



Construction and Validation of a Lung Cancer Risk Prediction Model for Non-Smokers in China

Lan-Wei Guo¹, Zhang-Yan Lyu², Qing-Cheng Meng³, Li-Yang Zheng¹, Qiong Chen¹, Yin Liu¹, Hui-Fang Xu¹, Rui-Hua Kang¹, Lu-Yao Zhang¹, Xiao-Qin Cao¹, Shu-Zheng Liu¹, Xi-Bin Sun¹, Jian-Gong Zhang¹ and Shao-Kai Zhang^{1*}

¹ Department of Cancer Epidemiology and Prevention, Henan Engineering Research Center of Cancer Prevention and Control, Henan International Joint Laboratory of Cancer Prevention, The Affiliated Cancer Hospital of Zhengzhou University, Henan Cancer Hospital, Zhengzhou, China, ² Department of Cancer Epidemiology and Biostatistics, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy of Tianjin, Tianjin's Clinical Research Center for Cancer, Key Laboratory of Molecular Cancer Epidemiology of Tianjin, Key Laboratory of Breast Cancer Prevention and Therapy of the Ministry of Education, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, ³ Department of Radiology, The Affiliated Cancer Hospital of Zhengzhou University, Henan Cancer Hospital, Zhengzhou, China

OPEN ACCESS

Edited by:

Yutong He,
Fourth Hospital of Hebei Medical
University, China

Reviewed by:

Lei Yang,
Peking University Cancer Hospital,
China
Xiaolong Tang,
Shandong University, China

*Correspondence:

Shao-Kai Zhang
shaokaizhang@126.com

Specialty section:

This article was submitted to
Thoracic Oncology,
a section of the journal
Frontiers in Oncology

Received: 30 August 2021

Accepted: 13 December 2021

Published: 04 January 2022

Citation:

Guo L-W, Lyu Z-Y, Meng Q-C,
Zheng L-Y, Chen Q, Liu Y, Xu H-F,
Kang R-H, Zhang L-Y, Cao X-Q,
Liu S-Z, Sun X-B, Zhang J-G and
Zhang S-K (2022) Construction and
Validation of a Lung Cancer
Risk Prediction Model for
Non-Smokers in China.
Front. Oncol. 11:766939.
doi: 10.3389/fonc.2021.766939

Background: About 15% of lung cancers in men and 53% in women are not attributable to smoking worldwide. The aim was to develop and validate a simple and non-invasive model which could assess and stratify lung cancer risk in non-smokers in China.

Methods: A large-sample size, population-based study was conducted under the framework of the Cancer Screening Program in Urban China (CanSPUC). Data on the lung cancer screening in Henan province, China, from October 2013 to October 2019 were used and randomly divided into the training and validation sets. Related risk factors were identified through multivariable Cox regression analysis, followed by establishment of risk prediction nomogram. Discrimination [area under the curve (AUC)] and calibration were further performed to assess the validation of risk prediction nomogram in the training set, and then validated by the validation set.

Results: A total of 214,764 eligible subjects were included, with a mean age of 55.19 years. Subjects were randomly divided into the training (107,382) and validation (107,382) sets. Elder age, being male, a low education level, family history of lung cancer, history of tuberculosis, and without a history of hyperlipidemia were the independent risk factors for lung cancer. Using these six variables, we plotted 1-year, 3-year, and 5-year lung cancer risk prediction nomogram. The AUC was 0.753, 0.752, and 0.755 for the 1-, 3- and 5-year lung cancer risk in the training set, respectively. In the validation set, the model showed a moderate predictive discrimination, with the AUC was 0.668, 0.678, and 0.685 for the 1-, 3- and 5-year lung cancer risk.

Conclusions: We developed and validated a simple and non-invasive lung cancer risk model in non-smokers. This model can be applied to identify and triage patients at high risk for developing lung cancers in non-smokers.

Keywords: lung cancer, risk model, forecasting, validation, non-smokers

HIGHLIGHTS

1. Evidence before this study:
 - About 15% of lung cancers in men and 53% in women are not attributable to smoking worldwide.
 - Screening people at high risk for lung cancer by low-dose computed tomography (LDCT) has been approved to effective in reducing the burden of this disease.
 - Developing lung-cancer risk prediction tools for Chinese non-smokers in large-scale population-based lung screening programs is sparse.
2. Added value of this study:
 - Risk factors associated with lung cancer in Chinese non-smokers were identified.
 - The model developed has moderate discriminatory accuracy and goodness-of-fit for both men and women, non-passive smokers and passive smokers.
3. Implications of all the available evidence:
 - This model can be applied to identify and triage patients at high risk for developing lung cancer in non-smokers.
 - The model has potential utility for shared decision making and individualized risk assessment for tailored lung cancer screening in Chinese non-smokers.

INTRODUCTION

Lung cancer is the leading cause of cancer related deaths in both the world and China. The latest data from the International Agency for Research on Cancer (IARC) shows that in 2020, there were about 1.80 million lung cancer deaths worldwide, which China accounts for 39.8% (1). The majority of lung cancer cases in China were found to be clinically advanced, with 64.6% of stage III-IV lung cancers in 2012-2014 (2). The age standardized 5-year survival rate of lung cancer in China increased slightly between 2003 and 2015, but still did not exceed 20.0% (3). The prognosis of lung cancer is closely related to the diagnostic stage, and the 5-year survival rate after surgery is almost 0 for stage IV patients, but >80% for stage I lung cancer patients (4).

The results of the National Lung Screening Trial (NLST), initiated in 2002, suggested that low-dose computed tomography (LDCT) screening could reduce lung cancer mortality by 20% (5). However, this project only screened people at high risk for lung cancer based on age and smoking history (55-74 years, smoked no less than 30 pack-years, and had no more than 15 years of smoking quit time). It is well known that smoking significantly increases the risk of lung cancer. Meta-analysis showed that the risk of lung cancer was 13.1 times higher among smokers than non-smokers in Europe and the United States [Hazard Ratio (HR)=13.1, 95% CI= 9.9-17.3] (6), much higher than the 2.77 times risk in the Chinese population [Odds Ratio (OR)=2.77, 95% CI=2.26-3.40] (7). This suggests that the current international standards for lung cancer screening based on smoking as the main indicator for high-risk populations may

not be suitable for the Chinese population, especially for Chinese non-smokers. Therefore, how to effectively predict the risk of lung cancer in non-smokers and then guide the more cost-effective LDCT screening is an effective way to achieve efficient early diagnosis and treatment of lung cancer.

Previous studies have constructed several lung cancer risk prediction models based on different characteristics of populations (8–38), but there is few lung cancer risk prediction models based on non-smokers in mainland of China. To this end, developing lung-cancer risk prediction tools for Chinese non-smokers based on risk factors consistently identified in previous studies becomes a priority (39). However, this is difficult and challenging. Unlike the situation of tobacco-driven lung cancer, there is no established risk factors dominating the development of lung cancer among non-smokers. Numerous risk factors have been suggested and their effects vary greatly by geographical region (40–43). For example, we note that the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) models do not seem to be useful for Asian non-smokers because PLCO only included about 2000 never-smokers of Asian ethnicity, of which 7 cases of lung cancer occurred (44). Indeed, none of the non-smokers in the PLCO (n=65,711) had a six-year risk >0.0151, using the PLCO_{M2014} that is analogous to PLCO_{M2012} and included non-smokers.

