available from the corresponding author upon request.

References

1. H. Sung, J. Ferlay, R. L. Siegel *et al.*, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians.* 2021; **71**: 209-249.
2. R. Zheng, S. Zhang, H. Zeng *et al.*, Cancer incidence and mortality in China, 2016. *Journal of the National Cancer Center.* 2022; **2**: 1-9.
3. C. K. Toh, W. S. Ong, W. T. Lim *et al.*, A Decade of Never-smokers Among Lung Cancer Patients-Increasing Trend and Improved Survival. *Clinical lung cancer.* 2018; **19**: e539-e550.
4. H. A. Wakelee, E. T. Chang, S. L. Gomez *et al.*, Lung cancer incidence in never smokers. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* 2007; **25**: 472-478.
5. D. Planchard, B. Besse, Lung cancer in never-smokers. *The European respiratory journal.* 2015; **45**: 1214-1217.
6. J. Subramanian, R. Govindan, Lung cancer in never smokers: a review. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* 2007; **25**: 561-570.
7. S. Sun, J. H. Schiller, A. F. Gazdar, Lung cancer in never smokers--a different disease. *Nature reviews. Cancer.* 2007; **7**: 778-790.
8. X. Lu, Z. Liu, Q. Cui *et al.*, A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study. *European heart journal.* 2022; **43**: 1702-1711.
9. N. Mavaddat, K. Michailidou, J. Dennis *et al.*, Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *American journal of human genetics.* 2019; **104**: 21-34.
10. A. V. Khera, M. Chaffin, K. G. Aragam *et al.*, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics.* 2018; **50**: 1219-1224.
11. G. Jin, J. Lv, M. Yang *et al.*, Genetic risk, incident gastric cancer, and healthy lifestyle: a meta-analysis of genome-wide association studies and prospective cohort study. *The Lancet. Oncology.* 2020; **21**: 1378-1386.
12. N. Zeinomar, W. K. Chung, Cases in Precision Medicine: The Role of Polygenic Risk Scores in Breast Cancer Risk Assessment. *Annals of internal medicine.* 2021; **174**: 408-412.

13. W. J. Seow, K. Matsuo, C. A. Hsiung *et al.*, Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. *Human molecular genetics.* 2017; **26**: 454-465.
14. Q. Lan, C. A. Hsiung, K. Matsuo *et al.*, Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nature genetics.* 2012; **44**: 1330-1335.
15. K. Shiraishi, H. Kunitoh, Y. Daigo *et al.*, A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nature genetics.* 2012; **44**: 900-903.
16. C. A. Hsiung, Q. Lan, Y. C. Hong *et al.*, The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia. *PLoS Genet.* 2010; **6**.
17. B. Blechter, J. Y. Y. Wong, C. Agnes Hsiung *et al.*, Sub-multiplicative interaction between polygenic risk score and household coal use in relation to lung adenocarcinoma among never-smoking women in Asia. *Environment international.* 2021; **147**: 105975.
18. Z. Hu, C. Wu, Y. Shi *et al.*, A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nature genetics.* 2011; **43**: 792-796.
19. J. Dai, J. Lv, M. Zhu *et al.*, Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *The Lancet. Respiratory medicine.* 2019; **7**: 881-891.
20. Z. Chen, J. Chen, R. Collins *et al.*, China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol.* 2011; **40**: 1652-1666.
21. O. Delaneau, J. F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods.* 2013; **10**: 5-6.
22. O. Delaneau, J. Marchini, J. F. Zagury, A linear complexity phasing method for thousands of genomes. *Nature methods.* 2011; **9**: 179-181.
23. B. N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics.* 2009; **5**: e1000529.
24. J. Marchini, B. Howie, S. Myers *et al.*, A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics.* 2007; **39**: 906-913.

