




ORIGINAL ARTICLE

Lung cancer risk score for ever and never smokers in China

Zhimin Ma^{1,2,3,#} | Jun Lv^{4,5,#} | Meng Zhu^{1,2,#} | Canqing Yu⁴ | Hongxia Ma^{1,2}  | Guangfu Jin^{1,2} | Yu Guo⁶ | Zheng Bian⁶ | Ling Yang⁷ | Yiping Chen⁷ | Zhengming Chen⁷ | Zhibin Hu^{1,2}  | Liming Li^{4,7} | Hongbing Shen^{1,2,8} 

¹Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, P. R. China

²Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, Jiangsu, P. R. China

³Department of Epidemiology, School of Public Health, Southeast University, Nanjing, Jiangsu, P. R. China

⁴Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing, P. R. China

⁵Ministry of Education, Key Laboratory of Molecular Cardiovascular Sciences (Peking University), Beijing, P. R. China

⁶Chinese Academy of Medical Sciences, Beijing, P. R. China

⁷Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford, Oxfordshire, UK

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

Correspondence

Hongbing Shen, Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, Jiangsu, P. R. China.

Email: hbshen@njmu.edu.cn

Liming Li, Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing 100191, P. R. China

Email: lmlee@vip.163.com

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 81820108028, 81922061, 81973123, 82273714,

Abstract

Background: Most lung cancer risk prediction models were developed in European and North-American cohorts of smokers aged ≥ 55 years, while less is known about risk profiles in Asia, especially for never smokers or individuals aged < 50 years. Hence, we aimed to develop and validate a lung cancer risk estimate tool for ever and never smokers across a wide age range.

Methods: Based on the China Kadoorie Biobank cohort, we first systematically selected the predictors and explored the nonlinear association of predictors with lung cancer risk using restricted cubic splines. Then, we separately developed risk prediction models to construct a lung cancer risk score (LCRS) in 159,715 ever smokers and 336,526 never smokers. The LCRS was further validated in an independent cohort over a median follow-up of 13.6 years, consisting of 14,153 never smokers and 5,890 ever smokers.

List of abbreviations: USPSTF, US Preventive Services Task Force; AUC, area under the receiver operating curve; ROC, receiver operating characteristic; CKB, China Kadoorie Biobank; CI, confidence interval; HR, hazard ratio; ICD-10, International Statistical Classification of Diseases and Related Health Problems 10th Revision; BMI, body-mass index; MET, metabolic equivalent of task; LCRS, lung cancer risk score; LCKEY, Lung Cancer Risk Evaluation by Yourself; NNS, number needed to be screened; PLCO, Prostate Lung Colorectal and Ovarian Cancer Screening Trial; N/A, not applicable; SD, standard deviation.

#These authors contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Cancer Communications* published by John Wiley & Sons Australia, Ltd. on behalf of Sun Yat-sen University Cancer Center.

82192901, 82192904, 82192900; Excellent Youth Foundation of Jiangsu Province, Grant/Award Number: BK20220100; Research Unit of Prospective Cohort of Cardiovascular Diseases and Cancer; Chinese Academy of Medical Sciences, Grant/Award Number: 2019RU038; Science and Technology Service Network Initiative of Chinese Academy of Sciences, Grant/Award Number: No.KFJ-STS-QYZD-2021-08-001; the National Key Research and Development Program of China, Grant/Award Number: 2016YFC0900500

Results: A total of 13 and 9 routinely available predictors were identified for ever and never smokers, respectively. Of these predictors, cigarettes per day and quit years showed nonlinear associations with lung cancer risk ($P_{\text{non-linear}} < 0.001$). The curve of lung cancer incidence increased rapidly above 20 cigarettes per day and then was relatively flat until approximately 30 cigarettes per day. We also observed that lung cancer risk declined sharply within the first 5 years of quitting, and then continued to decrease but at a slower rate in the subsequent years. The 6-year area under the receiver operating curve for the ever and never smokers' models were respectively 0.778 and 0.733 in the derivation cohort, and 0.774 and 0.759 in the validation cohort. In the validation cohort, the 10-year cumulative incidence of lung cancer was 0.39% and 2.57% for ever smokers with low (< 166.2) and intermediate-high LCRS (≥ 166.2), respectively. Never smokers with a high LCRS (≥ 21.2) had a higher 10-year cumulative incidence rate than those with a low LCRS (< 21.2 ; 1.05% vs. 0.22%). An online risk evaluation tool (LCKEY; <http://ccra.njmu.edu.cn/lckey/web>) was developed to facilitate the use of LCRS.

Conclusions: The LCRS can be an effective risk assessment tool designed for ever and never smokers aged 30 to 80 years.

KEYWORDS

early-onset cancer, lung cancer screening, lung cancer, never smokers, prediction model

1 | BACKGROUND

Lung cancer is the leading cause of cancer-related death. In 2020, almost one-third of the world's lung cancer cases and cancer-related deaths occurred in China [1, 2]. Two large lung cancer screening trials showed that high-risk populations screened by low-dose computed tomography could reduce lung cancer mortality [3, 4]. Current screening programs use simplified inclusion criteria to select high-risk individuals. The US Preventive Services Task Force (USPSTF) criteria used age (50-80 years), smoking (≥ 20 pack-years), and quit years (< 15 years) to screen high-risk individuals [5]. In China, the criterion for lung cancer screening was defined as age 50-74 years, smoked ≥ 30 pack-years, and quit smoking < 15 years ago; or one who passively smoked > 20 years; or one with chronic obstructive pulmonary disease; or one with occupational exposure; or one with a family history of lung cancer [6].

Recent studies have showed that the existing lung cancer risk prediction models, which include well-known risk factors, are more sensitive for early detection than the simplified criteria [7-9]. Furthermore, risk prediction models could provide a risk estimate tool to inform individuals about their specific risk [10]. Most of existing models were developed and validated in the West [9, 11-14], while few lung cancer risk prediction models were derived from

Asian populations, and they have rarely been externally validated [15, 16]. Moreover, lung cancer incidence and risk associated with tobacco consumption in Asian populations differ from those in Western populations [17, 18]. Thus, there is still a need to develop a Chinese lung cancer risk prediction model to provide information on risk assessment, in order to promote the personalized prevention of lung cancer in China.

To make wider populations benefit from smoking cessation or early intervention on other modifiable risk factors for lung cancer, the existing risk prediction models and lung cancer screening criteria still require further improvement. First, early-onset cases (< 50 years at diagnosis) represented 6.7% to 13.4% of diagnosed lung cancers [19, 20], but most lung cancer risk models were designed for smokers aged 55-74 years [14, 21, 22], and individuals aged less than 50 were precluded from benefiting from personalized risk assessment or screening. Previous randomized controlled trials and meta-analysis studies showed that communication of personalized disease risk can motivate smoking cessation [23, 24]. Second, there is an increasing proportion of lung cancer among never smokers [25], especially in Asia with over 40% of lung cancers occurring among never smokers [26]. Thus, a tool for accurately selecting individuals at high risk of lung cancer was also warranted, regardless of their smoking status.

Therefore, we used a large nationwide prospective cohort of China to develop a lung cancer risk prediction model, separately for ever smokers and never smokers across a wide age range. Furthermore, each model was validated in another prospective Chinese cohort. This study aimed to facilitate risk assessment for lung cancer screening, ultimately, leading to reductions in lung cancer morbidity and mortality.

2 | MATERIALS AND METHODS

2.1 | Study population

2.1.1 | Data for model development

For model development, we used data from the China Kadoorie Biobank (CKB) cohort. Briefly, the CKB is a prospective cohort study with 512,714 adults (aged 30–80 years) recruited during 2004–2008 from ten areas (5 urban and 5 rural) across China; these areas were selected from China's nationally representative Disease Surveillance Points to maximize geographical and social diversity [27]. Incident lung cancer events (ICD-10 code C33–34) were ascertained through linkage with the mortality and disease registries and national health insurance claim database, supplemented with local residential records and annual active confirmation. Details on the CKB cohort are described in the [Supplementary Materials](#).

In this study, we excluded lung cancer cases at baseline ($n = 130$) and individuals with missing predictors ($n = 16,344$), leaving 496,241 individuals eligible for analysis. The workflow chart for the study design is illustrated in Figure 1A.

2.1.2 | Data for external validation

For external validation, we tested the performance of the model in an independent prospective cohort from Changzhou, Jiangsu province, China. There were 20,803 participants over the age of 35 enrolled between April, 2004, and August, 2005. After excluding 9 baseline lung cancer cases and 751 individuals with missing predictors, a total of 20,043 participants were included for model validation (Figure 1B). Details on the Changzhou cohort are provided in the [Supplementary Materials](#).

Lung cancer incidence (ICD-10: C33–34) was identified via the disease and mortality registries, as well as follow-up questionnaires in 2008–2009, 2012–2013, and 2018–2019. Furthermore, the suspected cases of nonfatal cancer were identified by local medical records or doctor consultation.

2.2 | Candidate predictors

Candidate predictors were selected based on previous lung cancer prediction models [14, 28] and availability within the data. Besides, the predictors should be easily ascertained by community health service staff during a standard consultation. Overall, there were 4 categories of predictors, including demographics, lifestyles, health conditions, and family history. For demographics, age was divided into 8 groups: 1 group aged 30–39 years and 7 groups aged 40–80 years with a 5-year interval. Considering the difference in lung cancer prevalence between urban and rural areas of China, residential location (urban or rural) was also investigated in this study. Education was categorized into college or above, high, and middle school or below based on the highest degree obtained. Body mass index (BMI) was calculated by dividing the weight (kg) by the square of height (m) and classified into underweight ($< 18.5 \text{ kg/m}^2$), normal weight ($18.5\text{--}23.9 \text{ kg/m}^2$), and overweight/obesity ($\geq 24 \text{ kg/m}^2$) groups.