The model was developed based on the Cancer Screening Program in Urban China (CanSPUC) (45). With the focus on established risk factors for lung cancer routinely available in general cancer screening settings, we aimed to develop and internally validated a risk prediction model for lung cancer in Chinese non-smokers.

METHODS

Data Source and Subjects

This study was conducted within the framework of CanSPUC, an ongoing, nationwide, population-based cancer screening program in urban China. The purpose of CanSPUC is to screen five most prevalent cancers, including lung cancer, female breast cancer, liver cancer, upper gastrointestinal cancer, and colorectal cancer. The methodology of the CanSPUC has been previously described (45, 46). In brief, after signing a written informed consent, all eligible participants (40-74 years old) were interviewed by trained staffs to collect data on their exposure to risk factors and to evaluate their cancer risk using a defined clinical cancer risk score system. CanSPUC was launched in Henan province of China in October 2013, covering eight cities (Zhengzhou, Zhumadian, Anyang, Luoyang, Nanyang, Jiaozuo, Puyang, and Xinxiang). In this study, we used data from the first six years (from October 2013 to October 2019) in Henan province. Only those non-smokers (except former smokers) were included in this study. Subjects would be excluded if they have been already diagnosed with lung cancer.

Outcome, Variables and Measurements

All new cases of lung cancer in the study were ascertained through local cancer registry databases with a histologically confirmed diagnosis from October 1, 2013 to March 10, 2020

in mainland of China. Newly diagnosed lung cancers were classified by sites according to International Classification of Diseases, 10th version (ICD-10). Lung cancers were identified by ICD-10 of C33-C34. To identify potential risk factors for lung cancer, the following data were collected by self-report:

- (1) Demographic characteristics: including age, gender, race, height, weight and level of education. A low education was defined as primary school or below, medium education was defined junior or senior high school, and high-level education was defined as undergraduate or over. Body mass index (BMI) was calculated according to height and weight, and classified as “<18.5 kg/m²”, “18.5-23.9 kg/m²”, “24.0-27.9 kg/m²” and “≥28.0 kg/m²”.
- (2) Dietary habit: a) Dietary intake of the following food in the past two years: vegetables intake (<2.5kg/week, ≥2.5kg/week), fruit intake (<1.25kg/week, ≥1.25kg/week), roughage intake (<0.5kg/week, ≥0.5kg/week). Vegetables referred to green leaf plants and fungi, except for potato, sweet potato, and other starch. Roughage referred to the grains except white flour and rice. Food weight was determined before cooking. b) Taste preferences: heavy-salt diet (yes, no) and heavy-grease diet (yes, no).
- (3) Living environment, behavior and habits: a) Cooking oil fume (COF) exposure: exposure is considered as “none or a little”, if chimneys, fume extractors, or smoke-less pots was used during cooking; otherwise, it is considered as “a lot”. b) Physical activity: activities were categorized as Taijiquan/Qigong/Walking, long distance running/aerobics, ball games (basketball, table tennis, badminton, etc.), fast walking/yangko dance, swimming and other physical exercises (such as mountain climbing, rope skipping, shuttlecock kicking). Subjects who did exercise for at least three days with a total time ≥90 mins per week were categorized as “heavy physical activity”; otherwise, were categorized as “moderate or no physical activity”.
- (4) Comorbidities: including history of chronic respiratory disease, tuberculosis, chronic bronchitis, emphysema, asthma bronchiectasis and hyperlipidemia. All self-reported comorbidities required a diagnosis from professional medical institutions.
- (5) Family history of lung cancer: whether first-degree relatives, second-degree relatives or third-degree relatives had lung cancer or not.

Statistical Analysis

All statistical analyses were performed with the statistical software SAS version 9.4 (SAS Institute, Cary, NC) and R version 4.0.3 (The Free Software Foundation, Boston, MA, USA). The “rms” package was used to draw the nomogram. The “survivalROC” package was used to draw the ROC curves. The “ggplot2” package was used to draw the calibration curves. All tests were two-sided and p-values of 0.05 or less were considered to be statistically significant.

With the help of randomization codes produced by means of the PROC PLAN of the SAS system, the dataset was randomly

divided into training set and validation set with a 1:1 assignment ratio. The training set was used to create the risk prediction model, while the validation set was used to validate the performance of the model.

Descriptive statistics, expressed as proportions for categorical variables, were used to compare the characteristics of those with and without the outcome of developing lung cancer. Chi-squared tests for categorical variables were used to determine the univariate association between the baseline factors and lung cancer development. Continuous variables were described by means (standard deviation) or median (interquartile range, IQR).

In this study, the combined model based on all independent prognostic factors selected by the stepwise multivariable Cox regression ($P_{\text{entry}}=0.15$, $P_{\text{stay}}=0.10$) was used to construct a nomogram to assess the 1-, 3-, and 5-year estimates of the lung cancer risk in the training set. The calibration curves were used to evaluate the validity of the nomogram. The Kaplan–Meier curves were plotted for low-, medium-, and high-risk groups using the 33% and 66% quantiles for lung cancer according to the risk prediction model, and differences among the three curves were tested according to the log-rank test. The prediction performance of the 1-, 3- and 5-year estimates of the lung cancer risk was quantified by receiver operating characteristic (ROC) curves and the area under the curve (AUC) in the training set and validation set. The bootstrap sampling approach was used to evaluate the calibration of the present model by comparing the observed and predicted probabilities. Correction for deviation of estimates from observations (overfitting correction) estimates were based on predictions for a subset of the interval. The median absolute error is also used to evaluate the calibration performance.