25. C. J. Willer, Y. Li, G. R. Abecasis, METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England).* 2010; **26**: 2190-2191.
26. H. Heinzl, A. Kaider, Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer methods and programs in biomedicine.* 1997; **54**: 201-208.
27. M. C. Tammemägi, T. R. Church, W. G. Hocking *et al.*, Evaluation of the lung cancer risks at which to screen ever- and never-smokers: screening rules applied to the PLCO and NLST cohorts. *PLoS medicine.* 2014; **11**: e1001764.
28. L. H. Chien, C. H. Chen, T. Y. Chen *et al.*, Predicting Lung Cancer Occurrence in Never-Smoking Females in Asia: TNSF-SQ, a Prediction Model. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* 2020; **29**: 452-459.
29. F. Wang, F. Tan, S. Shen *et al.*, Risk-stratified Approach for Never- and Ever-Smokers in Lung Cancer Screening: A Prospective Cohort Study in China. *American journal of respiratory and critical care medicine.* 2023; **207**: 77-88.
30. V. A. Moyer, Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of internal medicine.* 2014; **160**: 330-338.
31. A. Lee, N. Mavaddat, A. N. Wilcox *et al.*, BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in medicine : official journal of the American College of Medical Genetics.* 2019; **21**: 1708-1718.
32. J. Dong, M. F. Buas, P. Gharahkhani *et al.*, Determining Risk of Barrett's Esophagus and Esophageal Adenocarcinoma Based on Epidemiologic Factors and Genetic Variants. *Gastroenterology.* 2018; **154**: 1273-1281.e1273.
33. R. J. Hung, M. T. Warkentin, Y. Brhane *et al.*, Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model. *Cancer research.* 2021; **81**: 1607-1615.
34. M. Xu, Y. Xu, M. Chen *et al.*, Association study confirms two susceptibility loci for breast cancer in Chinese Han women. *Breast cancer research and treatment.* 2016; **159**: 433-442.
35. J. B. Permuth, A. Pirie, Y. Ann Chen *et al.*, Exome genotyping arrays to identify rare and low frequency variants associated with epithelial ovarian cancer risk. *Human molecular genetics.* 2016; **25**: 3600-3612.

36. D. C. Lin, L. Xu, L. W. Ding *et al.*, Genomic and functional characterizations of phosphodiesterase subtype 4D in human cancers. *Proceedings of the National Academy of Sciences of the United States of America.* 2013; **110**: 6109-6114.
37. N. He, N. Kim, M. Song *et al.*, Integrated analysis of transcriptomes of cancer cell lines and patient samples reveals STK11/LKB1-driven regulation of cAMP phosphodiesterase-4D. *Molecular cancer therapeutics.* 2014; **13**: 2463-2473.
38. S. S. Pullamsetti, G. A. Banat, A. Schmall *et al.*, Phosphodiesterase-4 promotes proliferation and angiogenesis of lung cancer by crosstalk with HIF. *Oncogene.* 2013; **32**: 1121-1134.
39. F. Baty, D. Klingbiel, F. Zappa *et al.*, High-throughput alternative splicing detection using dually constrained correspondence analysis (DCCA). *Journal of biomedical informatics.* 2015; **58**: 175-185.
40. E. Kolosionek, R. Savai, H. A. Ghofrani *et al.*, Expression and activity of phosphodiesterase isoforms during epithelial mesenchymal transition: the role of phosphodiesterase 4. *Molecular biology of the cell.* 2009; **20**: 4751-4765.

502 **Table 1. Association of GWAS identified SNPs and risk of lung cancer in never-smoking women**

Locus	Variants	Position [a]	Gene	EA	OA	EAF	OR (95%CI)	*P* value
New loci								
5q11.2	rs1498606	58337549	PDE4D	C	T	0.29	1.15 (1.09-1.20)	4.79E-08
Known loci								
3q28	rs11928222	189350265	TP63	G	T	0.43	1.21 (1.15-1.26)	3.09E-16
5p15.33	rs7726159	1282319	CMTPM1L/TERT	A	C	0.38	1.42 (1.35-1.49)	1.10E-46
6p21.32	rs9275164	32652929	HLA-DQB1	C	T	0.34	1.20 (1.14-1.26)	8.43E-13
6p21.1	rs2496646	41483367	FOXP4	T	C	0.35	1.18 (1.12-1.24)	3.21E-10
6q22.1	rs9374663	117782634	ROS1/DCBLD1	A	G	0.50	1.15 (1.10-1.20)	1.56E-09
10q25.2	rs11196063	114460013	VTI1A	C	A	0.28	1.27 (1.21-1.33)	1.89E-21
11q23.3	rs55768116	118108331	AMICA1	C	A	0.45	1.15 (1.09-1.20)	3.92E-09
17q24.2	rs72845278	65877076	BPTF	T	C	0.71	1.20 (1.13-1.26)	9.09E-11