In terms of lifestyles, smoking status was classified as never and ever smokers (which included former and current smokers). Never smokers refer to individuals who smoked fewer than 100 cigarettes during their lifetime. Data on the smoking years for former and current smokers were also calculated. Cigarettes per day = cigarettes + $2 \times$ cigar + $5/3 \times$ (hand-rolled cigarettes + pipes or water pipe). In addition, people smoking similar numbers of cigarettes may have different nicotine intake levels, depending on depth and volume of inhalation [29]. A previous study showed that inhalation of cigarette smoke is a risk factor for lung cancer independent of pack-years [30]. Therefore, smoking inhalation to the lung was considered one of the candidate predictors. A high level of physical activity referred to the sex-specific upper quarter of total physical activity level (metabolic equivalent of task [MET]-hours/day) in the CKB cohort. Since information on calculating MET hours was not available in the Changzhou cohort, frequent exercise was used as a surrogate for a high level of physical activity. Frequent exercise referred to exercising for a minimum of 30 minutes at least 3 times a week.

For health conditions, history of emphysema and/or bronchitis, history of asthma, and frequent cough were considered candidate predictors. Since these lung diseases have low awareness and diagnosis rates but similar clinical manifestations (such as cough) [31], frequent cough might provide additional risk-discriminative information for lung cancer risk. In this study, individuals who reported frequent coughing during the day or at night (lasting 3 months or more) in the past 12 months were classified as having a frequent cough.

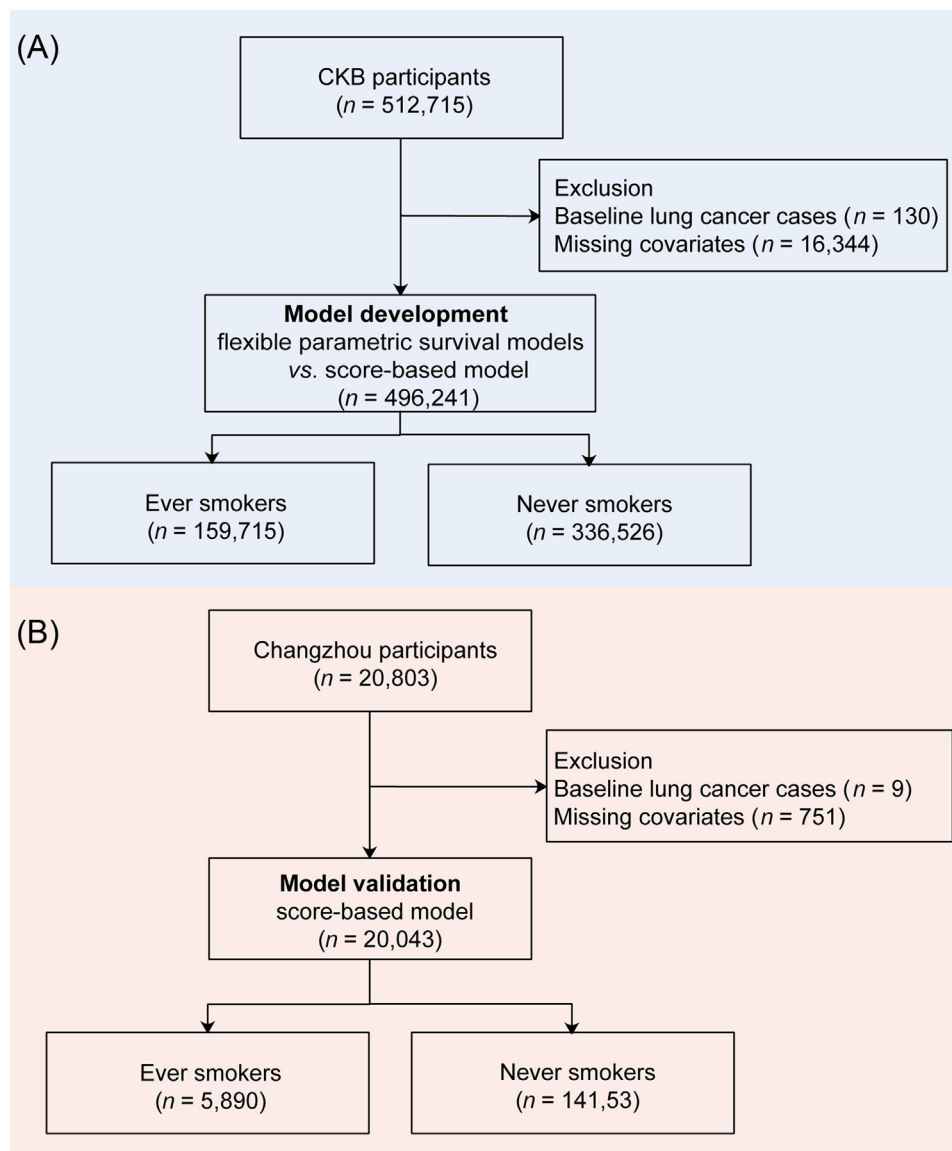


FIGURE 1 Study design and eligible participants' selection procedure. (A) Eligible participants' selection procedure in the CKB cohort. (B) Eligible participants' selection procedure in the Changzhou cohort. Abbreviation: CKB, China Kadoorie Biobank.

Because the family history of lung cancer was not surveyed in the CKB cohort, a family history of cancer was used as a surrogate. First-degree relatives include parents and siblings. Based on a previous study [12], we considered the variable of people with two or more first-degree relatives with cancer as a potential predictor.

Most predictors were assessed in both the CKB and Changzhou cohorts, except for frequent cough, history of emphysema and/or chronic bronchitis, and smoke inhalation to the lungs. Therefore, these missing predictors in the Changzhou cohort were imputed using the corresponding average point in the CKB cohort.

2.3 | Statistical analysis

Considering the significant difference in lung cancer incidences between ever smokers and never smokers, we fitted models for ever and never smokers respectively. We used a Cox regression model to assess the association of candidate predictors with lung cancer risk, giving hazard ratio (HR) and 95% confidence interval (CI). Variables with $P < 0.05$ in the univariable Cox regression were kept for further model development. The remaining variables were further selected by stepwise backward regression ($P < 0.05$). To ensure that no omitted predictor statistically significantly

improved the fit of the model, variables excluded in the first step were re-entered in the multivariable model [32, 33].

Subsequently, we used restricted cubic spline analysis to assess the nonlinear associations between the predictors and lung cancer risk. According to the results of restricted cubic spline analysis, we used flexible parametric survival models on the cumulative hazard scale to estimate HR, while considering the nonlinear associations [28]. Additionally, the lung cancer risk score (LCRS) was calculated based on the Cox regression coefficient. The risk scores of each predictor were calculated by dividing the minimum β -coefficient from the Cox regression model. The total risk score of each participant was calculated by summing the scores of each risk factor [33].

The discriminative ability of the model was assessed by the area under the receiver operating curve (AUC) and the model calibration was demonstrated by plotting observed probability (the Kaplan-Meier estimates) against the mean predicted probability by a tenth of the 6-year predicted absolute risk. Similar methods were used to assess the 3-, 5-, and 10-year lung cancer risk of discrimination and calibration. In addition, we conducted sensitivity analyses by reincluding participants with missing covariates on the basis of multiple imputed data, conducting a simple model based on the most important predictors (Supplementary Materials), and rebuilding a competing risk model by considering death as a competing event. Considering the sex differences in lung cancer risk, gender-stratified models were also examined. Furthermore, the models were verified through 10-fold cross-validation and external validation.

To calculate absolute risk of an individual developing lung cancer over 3, 5, 6, and 10 years, the baseline hazard function [$S_0(t)$] of each time was derived from the Cox regression model [11]. LCRS was used to calculate the probability (\hat{P}) of lung cancer during the next 3, 5, 6, and 10 years based on the following formula (the probability estimated by this score-based prediction model was almost the same as the risk calculated by the coefficient-based full model, as described in Supplementary Materials):

$$\hat{P} = 1 - S_0(t)^{\exp[\beta * (LCRS - \overline{LCRS})]}$$

Where $S_0(t)$ is the baseline survival, $S_0(t = 3) = 0.997,500,5$, $S_0(t = 5) = 0.995,303,7$, $S_0(t = 6) = 0.993,968,4$, $S_0(t = 10) = 0.987,733,3$ for ever smokers; and $S_0(t = 3) = 0.998,940,7$, $S_0(t = 5) = 0.998,059,5$, $S_0(t = 6) = 0.997,527,0$, $S_0(t = 10) = 0.994,892,9$ for never smokers.

Where β is the estimated regression coefficient of LCRS from a Cox regression model, $\beta = 0.022$ for ever smokers; and $\beta = 0.122$ for never smokers.

Where \overline{LCRS} is the corresponding mean for LCRS, $\overline{LCRS} = 157.3$ for ever smokers; and $\overline{LCRS} = 14.7$ for never smokers.

Furthermore, we used X-tile software to determine the optimal cutoff points for separating low-risk from high-risk groups. This method provides a comprehensive approach to determine the optimal threshold based on the time-dependence of the outcomes to distinguish the effects at various time points [34]. We compared the effectiveness of the cutoff points with USPSTF and Chinese screening criteria using sensitivity, specificity, Youden's index (which measures the accuracy of the prediction model), and the number needed to undergo low-dose computed tomography (NNS) to confirm one case in the next 6 years.

To compare the LCRS's performance with that of the PLCO₂₀₁₄ model, we presented the model's performance and the absolute risk of lung cancer for the 6-year period in the primary results, and reported these for the 3-, 5-, and 10-year in the supplementary materials. All two-sided P values were < 0.05 . All statistical analyses were conducted using R version 3.6.3 (R Core Team, Vienna, Austria) and X-tile 3.6.1 (Yale University School of Medicine, New Haven, Connecticut, USA).

3 | RESULTS

3.1 | Demographic characteristics

In the CKB cohort with a median follow-up of 10.1 years (interquartile range: 9.2-11.1 years), 3,133 lung cancer cases were identified among 159,715 ever smokers, and 2,428 cases were identified among 336,526 never smokers. In the Changzhou cohort, during a median follow-up of 13.6 years (interquartile range: 13.5-14.4 years), 104 and 99 new lung cancer cases were identified in 5,890 ever smokers and 14,153 never smokers, respectively. Table 1 demonstrates the demographic characteristics of participants in the development and validation cohorts.