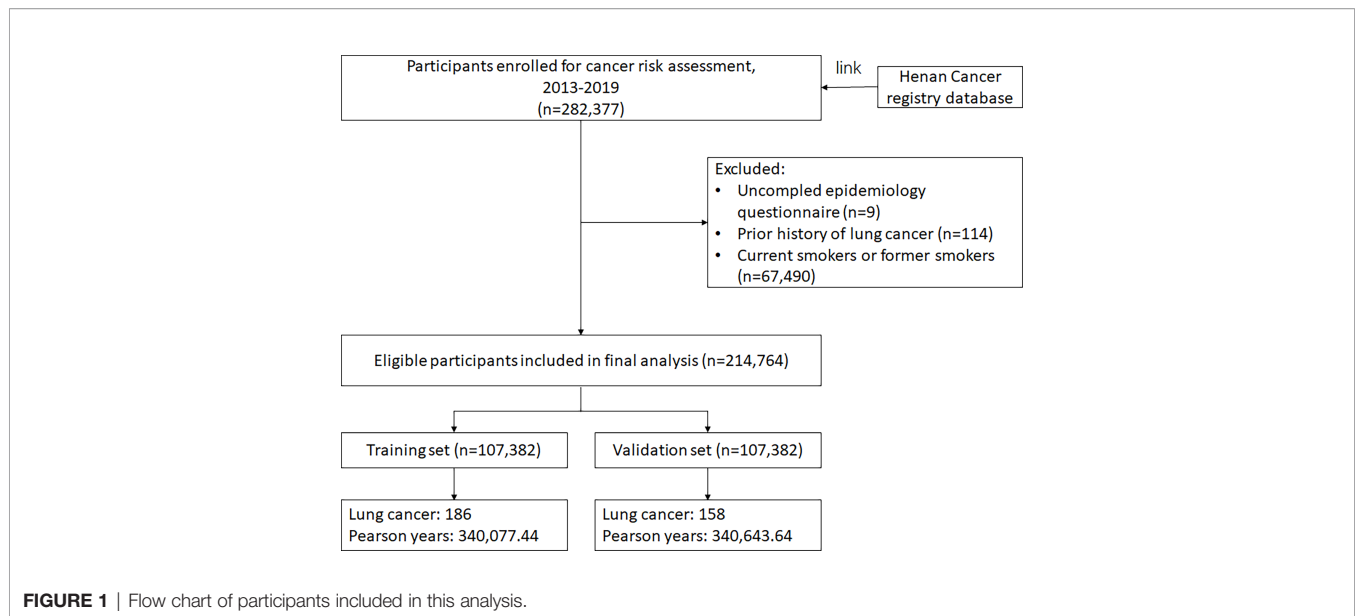
RESULTS

Characteristics of the Study Population

A total of 214,764 eligible subjects with a mean age of 55.19 years were included into this study, and 70.70% were females. Subjects were randomly divided into the training set (107,382 subjects) and validation set (107,382 subjects) (**Figure 1**). By March 2020, among 214,764 eligible participants, 344 lung cancer cases occurred in the follow-up yielding an incident density of 50.53/100,000 person-years. Compared with participants without lung cancer, lung cancer cases were more likely to have a low education, without passive smoking exposure, have a heavy physical activity and have a family history of lung cancer (all P values <0.05). Additional characteristics are presented in **Supplementary Table 1** and **Table 1**.

Development of the Lung Cancer Risk Assessment Model

Table 2 presents the HRs (95% CI) for each predictor. In the training set, age (≥55 years: 3.68, 1.60–8.43; ≥60 years: 5.51, 2.48–12.26; ≥65 years: 7.62, 3.43–16.92; ≥70 years: 9.03, 3.79–21.54), gender (male: 2.07, 1.53–2.79), education (low: 1.87, 1.05–3.33;



medium: 1.36, 0.81-2.31), family history of lung cancer (2.00, 1.25-3.20), history of tuberculosis (2.16, 0.87-5.37), and history of hyperlipidemia (0.61, 0.40-0.95) were independent risk factors of lung cancer. Thus, we used these variables to build the model. We plotted 1-year, 3-year, and 5-year lung cancer risk prediction nomogram (**Figure 2A**).

Predictive Performance of the Model

The risk predictions were stratified into low-, medium-, and high-risk groups and visualized by Kaplan–Meier curves, showing statistically significant differences between the groups by a log-rank test (**Figure 2B**, $P < 0.001$).

Using this model, the AUC was 0.753, 0.752, and 0.755 for 1-year, 3-year, and 5-year lung cancer risk in the training set, respectively. Stratified analysis by gender showed that the AUC of the model was higher among men (1-year: 0.776, 3-year: 0.780, and 5-year: 0.816) than women (1-year: 0.724, 3-year: 0.707, and 5-year: 0.694). Stratified analysis by age showed that the AUC of the model was higher among younger participants (<60 years) (1-year: 0.740, 3-year: 0.705, and 5-year: 0.664) than elder participants (≥ 60 years) (1-year: 0.628, 3-year: 0.648, and 5-year: 0.661). When examined by passive smoking status, the model yielded higher AUC for non-passive smokers (1-year: 0.762, 3-year: 0.756, and 5-year: 0.757) than passive smokers (1-year: 0.711, 3-year: 0.726, and 5-year: 0.738) (**Figure 3**). Calibration was satisfactory, with observed risks awfully close to the predicted risks (**Figure 4**).

Validation of the Lung Cancer Risk Model

The model showed a moderate predictive discrimination in the validation set, with the AUC was 0.668, 0.678, and 0.685 for 1-year, 3-year, and 5-year lung cancer risk (**Supplementary Figure 1**) and the satisfactory calibration of relative risk (**Supplementary Figure 2**).

DISCUSSION

In this study, using data from a large perspective lung cancer screening cohort studies, we developed and internally validated a simple risk prediction model for lung cancer in non-smokers, based on six widely available variables, including demographics (age, gender, education), comorbidities (tuberculosis, hyperlipidemia) and family history of lung cancer. Our results showed that the model has good discriminatory accuracy and goodness-of-fit for both men and women, non-passive smokers and passive smokers.

For non-smokers, several risk factors for lung cancer have been identified, including passive smoking (47, 48), previous lung diseases [tuberculosis, chronic bronchitis, emphysema, previous lung diseases (COPD)] (49), indoor radon (50), cooking oil fumes (51) and family history of lung cancer (52). The risk factors for lung cancer identified in our study, such as age, gender, family history of lung cancer, history of tuberculosis, are consistent with the findings. The most dominant risk factors for lung cancer in non-smokers is age, and our study showed that elder age was the main risk factor for lung cancer and the risk was more than 9 times higher in age group of 70-74 years than in the age group of 40-44 years. Besides, being male remains a risk factor for lung cancer in non-smokers in our study, even though more than 50% lung cancers were non-smokers in women in Southeast Asia compared to approximately 2-6% in men in Western series (41, 42, 53). Just like other prediction models, such as Bach model (8), LLP (Liverpool Lung Project) model (10) and PLCO_{M2012} model (54), education levels was included in our model as a protection factor.

Another important finding was that history of hyperlipidemia [increased total cholesterol (TC), or triglycerides (TG), or low-density lipoprotein cholesterol (LDL-C), or decreased high-density lipoprotein cholesterol (HDL-C)] exposure might

TABLE 1 | Baseline characteristics of the study population in the training set.