503 *Abbreviations*: EA, effect allele; OA, other allele; EAF, effect allele frequency; OR, odds (log-additive)
504 ratio; 95% CI, 95% confidence interval.
505 [a] Chromosome position, hg19/GRCh37 build.
506

Table 2. Association of different polygenic risk scores with risk of lung cancer in never-smoking women in the CKB cohort

PRS development methods	N SNPs in the PRS	HR (95% CI) [a]	*P* value	AUC
Clumping and thresholding approach [b]				
5×10^{-8}	12	1.24 (1.13-1.35)	2.75E-06	0.565
5×10^{-7}	16	1.28 (1.17-1.40)	9.37E-08	0.573
5×10^{-6}	44	1.25 (1.14-1.36)	1.60E-06	0.568
5×10^{-5}	202	1.16 (1.06-1.26)	1.77E-03	0.541
5×10^{-4}	1448	1.08 (0.98-1.18)	1.16E-01	0.524
Integrate PRS				
PRS-21	21	1.29 (1.18-1.41)	4.40E-08	0.576
Published PRS				
PRS-10(*17*)	10	1.21 (1.11-1.33)	3.77E-05	0.557
PRS-19(*19*)	19	1.20 (1.09-1.09)	1.31E-04	0.547

Abbreviations: CI, confidence interval; HR, hazard ratio; PRS, polygenic risk score; SNP, single nucleotide polymorphism.

[a] Adjusted for age, education, BMI, COPD history, personal history of cancer, family history of cancer, and the first ten principal components of ancestry. HR denotes the increase of hazard risk of lung cancer in never-smoking women per SD for PRS

[b] Polygenic risk score (PRS) derived from different significance thresholds in the GWAS meta-analysis of lung cancer in never-smoking women

Figure 1. Regional plot and forest plot for rs1498606. (A) Regional plot of chromosome 5q11.2 locus. The index SNP rs1498606 is colored purple, with other SNPs colored according to the degree of linkage disequilibrium r^2 with the index SNP. The r^2 values are calculated based on the 1000 Genome data. Each point represents a variant with chromosomal position on the x axis and the –log (P value) on the y axis. (B) Forest plot for rs1498606 association with risk of lung cancer in never-smoking women in each GWAS study.

Figure 2. The distribution and relationship of polygenic risk score (PRS) with incident lung cancer in never-smoking women in the CKB cohort. (A) Distribution of PRS in participants affected with or without lung cancer in never-smoking women in the CKB cohort; (B) Linear relationship between PRS and risk of lung cancer in never-smoking women was assessed using a restricted cubic spline analysis in the CKB cohort; (C) Participants in the CKB cohort were divided into five equal groups according to their polygenic risk score, and the HRs for each group were compared with those in quintile 1of the polygenic risk score; error bars show the 95% CIs; (D) The standardized cumulative incidence of lung cancer in never-smoking women in low (quintile 1), intermediate (quintiles 2–4), and high (quintile 5) genetic risk groups in the CKB cohort during median 10.5 years of follow-up; shaded regions represent the 95% CIs, and hazard ratios (HRs) were estimated with adjustment for age, education, BMI, COPD history, personal history of cancer, family history of cancer, and the first ten principal components of ancestry.

Figure 3. Absolute risk estimates of lung cancer in never-smoking women in the CKB cohort. (A) Five-year absolute risk; (B) Cumulative absolute risk until age 80. Cox proportional hazard model was used to estimate the HRs (95% CIs) and the cumulative risk adjusted for age, education, BMI, COPD history, personal history of cancer, family history of cancer, and the first ten principal components of ancestry. Genetic risk categories: low (quintile 1), intermediate (quintiles 2-4), and high (quintile 5) risk according to quintiles of PRS.