3.2 | Predictor selection

In the CKB cohort, Table 2 shows that in univariate Cox regression, most candidate predictors were significantly associated with the risk of lung cancer. After predictor selection, there were 13 and 9 predictors ($P < 0.001$) in the ever smokers' and never smokers' models, respectively (Table 3). Of these predictors, age, cigarettes per day, smoking years, and quit years were nonlinearly associated with lung cancer risk, whereas height and BMI were linearly associated with lung cancer risk (Figure 2). The curve of the lung cancer incidence increased rapidly above 20

TABLE 1 Distribution of participants by select variables in the CKB and Changzhou cohorts.

Variable	CKB development cohort		Changzhou validation cohort	
	Ever smoker	Never smoker	Ever smoker	Never smoker
Total	159,715	336,526	5,890	14,153
Lung cancer cases	3,133	2,428	104	99
Age at baseline (years, mean ± SD)	53.0 ± 10.7	51.2 ± 10.5	50.2 ± 12.7	49.9 ± 14.6
Age group (years, n [%])				
30-39	20,289 (12.7)	56,331 (16.7)	1,335 (22.7)	3,925 (27.7)
40-44	24,076 (15.1)	59,017 (17.5)	762 (12.9)	1,700 (12.0)
45-49	21,541 (13.5)	45,842 (13.6)	842 (14.3)	1,587 (11.2)
50-54	28,107 (17.6)	57,305 (17.0)	937 (15.9)	1,842 (13.0)
55-59	22,371 (14.0)	45,064 (13.4)	789 (13.4)	1,679 (11.9)
60-64	16,818 (10.5)	30,288 (9.0)	464 (7.9)	1,130 (8.0)
65-69	14,743 (9.2)	24,009 (7.1)	355 (6.0)	847 (6.0)
70-80	11,770 (7.4)	18,670 (5.5)	406 (6.9)	1,443 (10.2)
Residential area (n [%])				
Rural	93,985 (58.8)	182,939 (54.4)	5,890 (100.0)	14,153 (100.0)
Urban	65,730 (41.2)	153,587 (45.6)	0 (0)	0 (0)
Education (n [%])				
College or above	8,711 (5.5)	20,380 (6.1)	55 (0.9)	186 (1.3)
High school	25,080 (15.7)	50,498 (15.0)	647 (11.0)	1,375 (9.7)
Middle school or below	125,924 (78.8)	265,648 (78.9)	5,188 (88.1)	12,592 (89.0)
Height (cm, mean ± SD)	164.5 ± 7.2	156.0 ± 7.2	167.0 ± 6.3	157.3 ± 7.4
BMI (kg/m², mean ±SD)	23.3 ± 3.3	23.8 ± 3.4	23.0 ± 3.1	23.3 ± 3.5
BMI group (n [%])				
≥ 24 (overweight/obesity)	63,788 (39.9)	154,085 (45.8)	2,108 (35.8)	5,576 (39.4)
18.5-23.9 (normal weight)	87,998 (55.1)	169,114 (50.3)	3,425 (58.1)	7,639 (54.0)
< 18.5 (underweight)	7,929 (5.0)	13,327 (4.0)	357 (6.1)	938 (6.6)
Physical activity^a (n [%])				
High level/frequent	40,537 (25.4)	83,488 (24.8)	767 (13.0)	1,451 (10.3)
Low level/occasional	119,178 (74.6)	253,038 (75.2)	5,123 (87.0)	12,702 (89.7)
Frequent cough (n [%])				
No	139,492 (87.3)	316,132 (93.9)	N/A	N/A
Yes	20,223 (12.7)	20,394 (6.1)	N/A	N/A
History of emphysema and/or bronchitis (n [%])				
No	154,568 (96.8)	328,996 (97.8)	N/A	N/A
Yes	5,147 (3.2)	7,530 (2.2)	N/A	N/A
Two or more first degree relatives with cancer (n [%])				
No	155,989 (97.7)	329,127 (97.8)	5,705 (96.9)	13,773 (97.3)
Yes	3,726 (2.3)	7,399 (2.2)	185 (3.1)	380 (2.7)
Previous cancer diagnosis (n [%])				
No	159,033 (99.6)	334,843 (99.5)	5,846 (99.3)	14,021 (99.1)
Yes	682 (0.4)	1,683 (0.5)	44 (0.7)	132 (0.9)
Cigarettes per day (cigarettes/day, mean ± SD)	17.8 ± 10.8	N/A	18.7 ± 9.9	N/A

(Continues)

TABLE 1 (Continued)

Variable	CKB development cohort		Changzhou validation cohort	
	Ever smoker	Never smoker	Ever smoker	Never smoker
Cigarettes per day group (n [%])				
< 10	32,217 (20.2)	N/A	670 (11.4)	N/A
10-14	27,354 (17.1)	N/A	1,078 (18.3)	N/A
15-19	14,945 (9.4)	N/A	432 (7.3)	N/A
20-24	56,618 (35.4)	N/A	2,729 (46.3)	N/A
25-29	4,626 (2.9)	N/A	139 (2.4)	N/A
30-34	11,243 (7.0)	N/A	356 (6.0)	N/A
≥ 35	12,712 (8.0)	N/A	486 (8.3)	N/A
Smoking years (years, mean ± SD)	28.5 ± 11.5	N/A	25.4 ± 11.7	N/A
Smoking years group (n [%])				
< 10	7,667 (4.8)	N/A	462 (7.8)	N/A
10-19	29,655 (18.6)	N/A	1,296 (22.0)	N/A
20-29	52,729 (33.0)	N/A	1,981 (33.6)	N/A
30-39	43,407 (27.2)	N/A	1,428 (24.2)	N/A
40-49	19,951 (12.5)	N/A	509 (8.6)	N/A
≥ 50	6,306 (3.9)	N/A	214 (3.6)	N/A
Smoke inhalation to the lungs (n [%])				
No	83,502 (52.3)	N/A	N/A	N/A
Yes	76,213 (47.7)	N/A	N/A	N/A
Quit smoking (n [%])				
No	130,544 (81.7)	N/A	5,211 (88.5)	N/A
Yes	29,171 (18.3)	N/A	679 (11.5)	N/A
Quit years (years, mean ± SD)				
	1.8 ± 5.2	N/A	0.9 ± 3.6	N/A
Quit years group (n [%])				
> 5	17,465 (10.9)	N/A	319 (5.4)	N/A
≤ 5	142,250 (89.1)	N/A	5,571 (94.6)	N/A

^aSince information on calculating MET hours was not available in the Changzhou cohort, frequent exercise was used as a surrogate for a high level of physical activity.

Data are presented as the mean ± SD for continuous variables and *n* (%) for categorical variables.

Abbreviations: BMI, body-mass index; CKB, China Kadoorie Biobank; CI, confidence interval; HR, hazard ratio; MET, metabolic equivalent of task; N/A, not applicable; SD, standard deviation.

cigarettes per day and then was relatively flat above approximately 30 cigarettes per day ($P_{\text{nonlinear}} < 0.001$; Figure 2G). Above 30 smoking years, lung cancer risk increased with smoking years and then increased slightly above approximately 40 smoking years ($P_{\text{nonlinear}} = 0.004$; Figure 2H). As shown in Figure 2I, lung cancer risk decreased rapidly within the first 5 quit years and then started to decrease at a slower rate ($P_{\text{nonlinear}} = 0.001$).

3.3 | Development of the LCRS in the CKB cohort

Compared with a low LCRS, a higher LCRS was associated with an increased risk of lung cancer (Figure 3A-B). The

incidence of lung cancer increased with LCRS ($P_{\text{overall}} < 0.001$; Figure 4A-B, Supplementary Table S1). The LCRS showed good discrimination for the 6-year risk of lung cancer, with an AUC of 0.778 for ever smokers and an AUC of 0.733 for never smokers (Figure 3C-D). The LCRS-predicted incidence of lung cancer at 6 years was highly consistent with the observed predicted probability in the calibration curves of never and ever smokers (Figure 3E-F). Compared to the flexible parametric survival models considering the nonlinearity, our LCRS showed similarly excellent discrimination (AUC: 0.778 vs. 0.779, $P = 0.919$ for ever smokers; AUC: 0.733 vs. 0.733, $P = 0.828$ for never smokers; Supplementary Table S2, Supplementary Figure S1). Considering that LCRS was easily applied to lung cancer screening, further

TABLE 2 Univariate Cox regression analysis in the CKB cohort.

Variable	Ever smoker		Never smoker	
	HR (95% CI)	P value	HR (95% CI)	P value
Age group (years)				
30-39	Reference	N/A	Reference	N/A
40-44	1.94 (1.40-2.68)	< 0.001	2.22 (1.70-2.89)	< 0.001
45-49	3.43 (2.53-4.65)	< 0.001	3.27 (2.52-4.24)	< 0.001
50-54	6.54 (4.92-8.69)	< 0.001	4.96 (3.90-6.32)	< 0.001
55-59	9.69 (7.30-12.85)	< 0.001	6.72 (5.28-8.55)	< 0.001
60-64	13.72 (10.34-18.19)	< 0.001	10.92 (8.59-13.88)	< 0.001
65-69	20.84 (15.75-27.56)	< 0.001	13.80 (10.85-17.56)	< 0.001
70-80	25.36 (19.14-33.6)	< 0.001	14.69 (11.48-18.80)	< 0.001
Sex				
Female	Reference	N/A	Reference	N/A
Male	0.65 (0.57-0.74)	< 0.001	1.26 (1.14-1.40)	< 0.001
Residential area				
Rural	Reference	N/A	Reference	N/A
Urban	1.37 (1.28-1.47)	< 0.001	1.42 (1.31-1.54)	< 0.001
Education				
College or above	Reference	N/A	Reference	N/A
High school	1.10 (0.89-1.36)	0.395	0.92 (0.75-1.13)	0.403
Middle school or below	1.71 (1.41-2.07)	< 0.001	1.17 (0.98-1.39)	0.089
Height (cm)^a				
< 160 (< 150)	Reference	N/A	Reference	N/A
160-164 (150-154)	0.82 (0.74-0.90)	< 0.001	0.92 (0.82-1.03)	0.129
165-169 (155-159)	0.76 (0.69-0.84)	< 0.001	0.81 (0.72-0.91)	< 0.001
≥ 170 (≥ 160)	0.60 (0.54-0.67)	< 0.001	0.85 (0.76-0.96)	0.008
BMI (kg/m²)				
≥ 24 (overweight/obesity)	Reference	N/A	Reference	N/A
18.5-23.9 (normal weight)	1.27 (1.18-1.38)	< 0.001	1.02 (0.94-1.11)	0.562
< 18.5 (underweight)	2.51 (2.20-2.86)	< 0.001	1.57 (1.32-1.88)	< 0.001
Drinking				
No	Reference	N/A	Reference	N/A
Yes	0.90 (0.84-0.97)	0.004	1.01 (0.87-1.18)	0.860
Intake of fresh vegetables and fruits				
Frequent	Reference	N/A	Reference	N/A
Occasional	0.94 (0.86-1.02)	0.138	0.89 (0.82-0.97)	0.005
Physical activity				
High level/frequent	Reference	N/A	Reference	N/A
Low level/occasional	1.88 (1.71-2.07)	< 0.001	1.96 (1.75-2.19)	< 0.001
Often cooking				
No	Reference	N/A	Reference	N/A
Yes	1.31 (1.21-1.41)	< 0.001	1.05 (0.95-1.15)	0.352
Frequent cough				
No	Reference	N/A	Reference	N/A
Yes	1.53 (1.40-1.68)	< 0.001	1.48 (1.29-1.71)	< 0.001