	Total no. (%)	No lung cancer, n (%)	Lung cancer, n (%)	χ^2	P-value
All participants	107382	107196	186		
Person-years, median(IQR)	2.95 (1.71-4.83)	2.95 (1.71-4.83)	1.47 (0.78-2.33)		
Age, mean \pm SD, years	55.16 \pm 8.78	55.14 \pm 8.78	61.53 \pm 7.47		
Age (years)				98.24	<0.001
40-44	14028 (13.06)	14021 (99.95)	7 (0.05)		
45-49	20109 (18.73)	20101 (99.96)	8 (0.04)		
50-54	19769 (18.41)	19750 (99.90)	19 (0.10)		
55-59	16083 (14.98)	16054 (99.82)	29 (0.18)		
60-64	17556 (16.35)	17508 (99.73)	48 (0.27)		
65-69	14356 (13.37)	14302 (99.62)	54 (0.38)		
70-74	5481 (5.10)	5460 (99.62)	21 (0.38)		
Gender				16.73	<0.001
Male	31531 (29.36)	31451 (99.75)	80 (0.25)		
Female	75851 (70.64)	75745 (99.86)	106 (0.14)		
Race				0.44	0.506
Han nationality	105549 (98.29)	105365 (99.83)	184 (0.17)		
Others	1833 (1.71)	1831 (99.89)	2 (0.11)		
Education ^a				14.34	0.001
Low	20139 (18.75)	20086 (99.74)	53 (0.26)		
Medium	71634 (66.71)	71517 (99.84)	117 (0.16)		
High	15609 (14.54)	15593 (99.90)	16 (0.10)		
BMI (kg/m ²)				2.82	0.419
<18.5	1381 (1.29)	1377 (99.71)	4 (0.29)		
18.5-23.9	47588 (44.32)	47498 (99.81)	90 (0.19)		
24.0-28.0	46882 (43.66)	46810 (99.85)	72 (0.15)		
\geq 28.0	11531 (10.74)	11511 (99.83)	20 (0.17)		
Vegetables intake				0.03	0.861
\geq 2.5kg/week	56467 (52.59)	56368 (99.82)	99 (0.18)		
<2.5kg/week	50915 (47.41)	50828 (99.83)	87 (0.17)		
Fruit intake				0.07	0.785
\geq 1.25kg/week	63026 (58.69)	62915 (99.82)	111 (0.18)		
<1.25kg/week	44356 (41.31)	44281 (99.83)	75 (0.17)		
Roughage intake				0.47	0.492
\geq 0.5kg/week	73524 (68.47)	73401 (99.83)	123 (0.17)		
<0.5kg/week	33858 (31.53)	33795 (99.81)	63 (0.19)		
Heavy-salt diet				1.57	0.209
No	88623 (82.53)	88463 (99.82)	160 (0.18)		
Yes	18759 (17.47)	18733 (99.86)	26 (0.14)		
Heavy-grease diet				0.07	0.793
No	89672 (83.51)	89518 (99.83)	154 (0.17)		
Yes	17710 (16.49)	17678 (99.82)	32 (0.18)		
Cooking oil fume exposure				0.03	0.853
None or a little	95144 (88.60)	94980 (99.83)	164 (0.17)		
A lot	12238 (11.40)	12216 (99.82)	22 (0.18)		
Passive smoking				4.65	0.031
No	76618 (71.35)	76472 (99.81)	146 (0.19)		
Yes	30764 (28.65)	30724 (99.87)	40 (0.13)		
Alcohol Drinking				0.46	0.794
Never	96004 (89.40)	95835 (99.82)	169 (0.18)		
Current	9598 (8.94)	9584 (99.85)	14 (0.15)		
Former	1780 (1.66)	1777 (99.83)	3 (0.17)		
Physical activity				3.38	0.066
Moderate or no	56293 (52.42)	56208 (99.85)	85 (0.15)		
Heavy	51089 (47.58)	50988 (99.80)	101 (0.20)		
Family history of lung cancer				4.69	0.030
No	99660 (92.81)	99495 (99.83)	165 (0.17)		
Yes	7722 (7.19)	7701 (99.73)	21 (0.27)		
History of chronic respiratory disease				1.93	0.165
No	93185 (86.78)	93030 (99.83)	155 (0.17)		
Yes	14197 (13.22)	14166 (99.78)	31 (0.22)		
History of tuberculosis				4.58	0.032
No	106230 (98.93)	106049 (99.83)	181 (0.17)		
Yes	1152 (1.07)	1147 (99.57)	5 (0.43)		

(Continued)

TABLE 1 | Continued

	Total no. (%)	No lung cancer, n (%)	Lung cancer, n (%)	χ^2	P-value
History of chronic bronchitis				1.55	0.213
No	96485 (89.85)	96323 (99.83)	162 (0.17)		
Yes	10897 (10.15)	10873 (99.78)	24 (0.22)		
History of emphysema				4.51	0.034
No	106543 (99.22)	106361 (99.83)	182 (0.17)		
Yes	839 (0.78)	835 (99.52)	4 (0.48)		
History of asthma bronchiectasis				0.01	0.925
No	104612 (97.42)	104431 (99.83)	181 (0.17)		
Yes	2770 (2.58)	2765 (99.82)	5 (0.18)		
History of hyperlipidemia				0.50	0.481
No	91561 (85.27)	91399 (99.82)	162 (0.18)		
Yes	15821 (14.73)	15797 (99.85)	24 (0.15)		

^aLow, primary school or below; Medium, junior or senior high school; High, undergraduate or over.
IQR, Interquartile range; BMI, body mass index.

decrease the risk of lung cancer, despite a small effect. Since the 1980s, several epidemiological studies have investigated the associations of TC, TG, and HDL-C with lung cancer risk in non-smokers but have shown markedly contrasting results due to differences in the classification of smoking status, lack of prospective cohort study designs, relatively modest sample sizes and other potential bias (55–58). Lyu et al. (58) conducted a prospective cohort study among over 100 thousand Chinese males and found that both low and high TC levels, both low and high TG levels, and low LDL-C levels increased lung cancer risk in non-smokers. Besides, many studies reported an inverse relationship between TC (56, 59), LDL-C (60) and lung cancer incidence, to some extent, consistent with our findings. More

epidemiologic, molecular and biochemical studies are needed to test this hypothesis.

In addition to credible predictors, a risk prediction model should also meet performance standards related to discrimination defined as the ability to distinguish lung cancer cases from controls, and calibration defined as the consistency between observed and predicted risk for lung cancer. The rapid increase in the number of lung cancer risk prediction model studies since 2010 reflects the current need for the use of predictive models to guide population splitting. Initially, models focused on the use of traditional epidemiological risk factors such as age, smoking history, personal history of disease and family history of cancer, such as the Bach model (8), Spitz model (9), LLP model (10) and PLCO_{M2012} model

TABLE 2 | Multivariable Cox-regression prediction model of lung cancer risk in training set.

Variables	β coefficient	se	HR (95% CI)	χ^2	P
Age (years)					
40-44			1.00		
45-49	-0.24	0.52	0.78 (0.28-2.16)	0.22	0.637
50-54	0.67	0.44	1.96 (0.82-4.66)	2.29	0.130
55-59	1.30	0.42	3.68 (1.60-8.43)	9.47	0.002
60-64	1.71	0.41	5.51 (2.48-12.26)	17.53	<0.001
65-69	2.03	0.41	7.62 (3.43-16.92)	24.92	<0.001
70-74	2.20	0.44	9.03 (3.79-21.54)	24.65	<0.001
Gender					
Male	0.73	0.15	2.07 (1.53-2.79)	22.63	<0.001
Female			1.00		
Education ^a					
Low	0.62	0.30	1.87 (1.05-3.33)	4.45	0.035
Medium	0.31	0.27	1.36 (0.81-2.31)	1.33	0.249
High			1.00		
Family history of lung cancer					
No			1.00		
Yes	0.69	0.24	2.00 (1.25-3.20)	8.24	0.004
History of tuberculosis					
No			1.00		
Yes	0.77	0.46	2.16 (0.87-5.37)	2.75	0.097
History of hyperlipidemia					
No			1.00		
Yes	-0.49	0.22	0.61 (0.40-0.95)	4.80	0.028

^aLow, primary school or below; Medium, junior or senior high school; High, undergraduate or over.

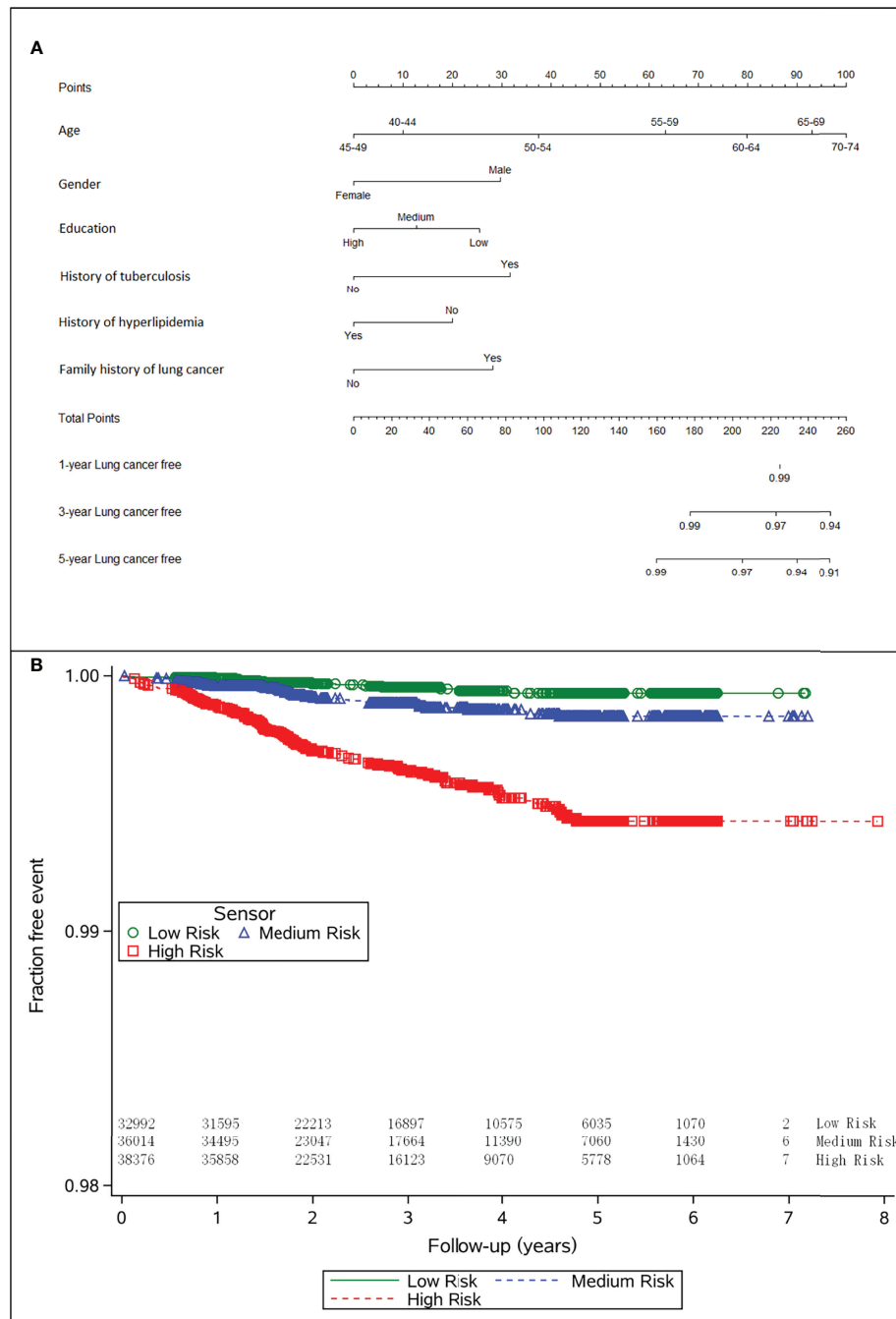


FIGURE 2 | (A) Nomogram to calculate the personal 1-, 3- and 5-year risk of lung cancer risk, and **(B)** the lung cancer incidence across different cancer risk categories.

(54). To our knowledge, the present study is one of the few studies to model lung cancer risk prediction among non-smoking men and women in mainland China. It is hard to directly compare the discriminatory performance of risk prediction models as each was developed in different populations with varying baseline risks or lengths of follow-up time. Nevertheless, each of the models' discriminative ability was relatively similar, with C-statistics

ranges from 0.72 to 0.86. Our model showed comparable predictive performance compared with previous studies.

Specific strengths and limitations deserve careful attention when interpreting our results. A major strength of our study is the fact that our analyses were based on a large-scale population-based cancer screening program in mainland of China. Furthermore, the variables included in this model could be