(Continues)

TABLE 2 (Continued)

Variable	Ever smoker		Never smoker	
	HR (95% CI)	P value	HR (95% CI)	P value
History of emphysema and/or bronchitis				
No	Reference	N/A	Reference	N/A
Yes	2.64 (2.31-3.02)	< 0.001	2.23 (1.84-2.69)	< 0.001
History of asthma				
No	Reference	N/A	Reference	N/A
Yes	1.77 (1.22-2.57)	0.003	1.39 (0.88-2.21)	0.161
Previous cancer diagnosis				
No	Reference	N/A	Reference	N/A
Yes	3.01 (2.11-4.29)	< 0.001	3.18 (2.27-4.47)	< 0.001
Two or more first-degree relatives with cancer				
No	Reference	N/A	Reference	N/A
Yes	1.65 (1.38-1.99)	< 0.001	1.69 (1.37-2.09)	< 0.001
Passive smoking for more than 20 years				
No	N/A	N/A	Reference	N/A
Yes	N/A	N/A	0.81 (0.75-0.88)	< 0.001
Cigarettes per day				
< 10	Reference	N/A	N/A	N/A
10-14	1.02 (0.90-1.16)	0.762	N/A	N/A
15-19	1.18 (1.02-1.36)	0.028	N/A	N/A
20-24	1.21 (1.09-1.34)	< 0.001	N/A	N/A
25-29	1.65 (1.36-2.01)	< 0.001	N/A	N/A
30-34	1.48 (1.28-1.72)	< 0.001	N/A	N/A
≥ 35	1.51 (1.31-1.74)	< 0.001	N/A	N/A
Smoking years				
< 10	Reference	N/A	N/A	N/A
10-19	1.13 (0.81-1.59)	0.474	N/A	N/A
20-29	1.98 (1.45-2.73)	< 0.001	N/A	N/A
30-39	4.77 (3.49-6.52)	< 0.001	N/A	N/A
40-49	9.81 (7.18-13.42)	< 0.001	N/A	N/A
≥ 50	14.62 (10.60-20.18)	< 0.001	N/A	N/A
Smoke inhalation to lungs				
No	Reference	N/A	N/A	N/A
Yes	1.21 (1.13-1.29)	< 0.001	N/A	N/A
Quit years				
> 5	Reference	N/A	N/A	N/A
≤ 5	0.94 (0.84-1.05)	0.295	N/A	N/A

^aFor ever smokers, height was classified into < 160, 160-164, 165-169, and ≥ 170 cm groups; for never smokers, height was divided into < 150, 150-154, 155-159, and ≥ 160 cm groups.

Abbreviations: BMI, body-mass index; CI, confidence interval; CKB, China Kadoorie Biobank; HR, hazard ratio; N/A, not applicable.

analysis was performed based on this method in this study.

To verify the stability of the models, internal 10-fold cross validation was used and exhibited good discriminating ability with average AUCs of 0.779 and 0.732 for

the ever smokers' and never smokers' models, respectively (Supplementary Table S3). In addition, similar discrimination was also observed in the 3-, 5-, and 10-year receiver operating characteristic (ROC) curves (Supplementary Figures S2-S4), as well as in a series of sensitivity

TABLE 3 Predictors for the lung cancer risk model in the CKB cohort and corresponding risk points.

Variable	Ever smoker				Never smoker			
	Regression coefficient	HR (95% CI)	P value	Points assigned	Regression coefficient	HR (95% CI)	P value	Points assigned
Age group (years)								
30-39	N/A	Reference	N/A	0	N/A	Reference	N/A	0
40-44	0.559	1.75 (1.25-2.44)	0.001	25.4	0.802	2.23 (1.71-2.91)	< 0.001	6.6
45-49	1.023	2.78 (2.02-3.83)	< 0.001	46.5	1.184	3.27 (2.52-4.24)	< 0.001	9.7
50-54	1.565	4.78 (3.53-6.49)	< 0.001	71.1	1.596	4.93 (3.87-6.29)	< 0.001	13.1
55-59	1.906	6.72 (4.95-9.14)	< 0.001	86.6	1.890	6.62 (5.19-8.44)	< 0.001	15.5
60-64	2.182	8.86 (6.49-12.10)	< 0.001	99.2	2.364	10.63 (8.34-13.55)	< 0.001	19.4
65-69	2.525	12.49 (9.13-17.11)	< 0.001	114.8	2.570	13.06 (10.23-16.69)	< 0.001	21.1
70-80	2.659	14.28 (10.34-19.72)	< 0.001	120.9	2.615	13.67 (10.63-17.58)	< 0.001	21.4
Residential area								
Rural	N/A	Reference	N/A	0	N/A	Reference	N/A	0
Urban	0.417	1.52 (1.41-1.64)	< 0.001	19.0	0.174	1.19 (1.09-1.29)	< 0.001	1.4
Education								
College or above	N/A	Reference	N/A	0	N/A	N/A	N/A	N/A
High school	0.393	1.48 (1.19-1.84)	< 0.001	17.9	N/A	N/A	N/A	N/A
Middle school or below	0.455	1.58 (1.30-1.92)	< 0.001	20.7	N/A	N/A	N/A	N/A
Height (cm)^a								
< 160 (< 150)	N/A	Reference	N/A	0	N/A	Reference	N/A	0
160-164 (150-154)	0.022	1.02 (0.93-1.12)	0.646	1.0	0.128	1.14 (1.01-1.27)	0.028	1.0
165-169 (155-159)	0.128	1.14 (1.03-1.25)	< 0.001	5.8	0.142	1.15 (1.02-1.30)	0.020	1.2
≥ 170 (≥ 160)	0.135	1.14 (1.02-1.28)	0.019	6.1	0.229	1.26 (1.12-1.42)	< 0.001	1.9
BMI (kg/m²)								
≥ 24 (overweight/obesity)	N/A	Reference	N/A	0	N/A	Reference	N/A	0
18.5-23.9 (normal weight)	0.201	1.22 (1.13-1.32)	< 0.001	9.1	0.159	1.17 (1.08-1.27)	< 0.001	1.3
< 18.5 (underweight)	0.552	1.74 (1.52-1.99)	< 0.001	25.1	0.371	1.45 (1.21-1.73)	< 0.001	3.0
Physical activity								
High level/frequent	N/A	N/A	N/A	N/A	N/A	Reference	N/A	0
Low level/occasional	N/A	N/A	N/A	N/A	0.122	1.13 (1.00-1.27)	0.044	1.0
Frequent cough								
No	N/A	Reference	N/A	0	N/A	Reference	N/A	0
Yes	0.282	1.33 (1.21-1.46)	< 0.001	12.8	0.186	1.20 (1.04-1.39)	0.012	1.5
History of emphysema and/or chronic bronchitis								
No	N/A	Reference	N/A	0	N/A	Reference	N/A	0
Yes	0.332	1.39 (1.21-1.60)	< 0.001	15.1	0.395	1.48 (1.22-1.81)	< 0.001	3.2
Previous cancer diagnosis								
No	N/A	Reference	N/A	0	N/A	Reference	N/A	0
Yes	0.589	1.80 (1.26-2.57)	< 0.001	26.8	0.792	2.21 (1.57-3.10)	< 0.001	6.5
Two or more first degree relatives with cancer								
No	N/A	Reference	N/A	0	N/A	Reference	N/A	0
Yes	0.300	1.35 (1.12-1.62)	< 0.001	13.6	0.277	1.32 (1.07-1.63)	0.011	2.3

(Continues)

TABLE 3 (Continued)

Variable	Ever smoker				Never smoker			
	Regression coefficient	HR (95% CI)	P value	Points assigned	Regression coefficient	HR (95% CI)	P value	Points assigned
Cigarettes per day								
< 10	N/A	Reference	N/A	0	N/A	N/A	N/A	N/A
10-14	0.122	1.13 (1.00-1.28)	0.057	5.5	N/A	N/A	N/A	N/A
15-19	0.232	1.26 (1.09-1.46)	0.002	10.5	N/A	N/A	N/A	N/A
20-24	0.402	1.49 (1.34-1.66)	< 0.001	18.3	N/A	N/A	N/A	N/A
25-29	0.546	1.73 (1.42-2.10)	< 0.001	24.8	N/A	N/A	N/A	N/A
30-34	0.582	1.79 (1.54-2.07)	< 0.001	26.5	N/A	N/A	N/A	N/A
≥ 35	0.636	1.89 (1.64-2.18)	< 0.001	28.9	N/A	N/A	N/A	N/A
Smoking years (years)								
< 10	N/A	Reference	N/A	0	N/A	N/A	N/A	N/A
10-19	0.345	1.41 (1.00-1.99)	0.048	15.7	N/A	N/A	N/A	N/A
20-29	0.495	1.64 (1.19-2.26)	0.003	22.5	N/A	N/A	N/A	N/A
30-39	0.773	2.17 (1.57-2.98)	< 0.001	35.1	N/A	N/A	N/A	N/A
40-49	1.046	2.85 (2.05-3.95)	< 0.001	47.5	N/A	N/A	N/A	N/A
≥ 50	1.199	3.32 (2.35-4.68)	< 0.001	54.5	N/A	N/A	N/A	N/A
Smoke inhalation to the lungs								
No	N/A	Reference	N/A	0	N/A	N/A	N/A	N/A
Yes	0.220	1.25 (1.16-1.34)	< 0.001	10.0	N/A	N/A	N/A	N/A
Quit years (years)								
> 5	N/A	Reference	N/A	0	N/A	N/A	N/A	N/A
≤ 5	0.188	1.21 (1.06-1.37)	0.004	8.5	N/A	N/A	N/A	N/A

^aFor ever smokers, height was classified into < 160, 160-164, 165-169, and ≥ 170 cm groups; for never smokers, height was divided into < 150, 150-154, 155-159, and ≥ 160 cm groups.

Abbreviations: BMI, body-mass index; CI: confidence interval; CKB, China Kadoorie Biobank; HR, hazard ratio; N/A, not applicable.

analyses, including multiple imputation-based analysis (Supplementary Table S4, Supplementary Figure S5), simple model (Supplementary Tables S5-S6, Supplementary Figure S6), competing risk model (Supplementary Table S7, Supplementary Figure S7), and gender stratification (Supplementary Figure S8).

3.4 | Validation of the LCRS

In the Changzhou cohort, the LCRS of lung cancer incidence was significantly higher than that in patients without lung cancer (Figure 3G-H). There was a significant gradient increase in lung cancer risk from deciles 1 to 10 of the LCRS (Figure 4C-D, Supplementary Table S8). The AUCs of LCRS were 0.774 and 0.759 for ever smokers and never smokers, respectively (Figure 3I-J). In addition, the LCRS was largely consistent with the observed risk of lung cancer (Figure 3K-L). Our prediction models had a higher net benefit compared to the situations without the prediction model (Supplementary Figure S9).

3.5 | LCRS categories and absolute risk of incident lung cancer

Among ever smokers, cutoffs (166.2 and 222.4) separated the individuals into low, intermediate, and high-risk groups based on the CKB cohort (Supplementary Figure S10A). Supplementary Figure S11A shows that the absolute risk of lung cancer started to increase when the LCRS was 166.2. Therefore, we used the cutoff value of 166.2 (intermediate-high risk) to identify individuals with high lung cancer risk. The 10-year cumulative incidence of lung cancer was 3.73% and 0.63% in the low and intermediate-high risk groups in the CKB cohort, respectively (Figure 5A). By using the same cutoff, individuals in the Changzhou cohort also showed a differentiated 10-year incidence of lung cancer across the two risk levels (2.57% vs. 0.39%) (Figure 5B).

In never smokers, individuals were divided into low and high-risk groups based on the cutoff value of 21.2 (Supplementary Figure S10B). Among individuals with an LCRS above 21.2, we observed an increased absolute

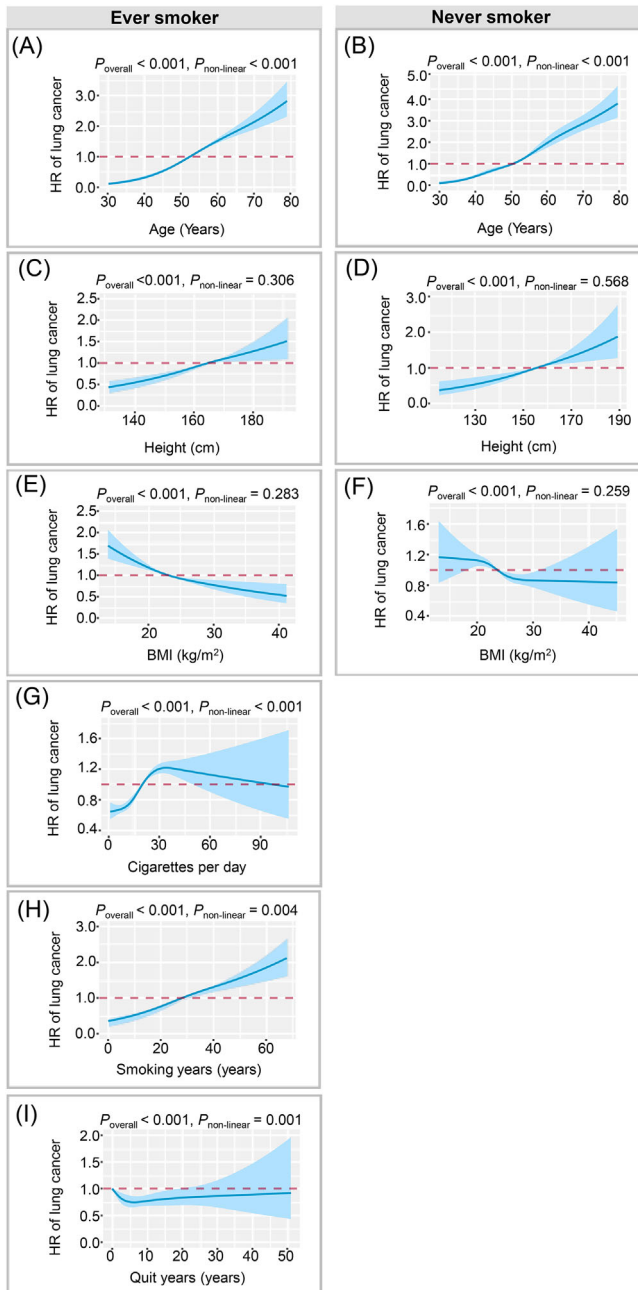


FIGURE 2 Linear and non-linear association between predictors and lung cancer risk. The linear association between age and lung cancer risk using restricted cubic splines in ever smokers (A) and never smokers (B). The linear association between height and lung cancer risk using restricted cubic splines in ever smokers (C) and never smokers (D). The linear association between BMI and lung cancer risk using restricted cubic splines in ever smokers (E) and never smokers (F). The non-linear association between cigarettes per day and lung cancer risk using restricted cubic splines in ever smokers (G). The non-linear association between smoking years and lung cancer risk using restricted cubic splines in ever smokers (H). The non-linear association between quit years and lung cancer risk using restricted cubic splines in ever smokers (I). Abbreviations: BMI, body-mass index; HR, hazard ratio.

risk of lung cancer (Supplementary Figure S11B). In addition, compared with the low-risk group, the high-risk group had a higher 10-year cumulative incidences of lung cancer (1.69% vs. 0.45%; Figure 5C). Similar results were also observed in the Changzhou cohort (1.05% vs. 0.22%; Figure 5D).

Compared with the USPSTF criteria, the LCRS had higher sensitivity (73.95% vs. 64.12%) and Youden's index (37.24% vs. 29.51%) with the same number of ever smokers screened in the CKB cohort (Supplementary Table S9). Moreover, the risk-based fixed Chinese screening criteria sample-size strategy was modeled to have higher sensitivity (69.42% vs. 58.25%) and Youden's index (37.18% vs. 28.00%) compared with Chinese screening criteria (Supplementary Table S10). Consistent results were observed in the Changzhou validation cohort.

The comparisons of LCRS at cutoff points with USPSTF and Chinese screening criteria are shown in Supplementary Table S11. Our LCRS had the highest Youden's index and the lowest NNS, which favored the effectiveness of the cutoff points.

When further compared with the PLCO₂₀₁₄ model in the Changzhou cohort, the LCRS showed higher discrimination (LCRS AUC vs. PLCO₂₀₁₄ AUC = 0.774 vs. 0.748 for ever smokers; LCRS AUC vs. PLCO₂₀₁₄ AUC = 0.759 vs. 0.651; Figure 3I-J, Supplementary Figure S12A-B), along with better calibration (Figure 3K-L, Supplementary Figure S12C-D). Besides, our LCRS showed a higher sensitivity (73.08% vs. 53.85%) and higher specificity (66.33% vs. 56.76%) than PLCO₂₀₁₄ in ever smokers.

3.6 | Web-based tool for lung risk assessment: LCKEY

To facilitate risk evaluation, a website, namely Lung Cancer Risk Evaluation by Yourself (LCKEY), was conducted based on the LCRS. Individuals could easily use LCKEY to calculate the risk of developing lung cancer over the next 3-, 5-, 6-, and 10-year. LCKEY also provided the corresponding recommendations for risk reduction and screening based on their responses. The LCKEY tool is available at <http://ccra.njmu.edu.cn/lckey/web>.