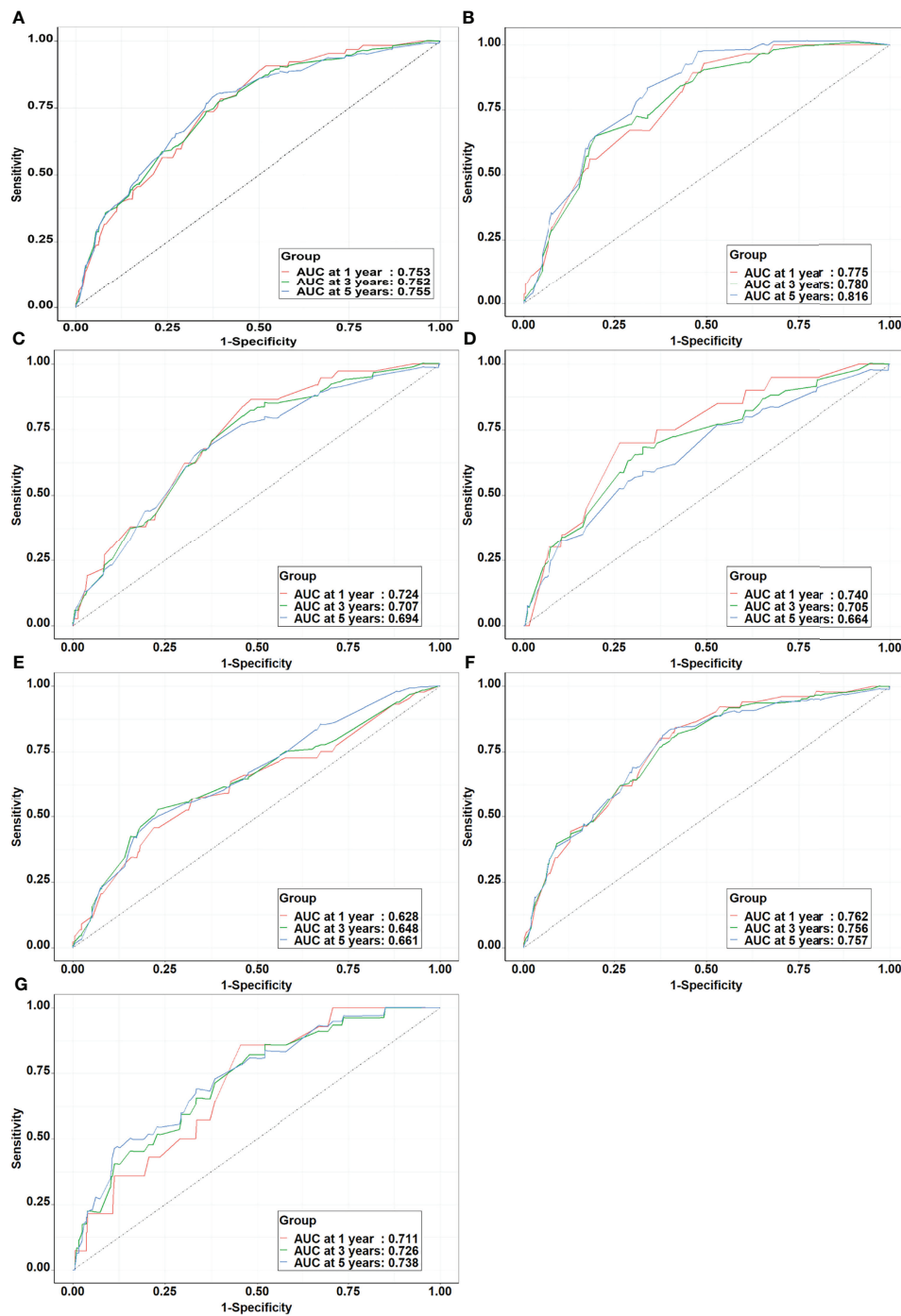


FIGURE 3 | The receiver operating characteristic curves of prediction models in the training set. (A) Whole population; (B) Male; (C) Female; (D) Age < 60 years; (E) Age ≥ 60 years; (F) Non-passive smokers; (G) Passive smokers.

easily collected and updated without any imaging, sophisticated testing or calculation. Moreover, the model will not only be used as a practical tool to triage high risk patients in non-smokers, but also have implications for public health measures, such as guidelines for the prevention of lung cancer in non-smokers.

However, limitations include that the self-report data might be subject to social desirability and recall bias. However, given the good data acquisition and quality control, most information is believed to be reliable. Secondly, the performance of our risk prediction model was not validated on an external dataset.

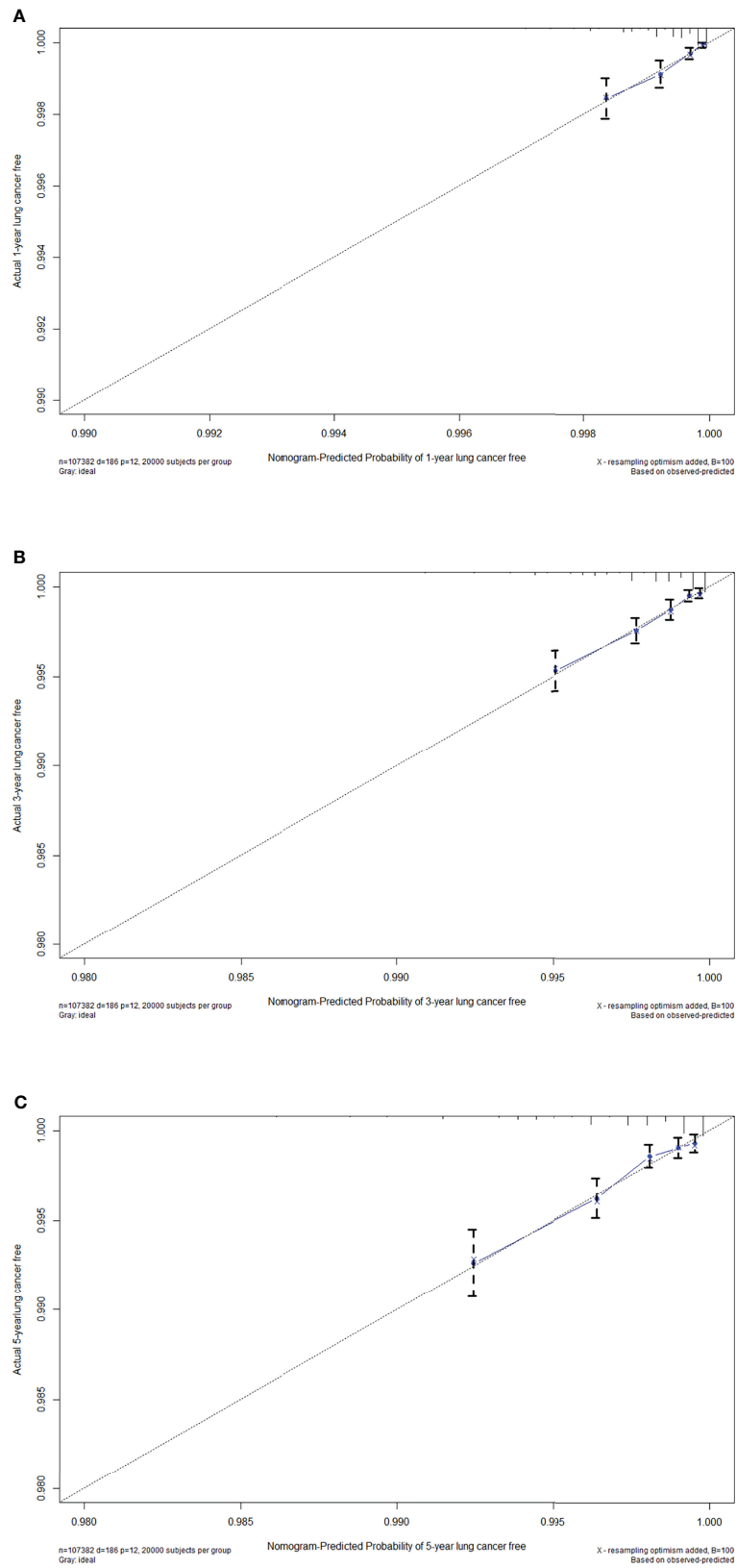


FIGURE 4 | Calibration curves of the nomogram for (A) 1-year, (B) 3-year and (C) 5-year lung cancer free in the training set.

However, the results of the internal validation suggest promisingly that this model will obtain well performance when applied to other populations.

CONCLUSIONS

In summary, we developed and internally validated a simple risk prediction model for lung cancer in non-smokers based on a large-scale lung cancer screening program in China. The model has good discrimination and could be used as a tool for triaging high-risk patients to prevent lung cancer in non-smokers. Further prospective studies are required to validate the model in external populations.

DATA AVAILABILITY STATEMENT

The datasets for this manuscript are not publicly available because all our data are under regulation of both the National Cancer Center of China and Henan Cancer Hospital. Requests to access the datasets should be directed to Shaokai Zhang, shaokaizhang@126.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Henan Cancer Hospital

REFERENCES

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
- Shi JF, Wang L, Wu N, Li JL, Hui ZG, Liu SM, et al. Clinical Characteristics and Medical Service Utilization of Lung Cancer in Chin-2014: Overall Design and Results From a Multicenter Retrospective Epidemiologic Survey. *Lung Cancer* (2019) 128:91–100. doi: 10.1016/j.lungcan.2018.11.031
- Zeng H, Chen W, Zheng R, Zhang S, Ji JS, Zou X, et al. Changing Cancer Survival in China During 2003–15: A Pooled Analysis of 17 Population-Based Cancer Registries. *Lancet Glob Health* (2018) 6(5):e555–67. doi: 10.1016/s2214-109x(18)30127-x
- Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. *Chest* (2017) 151(1):193–203. doi: 10.1016/j.chest.2016.10.010
- 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. doi: 10.1056/NEJMoa1102873
- Ordóñez-Mena JM, Schöttker B, Mons U, Jenab M, Freisling H, Bueno-de-Mesquita B, et al. Quantification of the Smoking-Associated Cancer Risk With Rate Advancement Periods: Meta-Analysis of Individual Participant Data From Cohorts of the CHANCES Consortium. *BMC Med* (2016) 14:62. doi: 10.1186/s12916-016-0607-5
- 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. doi: 10.3760/cma.j.cn112152-20210119-00060