4 | DISCUSSION

In this study, we systematically selected the predictors and explored the linear and nonlinear association between the final inclusive predictors and lung cancer risk. Further, we constructed the LCRS to calculate the absolute risk of lung cancer over 3-, 5-, 6-, and 10-year for never and ever smokers aged 30 to 80 years. Our LCRS was comparable to the

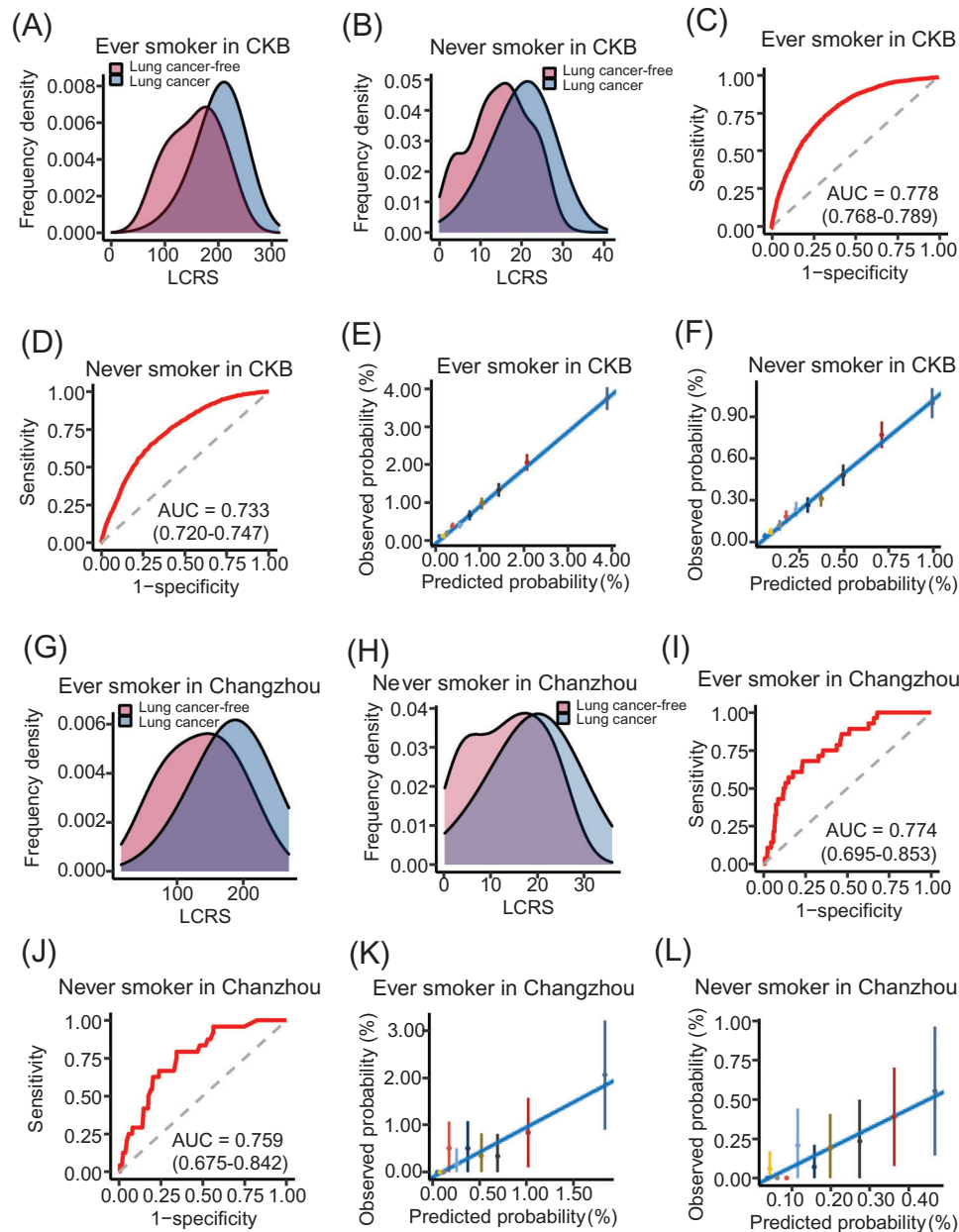


FIGURE 3 Distribution, discrimination, and calibration of the LCRS in the CKB and Changzhou cohorts. Distribution of the LCRS across incident lung cancer cases and lung cancer-free participants during follow-up in the CKB (A and B) and Changzhou cohorts (G and H). Receiver operating characteristic curve at six years in the CKB cohort (C and D) and Changzhou cohort (I and J). The observed 6-year probability of lung cancer with 95% CIs was estimated by the Kaplan-Meier method within deciles of LCRS predicted probability in the CKB (E and F) and Changzhou cohorts (K and L).

Abbreviations: AUC, area under the receiver operating curve; CI, confidence interval; CKB, China Kadoorie Biobank; LCRS, lung cancer risk score.

state-of-the-art method (flexible parametric survival models). The LCRS had good discrimination and calibration in both the development and validation cohorts. Furthermore, we also provided the justified cutoff points for ever (LCRS ≥ 166.2) and never smokers (LCRS ≥ 21.2) to select screening candidates. Finally, a web-based tool was conducted based on our models to facilitate risk assessment as

well as guide referral to screening and motivate behavior change.

Our LCRS was constructed using variables that could be easily obtained by questionnaire. Most of these variables have been reported in the established models [14, 28]. Our model for ever smokers included cigarettes per day, smoking years, quit years, and smoke inhalation

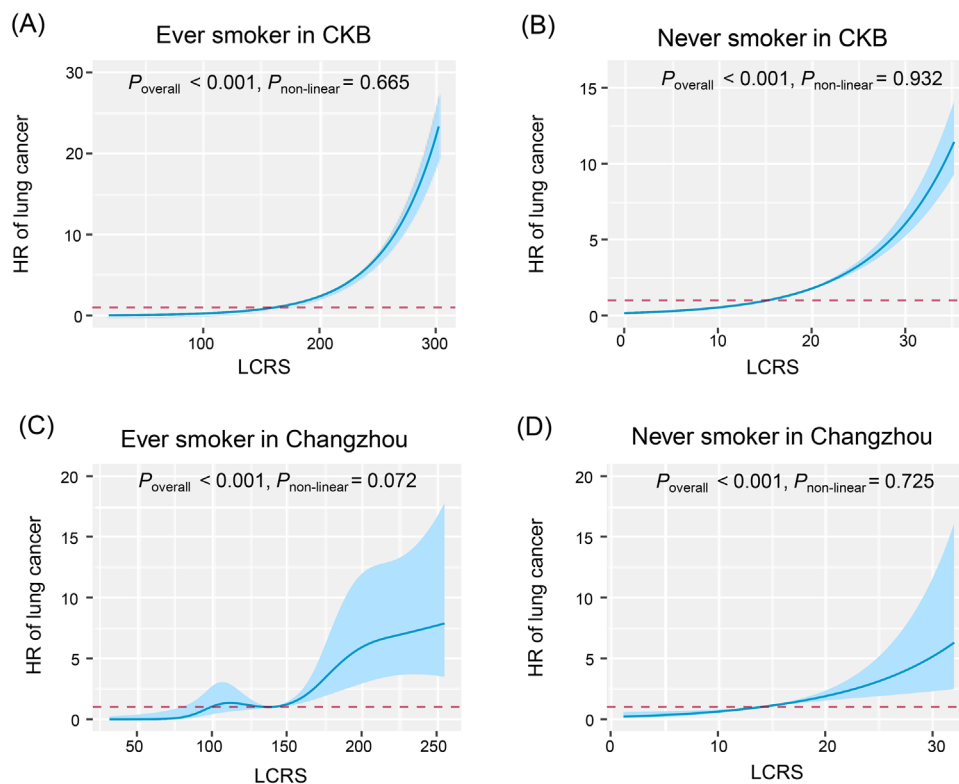


FIGURE 4 The association of the LCRS with incident lung cancer risk in the CKB cohort. The linear association of LCRS and lung cancer risk using restricted cubic splines in ever smokers (A) and never smokers (B) in the CKB cohort. The linear association of LCRS and lung cancer risk using restricted cubic splines in ever smokers (C) and never smokers (D) in the Changzhou cohort. Abbreviations: CKB, China Kadoorie Biobank; HR, hazard ratio; LCRS, lung cancer risk score.

to the lungs. The relative risks between cigarettes per day, smoking years, and lung cancer risk were weaker in our model than in the previous European or North American risk models [22, 32, 35], but were consistent with those in other Chinese models [36, 37]. For cigarettes per day, smoking years, and quit years, the associations with lung cancer risk were nonlinear in our study, which was also observed in European populations [28, 38-40]. We observed a flat lung cancer risk at 30 cigarettes per day, which was consistent with the previous hypothesis proposed by Doll and Peto that cigarettes per day and the risk of lung cancer should be upward curving [41]. A meta-analysis including 12 Chinese studies explored the 10 dose-response relationship models of pack-years with the risk of lung cancer development or death, and found that the best fit model exhibits a “ceiling effect”, with a steep curve at low exposures that smooths out at high exposures [42]. The possible reasons for these flat associations might be that less smoke is inhaled from each cigarette by men with high daily cigarette consumption than by men with lower consumption [43]. Furthermore, when plotted against smoking pack-years, mutations followed the linear increase in cancer risk until approximately 23 pack-years,

after which no further increase in mutation frequency was observed [40]. Moreover, the flat associations may partly be because of exaggeration by smokers who report heavy consumption.

A previous Chinese study, including 1,208 lung cancer cases vs. 1,069 controls, found that a rapidly decreasing odds ratio of lung cancer within the first 5 years of quitting; the odds ratio continued to decrease but at a slower rate in the subsequent years [39]. Similarly, a recent study based on CKB has reported that people who had stopped smoking had only small excess risks for overall mortality including lung cancer (HR = 1.06, 95% CI: 1.01-1.11) and morbidity (HR = 1.05, 95% CI: 1.03-1.08), with the risks approaching those among never-smokers about 5-10 years after quitting [44]. Besides, the aforementioned meta-analysis also showed the risk of lung cancer decreases significantly with quit-years, with the relative risk close to 1 after 7 years of abstinence [42]. Our result was in line with these findings, lung cancer risk reduced rapidly until approximately 5 quit years and then started to decrease at a slower rate. Although the association between quit years and lung cancer risk remains controversial, these findings could be persuasive to promote smoking cessation.