(no. 2021-KY-0028-001). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Conception and design: L-WG, Z-YL, J-GZ and S-KZ. Statistical Analyses: L-WG and L-YZ. Data acquisition and data interpretation: L-WG, Q-CM, L-YZ, QC, YL, H-FX, R-HK, L-YZ, X-QC, S-ZL, X-BS and S-KZ. Drafting of the article: L-WG. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Natural Science Foundation of Henan Province (No. 212300410261), the Henan Province Medical Science and Technology Tackling Program (SBGJ202001004), the National Key R&D Program of China (2018YFC1315000/2018YFC1315001) and the CAMS Innovation Fund for Medical Science (2019-I2M-2-002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2021.766939/full#supplementary-material>

- Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in Lung Cancer Risk Among Smokers. *J Natl Cancer Inst* (2003) 95(6):470–8. doi: 10.1093/jnci/95.6.470
- Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, et al. A Risk Model for Prediction of Lung Cancer. *J Natl Cancer Inst* (2007) 99(9):715–26. doi: 10.1093/jnci/djk153
- 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. doi: 10.1038/sj.bjc.6604158
- Etzel CJ, Kachroo S, Liu M, D'Amelio A, Dong Q, Cote ML, et al. Development and Validation of a Lung Cancer Risk Prediction Model for African-Americans. *Cancer Prev Res (Phila)* (2008) 1(4):255–65. doi: 10.1158/1940-6207.capr-08-0082
- Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, et al. An Expanded Risk Prediction Model for Lung Cancer. *Cancer Prev Res (Phila)* (2008) 1(4):250–4. doi: 10.1158/1940-6207.capr-08-0060
- Young RP, Hopkins RJ, Hay BA, Epton MJ, Mills GD, Black PN, et al. Lung Cancer Susceptibility Model Based on Age, Family History and Genetic Variants. *PLoS One* (2009) 4(4):e5302. doi: 10.1371/journal.pone.0005302
- D'Amelio AM Jr, Cassidy A, Asomaning K, Raji OY, Duffy SW, Field JK, et al. Comparison of Discriminatory Power and Accuracy of Three Lung Cancer Risk Models. *Br J Cancer* (2010) 103(3):423–9. doi: 10.1038/sj.bjc.6605759
- Raji OY, Agbaje OF, Duffy SW, Cassidy A, Field JK. Incorporation of a Genetic Factor Into an Epidemiologic Model for Prediction of Individual Risk of Lung Cancer: The Liverpool Lung Project. *Cancer Prev Res (Phila)* (2010) 3(5):664–9. doi: 10.1158/1940-6207.capr-09-0141
- Maisonneuve P, Bagnardi V, Bellomi M, Spaggiari L, Pelosi G, Rampinelli C, et al. Lung Cancer Risk Prediction to Select Smokers for Screening CT—A Model Based on the Italian COSMOS Trial. *Cancer Prev Res (Phila)* (2011) 4(11):1778–89. doi: 10.1158/1940-6207.capr-11-0026

17. 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. doi: 10.1093/jnci/djr173
18. Tammemagi MC, Lam SC, McWilliams AM, Sin DD. Incremental Value of Pulmonary Function and Sputum DNA Image Cytometry in Lung Cancer Risk Prediction. *Cancer Prev Res (Phila)* (2011) 4(4):552–61. doi: 10.1158/1940-6207.capr-10-0183
19. 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. doi: 10.1158/1940-6207.capr-11-0237
20. Li H, Yang L, Zhao X, Wang J, Qian J, Chen H, et al. Prediction of Lung Cancer Risk in a Chinese Population Using a Multifactorial Genetic Model. *BMC Med Genet* (2012) 13:118. doi: 10.1186/1471-2350-13-118
21. Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, et al. Predictive Accuracy of the Liverpool Lung Project Risk Model for Stratifying Patients for Computed Tomography Screening for Lung Cancer: A Case-Control and Cohort Validation Study. *Ann Intern Med* (2012) 157(4):242–50. doi: 10.7326/0003-4819-157-4-201208210-00004
22. Park S, Nam BH, Yang HR, Lee JA, Lim H, Han JT, et al. Individualized Risk Prediction Model for Lung Cancer in Korean Men. *PLoS One* (2013) 8(2):e54823. doi: 10.1371/journal.pone.0054823
23. Spitz MR, Amos CI, Land S, Wu X, Dong Q, Wenzlaff AS, et al. Role of Selected Genetic Variants in Lung Cancer Risk in African Americans. *J Thorac Oncol* (2013) 8(4):391–7. doi: 10.1097/JTO.0b013e318283da29
24. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection Criteria for Lung-Cancer Screening. [Erratum Appears in N Engl J Med. 2013 Jul 25;369(4):394]. *New Engl J Med* (2013) 368(8):728–36. doi: 10.1056/NEJMoa1211776
25. Veronesi G, Maisonneuve P, Rampinelli C, Bertolotti R, Petrella F, Spaggiari L, et al. Computed Tomography Screening for Lung Cancer: Results of Ten Years of Annual Screening and Validation of Cosmos Prediction Model. *Lung Cancer* (2013) 82(3):426–30. doi: 10.1016/j.lungcan.2013.08.026
26. El-Zein RA, Lopez MS, D'Amelio AM Jr, Liu M, Munden RF, Christiani D, et al. The Cytokinesis-Blocked Micronucleus Assay as a Strong Predictor of Lung Cancer: Extension of a Lung Cancer Risk Prediction Model. *Cancer Epidemiol Biomarkers Prev* (2014) 23(11):2462–70. doi: 10.1158/1055-9965.epi-14-0462
27. Li K, Huesing A, Sookthai D, Bergmann M, Boeing H, Becker N, et al. Selecting High-Risk Individuals for Lung Cancer Screening: A Prospective Evaluation of Existing Risk Models and Eligibility Criteria in the German EPIC Cohort. *Cancer Prev Res* (2015) 8(9):777–85. doi: 10.1158/1940-6207.CAPR-14-0424
28. Marcus MW, Chen Y, Raji OY, Duffy SW, Field JK. LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence. *Cancer Prev Res (Phila)* (2015) 8(6):570–5. doi: 10.1158/1940-6207.capr-14-0438
29. Wang X, Ma K, Cui J, Chen X, Jin L, Li W. An Individual Risk Prediction Model for Lung Cancer Based on a Study in a Chinese Population. *Tumori J January/February* (2015) 101(1):16–23. doi: 10.5301/tj.5000205
30. Marcus MW, Raji OY, Duffy SW, Young RP, Hopkins RJ, Field JK. Incorporating Epistasis Interaction of Genetic Susceptibility Single Nucleotide Polymorphisms in a Lung Cancer Risk Prediction Model. *Int J Oncol* (2016) 49(1):361–70. doi: 10.3892/ijo.2016.3499
31. Wang X, Ma K, Chi L, Cui J, Jin L, Hu J-F, et al. Combining Telomerase Reverse Transcriptase Genetic Variant Rs2736100 With Epidemiologic Factors in the Prediction of Lung Cancer Susceptibility. *J Of Cancer* (2016) 7(7):846–53. doi: 10.7150/jca.13437
32. 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. doi: 10.1038/srep36482
33. 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. doi: 10.1200/jco.2016.69.2467
34. Weber M, Yap S, Goldsbury D, Manners D, Tammemagi M, Marshall H, et al. Identifying High Risk Individuals for Targeted Lung Cancer Screening: Independent Validation of the PLCOm2012 Risk Prediction Tool. *Int J Cancer* (2017) 141(2):242–53. doi: 10.1002/ijc.30673
35. 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. doi: 10.1111/cas.13509
36. 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. doi: 10.7326/m17-2701
37. Markaki M, Tsamardinos I, Langhammer A, Lagani V, Hveem K, Roe OD. A Validated Clinical Risk Prediction Model for Lung Cancer in Smokers of All Ages and Exposure Types: A HUNT Study. *EBioMedicine* (2018) 31:36–46. doi: 10.1016/j.ebiom.2018.03.027
38. 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. doi: 10.1002/cam4.3025
39. Lam S. Lung Cancer Screening in Never-Smokers. *J Thorac Oncol* (2019) 14(3):336–7. doi: 10.1016/j.jtho.2018.12.019
40. Toh CK, Gao F, Lim WT, Leong SS, Fong KW, Yap SP, et al. Never-Smokers With Lung Cancer: Epidemiologic Evidence of a Distinct Disease Entity. *J Clin Oncol* (2006) 24(15):2245–51. doi: 10.1200/jco.2005.04.8033
41. Sun S, Schiller JH, Gazdar AF. Lung Cancer in Never Smokers—a Different Disease. *Nat Rev Cancer* (2007) 7(10):778–90. doi: 10.1038/nrc2190
42. Couraud S, Zalcman G, Milleron B, Morin F, Souquet PJ. Lung Cancer in Never Smokers—a Review. *Eur J Cancer* (2012) 48(9):1299–311. doi: 10.1016/j.ejca.2012.03.007
43. Sisti J, Boffetta P. What Proportion of Lung Cancer in Never-Smokers can be Attributed to Known Risk Factors? *Int J Cancer* (2012) 131(2):265–75. doi: 10.1002/ijc.27477
44. 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. doi: 10.1371/journal.pmed.1001764
45. Guo LW, Chen Q, Shen YC, Meng QC, Zheng LY, Wu Y, et al. Evaluation of a Low-Dose Computed Tomography Lung Cancer Screening Program in Henan, China. *JAMA Netw Open* (2020) 3(11):e2019039. doi: 10.1001/jamanetworkopen.2020.19039
46. Guo L, Zhang S, Liu S, Zheng L, Chen Q, Cao X, et al. Determinants of Participation and Detection Rate of Upper Gastrointestinal Cancer From Population-Based Screening Program in China. *Cancer Med* (2019) 8(16):7098–107. doi: 10.1002/cam4.2578
47. Hackshaw AK, Law MR, Wald NJ. The Accumulated Evidence on Lung Cancer and Environmental Tobacco Smoke. *Bmj* (1997) 315(7114):980–8. doi: 10.1136/bmj.315.7114.980
48. Stayner L, Bena J, Sasco AJ, Smith R, Steenland K, Kreuzer M, et al. Lung Cancer Risk and Workplace Exposure to Environmental Tobacco Smoke. *Am J Public Health* (2007) 97(3):545–51. doi: 10.2105/ajph.2004.061275
49. Brenner DR, McLaughlin JR, Hung RJ. Previous Lung Diseases and Lung Cancer Risk: A Systematic Review and Meta-Analysis. *PLoS One* (2011) 6(3):e17479. doi: 10.1371/journal.pone.0017479
50. Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, et al. Radon in Homes and Risk of Lung Cancer: Collaborative Analysis of Individual Data From 13 European Case-Control Studies. *Bmj* (2005) 330(7485):223. doi: 10.1136/bmj.38308.477650.63
51. Zhao Y, Wang S, Aunan K, Seip HM, Hao J. Air Pollution and Lung Cancer Risks in China—a Meta-Analysis. *Sci Total Environ* (2006) 366(2–3):500–13. doi: 10.1016/j.scitotenv.2005.10.010
52. Lisowska J, Foretova L, Dabek J, Zaridze D, Szeszenia-Dabrowska N, Rudnai P, et al. Family History and Lung Cancer Risk: International Multicentre Case-Control Study in Eastern and Central Europe and Meta-Analyses. *Cancer Causes Control* (2010) 21(7):1091–104. doi: 10.1007/s10552-010-9537-2
53. Scagliotti GV, Longo M, Novello S. Nonsmall Cell Lung Cancer in Never Smokers. *Curr Opin Oncol* (2009) 21(2):99–104. doi: 10.1097/CCO.0b013e328321049e
54. Tammemägi 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. doi: 10.1056/NEJMoa1211776
55. Ulmer H, Borena W, Rapp K, Klenk J, Strasak A, Diem G, et al. Serum Triglyceride Concentrations and Cancer Risk in a Large Cohort Study in Austria. *Br J Cancer* (2009) 101(7):1202–6. doi: 10.1038/sj.bjc.6605264

56. Kitahara CM, Berrington de González A, Freedman ND, Huxley R, Mok Y, Jee SH, et al. Total Cholesterol and Cancer Risk in a Large Prospective Study in Korea. *J Clin Oncol* (2011) 29(12):1592–8. doi: 10.1200/jco.2010.31.5200
57. Chandler PD, Song Y, Lin J, Zhang S, Sesso HD, Mora S, et al. Lipid Biomarkers and Long-Term Risk of Cancer in the Women's Health Study. *Am J Clin Nutr* (2016) 103(6):1397–407. doi: 10.3945/ajcn.115.124321
58. Lyu Z, Li N, Wang G, Feng X, Chen S, Su K, et al. Independent and Joint Associations of Blood Lipids and Lipoproteins With Lung Cancer Risk in Chinese Males: A Prospective Cohort Study. *Int J Cancer* (2019) 144(12):2972–84. doi: 10.1002/ijc.32051
59. Ahn J, Lim U, Weinstein SJ, Schatzkin A, Hayes RB, Virtamo J, et al. Prediagnostic Total and High-Density Lipoprotein Cholesterol and Risk of Cancer. *Cancer Epidemiol Biomarkers Prev* (2009) 18(11):2814–21. doi: 10.1158/1055-9965.Epi-08-1248
60. Benn M, Tybjaerg-Hansen A, Stender S, Frikke-Schmidt R, Nordestgaard BG. Low-Density Lipoprotein Cholesterol and the Risk of Cancer: A Mendelian Randomization Study. *J Natl Cancer Inst* (2011) 103(6):508–19. doi: 10.1093/jnci/djr008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Guo, Lyu, Meng, Zheng, Chen, Liu, Xu, Kang, Zhang, Cao, Liu, Sun, Zhang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.