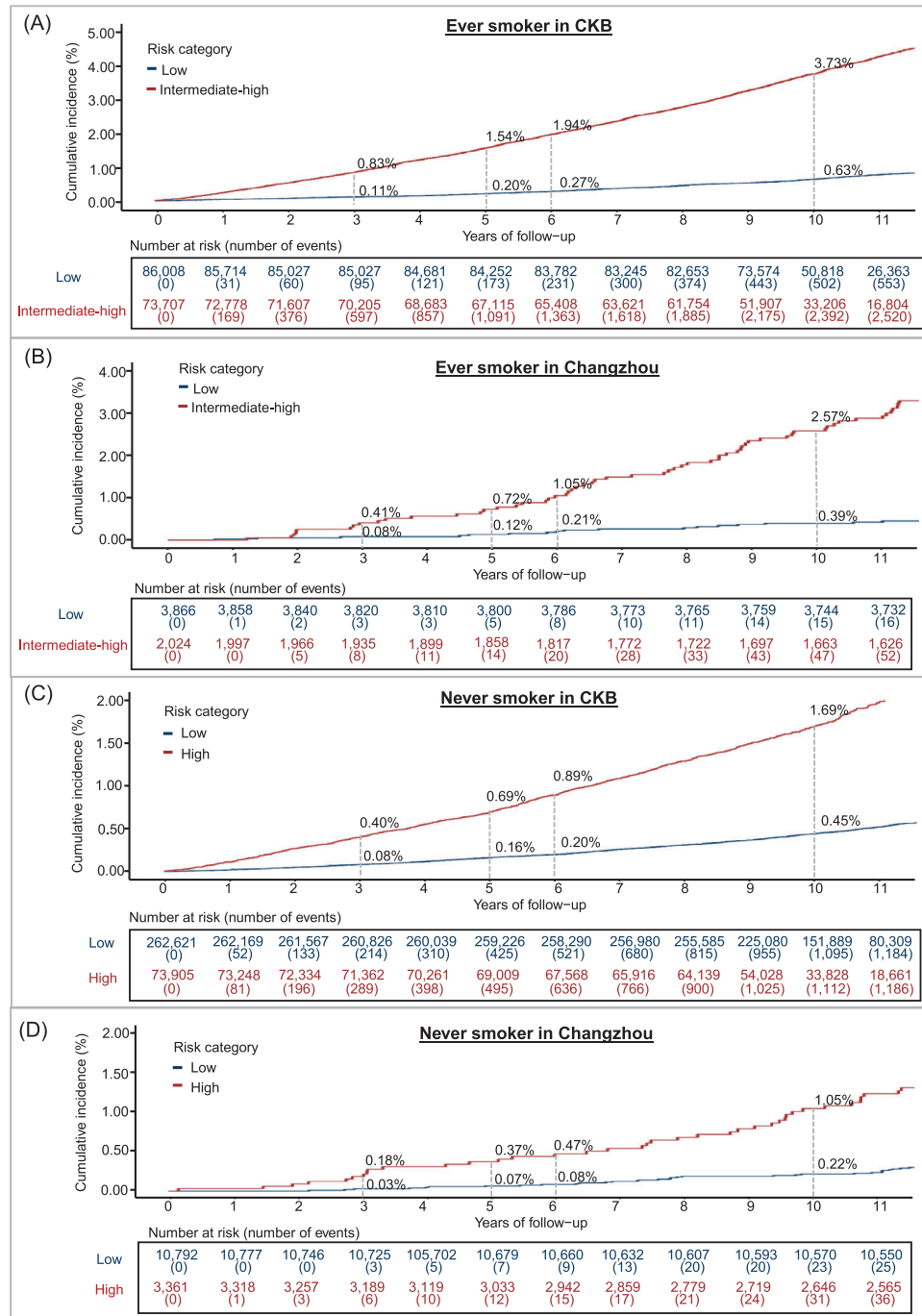


FIGURE 5 Inverted Kaplan-Meier plot of incident lung cancer in the CKB and Changzhou cohorts. Ever smokers were classified into low (LCRS < 166.2) and intermediate-high risk groups (LCRS ≥ 166.2); and never smokers were divided into low (LCRS < 21.2) and high (LCRS ≥ 21.2) risk groups. Abbreviations: CKB, China Kadoorie Biobank; LCRS, lung cancer risk score.

To our knowledge, we first included smoke inhalation to the lungs in the risk prediction model. Researchers found that inhalation of cigarette smoke is a risk factor for lung cancer independent from pack-years [30]. Therefore, the inclusion of smoke inhalation to the lungs could improve the discrimination of the model among ever smokers.

In addition, height and frequent cough were also considered important predictors in our models. A Mendelian randomization analysis showed that height is independently associated with lung cancer risk [45]. As reported in recent studies, subjects with lung diseases were at excess risk of lung cancer [46, 47]. Because these diseases have

low awareness and diagnosis rates but similar clinical manifestations (such as cough) [31], frequent cough during the last 12 months might provide additional information for lung cancer risk estimation.

Due to the difference in the prevalence of risk factors and the effect on lung cancer between Western populations and the Chinese population [17, 18], our study showed that the PLCO₂₀₁₄ could not be directly adapted to the Chinese population. To our knowledge, there were 10 lung cancer risk prediction models derived for the Chinese populations [36, 37, 48-55]. And only two of these was externally validated. Moreover, these two models performed not well in validation cohorts due to the limited number of predictors and short follow-up time [36, 52]. Therefore, our study first conducted a LCRS for a broader range of the Chinese population using 2 prospective cohorts with more than 10 years of follow-up. Our models also showed good performance in the validation cohort with all AUCs > 0.750 for ever smokers and never smokers. According to 2018 national smoking surveys, more than half of men aged ≥ 30 are smokers, and the total number of smokers exceeds 350 million. This might help a broader population, especially those aged 30-50 years, enhance their awareness of smoking cessation since the models inform them of the hazards of smoking, rather than a fuzzy sense of “smoking is harmful to health”.

Our study has several strengths. The predictors were derived from easily available predictors, implying that they could be straightforwardly applied in clinical practice. Furthermore, an easy-to-use online tool, LCKEY, allowed calculation of risks, along with the corresponding suggestions for lifestyle changes and tailored screening. Besides, the morbidity and mortality rates of cancer in the CKB cohort are consistent with those from the Cancer Registration System of China, which indicates good representativeness of the participants in this study [56]. Moreover, the number of lung cancer cases in this study was the largest ($n = 4,395$) compared to previous studies [7, 9, 21, 32, 36, 38].

However, some limitations require consideration. First, risk factors, in particular smoking behavior, might have changed during study follow-up, but such information on variable changes was unavailable for analysis. Second, frequent cough, history of emphysema and/or bronchitis, and smoke inhalation to the lungs were not assessed in the Changzhou cohort. We imputed these missing predictors in the Changzhou cohort with the corresponding average point in the CKB cohort, which possibly reduced the actual discrimination of our LCRS. Our models require further validation by studies with cohorts including all prediction indicators. Third, missing information on the stages of lung cancer may preclude us from assessing the performance of our models in different lung cancer staging.

Fourth, $LCRS \geq 166.2$ for ever smokers caused a nearly 50% false positive rate, and $LCRS \leq 21.2$ for never smokers led to approximately 50% false negative rate. However, our model had higher accuracy (Youden's index) and needed to screen fewer individuals with higher screening efficiency (NNS) compared with the selection criteria. The predictive model needs to be improved in lung cancer screening trials to determine the optimal threshold for better guidance on lung cancer risk stratification. Fifth, we only externally validated our models in a relatively small-scale cohort. Further validations in a large cohort are warranted.

5 | CONCLUSIONS

Using data from 2 prospective Chinese cohorts, we developed and validated the LCRS for ever smokers and never smokers with a wide age range. Moreover, an online risk assessment tool, LCKEY, was constructed and would be easily accessible to the general population without fees. This tool could help reduce lung cancer risk by encouraging current smokers to quit and by identifying high-risk individuals who might benefit from screening.

DECLARATIONS

AUTHORS CONTRIBUTIONS

Hongbing Shen contributed to study concept, study supervision, and critical revision of the manuscript. Liming Li contributed to the study design and supervised the whole project. Zhimin Ma contributed to data analysis and manuscript writing. Jun Lv contributed to the study design and data collection. Meng Zhu contributed to data interpretation of the present analysis and manuscript writing. Hongxia Ma, Guangfu Jin, and Zhibin Hu guided and supervised the study. Canqing Yu, Yu Guo, Zheng Bian, Ling Yang, Yiping Chen, and Zhengming Chen contributed to the study design and sample collection. All of the authors reviewed or revised the manuscript.

ACKNOWLEDGEMENTS

The most important acknowledgement is to the participants in the study and the members of the survey teams in each of the 10 regional centers, as well as to the project development and management teams based at Beijing, Oxford and the 10 regional centers. We thank all the study participants and research staff for their contributions and commitment to the present study. This work was supported by National Natural Science Foundation of China (81820108028, 81922061, 81973123, 82273714, 82192901, 82192904, and 82192900), the Excellent Youth Foundation of Jiangsu Province (BK20220100), Research Unit of Prospective Cohort of Cardiovascular Diseases and Cancer, Chinese Academy of Medical Sciences

(2019RU038), the Science and Technology Service Network Initiative of Chinese Academy of Sciences (No.KFJ-STS-QYZD-2021-08-001), and National Key Research and Development Program of China (2016YFC0900500).

CONFLICT OF INTERESTS STATEMENT

The authors declare that they have no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The CKB study was approved by the ethical committee of the Chinese Center for Disease Control and Prevention (Beijing, China: 005/2004) and the Oxford Tropical Research Ethics Committee, University of Oxford (Oxford, UK: 025-04). The Changzhou cohort have been approved by the Ethical Review Committee of the Nanjing Medical University (Nanjing, China [2003]068). All participants provided the written informed consent.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY STATEMENT

Details of how to access CKB data and details of the data release schedule are available from www.ckbiobank.org/site/Data + Access. Changzhou data can be obtained upon reasonable request to the corresponding author.

ORCID

Hongxia Ma  <https://orcid.org/0000-0002-2462-9693>

Zhibin Hu  <https://orcid.org/0000-0002-8277-5234>

Hongbing Shen  <https://orcid.org/0000-0002-2581-5906>

REFERENCES

- Wang JB, Jiang Y, Wei WQ, Yang GH, Qiao YL, Boffetta P. Estimation of cancer incidence and mortality attributable to smoking in China. *Cancer Causes Control*. 2010;21(6):959–65.
- Qiu H, Cao S, Xu R. Cancer incidence, mortality, and burden in China: a time-trend analysis and comparison with the United States and United Kingdom based on the global epidemiological data released in 2020. *Cancer Commun (Lond)*. 2021;41(10):1037–48.
- de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med*. 2020;382(6):503–13.
- Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395–409.
- Force USPST, Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, et al. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2021;325(10):962–70.
- He J, Li N, Chen WQ, Wu N, Shen HB, Jiang Y, et al. [China guideline for the screening and early detection of lung cancer(2021, Beijing)]. *Zhonghua Zhong Liu Za Zhi*. 2021;43(3):243–68.
- Tammemägi MC, Church TR, Hocking WG, Silvestri GA, Kvale PA, Riley TL, 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 Med*. 2014;11(12):e1001764.
- Lu MT, Raghu VK, Mayrhofer T, Aerts H, Hoffmann U. Deep Learning Using Chest Radiographs to Identify High-Risk Smokers for Lung Cancer Screening Computed Tomography: Development and Validation of a Prediction Model. *Ann Intern Med*. 2020;173(9):704–13.
- Tammemagi MC, Schmidt H, Martel S, McWilliams A, Goffin JR, Johnston MR, et al. Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study. *Lancet Oncol*. 2017;18(11):1523–31.
- Guida F, Sun N, Bantis LE, Muller DC, Li P, Taguchi A, et al. Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins. *JAMA Oncol*. 2018;4(10):e182078.
- Marcus MW, Chen Y, Raji OY, Duffy SW, Field JK. LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence. *Cancer Prev Res*. 2015;8(6):570–5.
- Katki HA, Kovalchik SA, Petito LC, Cheung LC, Jacobs E, Jemal A, et al. Implications of Nine Risk Prediction Models for Selecting Ever-Smokers for Computed Tomography Lung Cancer Screening. *Ann Intern Med*. 2018;169(1):10–9.
- Husing A, Kaaks R. Risk prediction models versus simplified selection criteria to determine eligibility for lung cancer screening: an analysis of German federal-wide survey and incidence data. *Eur J Epidemiol*. 2020;35(10):899–912.
- Tammemagi MC, Church TR, Hocking WG, Silvestri GA, Kvale PA, Riley TL, 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 Med*. 2014;11(12):e1001764.
- Park B, Kim Y, Lee J, Lee N, Jang SH. Risk-based prediction model for selecting eligible population for lung cancer screening among ever smokers in Korea. *Transl Lung Cancer Res*. 2021;10(12):4390–402.
- Charvat H, Sasazuki S, Shimazu T, Budhathoki S, Inoue M, Iwasaki M, et al. Development of a risk prediction model for lung cancer: The Japan Public Health Center-based Prospective Study. *Cancer Sci*. 2018;109(3):854–62.
- Chen ZM, Peto R, Iona A, Guo Y, Chen YP, Bian Z, et al. Emerging tobacco-related cancer risks in China: A nationwide, prospective study of 0.5 million adults. *Cancer*. 2015;121 Suppl 17(Suppl 17):3097–106.
- Agudo A, Bonet C, Travier N, González CA, Vineis P, Bueno-de-Mesquita HB, et al. Impact of cigarette smoking on cancer risk in the European prospective investigation into cancer and nutrition study. *J Clin Oncol*. 2012;30(36):4550–7.
- Gadgeel SM, Ramalingam S, Cummings G, Kraut MJ, Wozniak AJ, Gaspar LE, et al. Lung cancer in patients < 50 years of age: the experience of an academic multidisciplinary program. *Chest*. 1999;115(5):1232–6.
- Ak G, Metintas M, Metintas S, Yildirim H, Erginel S, Alatas F. Lung cancer in individuals less than 50 years of age. *Lung*. 2007;185(5):279–86.

21. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection Criteria for Lung-Cancer Screening. *N Engl J Med*. 2013;368(8):728–36.
22. Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening. *JAMA*. 2016;315(21):2300–11.
23. Gilbert H, Sutton S, Morris R, Petersen I, Galton S, Wu Q, et al. Effectiveness of personalised risk information and taster sessions to increase the uptake of smoking cessation services (Start2quit): a randomised controlled trial. *Lancet*. 2017;389(10071):823–33.
24. Adamson A, Portas L, Accordini S, Marcon A, Jarvis D, Baio G, et al. Communication of personalised disease risk by general practitioners to motivate smoking cessation in England: a cost-effectiveness and research prioritisation study. *Addiction*. 2022;117(5):1438–49.
25. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. *Nat Rev Cancer*. 2007;7(10):778–90.
26. Cho J, Choi SM, Lee J, Lee CH, Lee SM, Kim DW, et al. Proportion and clinical features of never-smokers with non-small cell lung cancer. *Chin J Cancer*. 2017;36(1):20.
27. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol*. 2011;40(6):1652–66.
28. Muller DC, Johansson M, Brennan P. Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study. *J Clin Oncol*. 2017;35(8):861–9.
29. Pennanen M, Broms U, Korhonen T, Haukkala A, Partonen T, Tuulio-Henriksson A, et al. Smoking, nicotine dependence and nicotine intake by socio-economic status and marital status. *Addict Behav*. 2014;39(7):1145–51.
30. Fukumoto K, Ito H, Matsuo K, Tanaka H, Yokoi K, Tajima K, et al. Cigarette smoke inhalation and risk of lung cancer: a case-control study in a large Japanese population. *Eur J Cancer Prev*. 2015;24(3):195–200.
31. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract*. 2011;61(592):e715–23.
32. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer*. 2008;98(2):270–6.
33. Chen W, Li H, Ren J, Zheng R, Shi J, Li J, et al. Selection of high-risk individuals for esophageal cancer screening: A prediction model of esophageal squamous cell carcinoma based on a multicenter screening cohort in rural China. *Int J Cancer*. 2021;148(2):329–39.
34. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bioinformatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res*. 2004;10(21):7252–9.
35. Hoggart C, Brennan P, Tjonneland A, Vogel U, Overvad K, Ostergaard JN, et al. A risk model for lung cancer incidence. *Cancer Prev Res (Phila)*. 2012;5(6):834–46.
36. Wang F, Tan F, Shen S, Wu Z, Cao W, Yu Y, et al. A Risk-Stratified Approach for Never- and Ever-Smokers in Lung Cancer Screening: A Prospective Cohort Study in China. *Am J Respir Crit Care Med*. 2023;207(1):77–88.
37. Guo LW, Lyu ZY, Meng QC, Zheng LY, Chen Q, Liu Y, et al. A risk prediction model for selecting high-risk population for computed tomography lung cancer screening in China. *Lung Cancer*. 2022;163:27–34.
38. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation. *J Natl Cancer Inst*. 2011;103(13):1058–68.
39. Tse LA, Yu IT, Qiu H, Au JS, Wang XR, Tam W, et al. Lung cancer decreased sharply in first 5 years after smoking cessation in Chinese men. *J Thorac Oncol*. 2011;6(10):1670–6.
40. Huang Z, Sun S, Lee M, Maslov AY, Shi M, Waldman S, et al. Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nat Genet*. 2022;54(4):492–8.
41. Doll R, Peto R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. *J Epidemiol Community Health* (1978). 1978;32(4):303–13.
42. Ai F, Zhao J, Yang W, Wan X. Dose-response relationship between active smoking and lung cancer mortality/prevalence in the Chinese population: a meta-analysis. *BMC Public Health*. 2023;23(1):747.
43. Law MR, Morris JK, Watt HC, Wald NJ. The dose-response relationship between cigarette consumption, biochemical markers and risk of lung cancer. *Br J Cancer*. 1997;75(11):1690–3.
44. Chan KH, Wright N, Xiao D, Guo Y, Chen Y, Du H, et al. Tobacco smoking and risks of more than 470 diseases in China: a prospective cohort study. *Lancet Public Health*. 2022;7(12):e1014–e26.
45. Khankari NK, Shu XO, Wen W, Kraft P, Lindström S, Peters U, et al. Association between Adult Height and Risk of Colorectal, Lung, and Prostate Cancer: Results from Meta-analyses of Prospective Studies and Mendelian Randomization Analyses. *PLoS Med*. 2016;13(9):e1002118.
46. 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.
47. Ramanakumar AV, Parent ME, Menzies D, Siemiatycki J. Risk of lung cancer following nonmalignant respiratory conditions: evidence from two case-control studies in Montreal, Canada. *Lung Cancer*. 2006;53(1):5–12.
48. Yeh MC-H, Wang Y-H, Yang H-C, Bai K-J, Wang H-H, Li Y-CJ. Artificial Intelligence-Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach. *J Med Internet Res*. 2021;23(8):e26256.
49. Guo L-W, Lyu Z-Y, Meng Q-C, Zheng L-Y, Chen Q, Liu Y, et al. Construction and Validation of a Lung Cancer Risk Prediction Model for Non-Smokers in China. *Front Oncol*. 2021;11:766939.
50. Warkentin MT, Tammemagi MC, Espin-Garcia O, Budhathoki S, Liu G, Hung RJ. Lung Cancer Absolute Risk Models for Mortality in Asian Population using China Kadoorie Biobank. *J Natl Cancer Inst*. 2022;114(12):1665–73.
51. Wu X, Wen CP, Ye Y, Tsai M, Wen C, Roth JA, et al. Personalized Risk Assessment in Never, Light, and Heavy Smokers in a prospective cohort in Taiwan. *Sci Rep*. 2016;6:36482.
52. Chien LH, Chen CH, Chen TY, Chang GC, Tsai YH, Hsiao CF, et al. Predicting Lung Cancer Occurrence in Never-Smoking Females in Asia: TNSF-SQ, a Prediction Model. *Cancer Epidemiol Biomarkers Prev*. 2020;29(2):452–9.

53. Li Y, Zou Z, Gao Z, Wang Y, Xiao M, Xu C, et al. Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting. *Cancer Med.* 2022;11(23):4469–78.
54. Tse LA, Wang F, Wong MC-S, Au JS-K, Yu IT-S. Risk assessment and prediction for lung cancer among Hong Kong Chinese men. *BMC Cancer.* 2022;22(1):585.
55. Lyu Z, Li N, Chen S, Wang G, Tan F, Feng X, et al. Risk prediction model for lung cancer incorporating metabolic markers: Development and internal validation in a Chinese population. *Cancer Med.* 2020;9(11):3983–94.
56. Pan R, Zhu M, Yu C, Lv J, Guo Y, Bian Z, et al. Cancer incidence and mortality: A cohort study in China, 2008-2013. *Int J Cancer.* 2017;141(7):1315–23.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ma Z, Lv J, Zhu M, Yu C, Ma H, Jin G, et al. Lung cancer risk score for ever and never smokers in China. *Cancer Commun.* 2023;1–19. <https://doi.org/10.1002/cac2.12463>