Additional file 1: Methods

1 Data source and study participants

CHinese Electronic health Records Research in Yinzhou (CHERRY), based on the regional health information system of Yinzhou District, Ningbo City, Zhejiang Province, integrates different administrative databases of general demographic characteristics, health examination information, inpatient and outpatient electronic medical records (EMRs), health insurance database, disease management and death certificates from the Yinzhou Center for Disease Control and Prevention through a unique and encoded identifier assigned to each individual, which has provided disease and health-related information to 98% of the resident population in Yinzhou District since 2009.

Using the retrospective cohort study design, we collected the individual participant data of all permanent residents in the residents' health files of the Yinzhou Regional Health Information Platform from 2006 to 2016. We used electronic medical records to obtain the follow-up on their hospital visits from 2017 to 2021. Potential high-risk population with all available information were included. Exclusion criteria were participants without nodular lung masses, asthma, interstitial lung disease, primary pneumonia, secondary pneumonia, tuberculosis, pulmonary dysfunction, chronic bronchitis, secondary lung dilatation, pneumatocele (one of these), or absent lung CT findings. Participants were defined as being at standard risk for lung cancer if they had exclusively one of the following conditions: primary pneumonia, asthma, chronic bronchitis, or active tuberculosis.

2 Definitions of outcomes

The primary outcome of the study was an ICD-10 code of C34 or a clinical imaging diagnosis of lung cancer.

3 Model development and performance

We randomly selected 70% (*N*=111269) of participants as the training cohort for model building and hyperparameter optimization and 30% (*N*=47688) as the validation cohort. In this study, we evaluated four model-building approaches: decision tree (DT), light gradient boosting machine (LightGBM), extreme gradient boosting (XGBoost), and logistic regression (LR). First, a univariate logistic regression analysis was performed to obtain a baseline model for predicting the risk of lung cancer among high-risk population. Variables with a p-value less than 0.1 in the univariate analysis were included in the Least absolute shrinkage and selection operator (LASSO) analysis to select the optimal predictors associated with lung cancer risk to reduce high-dimensional datasets among population. 24 variables that yielded non-zero coefficient values in the LASSO regression analysis were retained for model construction.

Multiple imputation by chained equations (MICE) method based on the random forest was used for multiple imputations of missing variables, including obsolete TB, uric acid, etc., achieving good convergence after one imputation with 30 iterations.

In the hyperparameter tuning process of each algorithm, we used a grid search to identify the

optimal values for possible combinations of parameter values for each algorithm. The model with the highest area under the curve (AUC) of the receiver operating characteristic (ROC) was identified and selected as the optimal model. After acquiring the parameters and achieving the desired accuracy in the training set, the produced model was tested again in the validation set to evaluate its predictive power. The optimized parameters of each algorithm are shown in Supplementary data, Table S1.

AUC of the ROC curve, sensitivity, specificity, and accuracy were used to assess the performance of the predictive model. Decision curve analyses were used to assess the clinical utility of the lung cancer risk prediction model by quantifying net benefit across different probability thresholds. Net benefit was calculated by subtracting the proportion of false positives from true positives and evaluating clinical utility. Calibration was measured using the calibration plot. For model explanations, we used SHAP scores to demonstrate the degree of contribution of variables to model prediction.

4 Statistical analysis

Descriptive statistics for categorical data were presented as counts and percentages, while the median and interquartile range (IQR) were used for continuous data. Pearson's Chi-square tests were used to compare categorical variables between different groups, while the nonparametric Wilcoxon rank test was used to compare continuous variables. Data for categorical variables were processed using the one-hot encoding method. Statistical significance was set at a level of 0.05, and all tests were conducted using a two-tailed approach. The logistic regression model and machine learning models were performed using R 4.0.5 (R Foundation for Statistical Computing).

Supplementary Table S1. Final selected hyperparameters

Model	Final selected hyperparameters after tuning
Decision Tree	
Tree Depth	7
Minimal Node Size	6
Cost-Complexity Parameter	0.00002
LightGBM	
Tree Depth	3
Trees	208
Learning Rate	0.04260
mtry	8
Minimal Node Size	8
Minimum Loss Reduction	0.16300
XGBoost	
Tree Depth	6
Learning Rate	0.02170
mtry	7
Minimal Node Size	20
Minimum Loss Reduction	0.00996
Proportion Observations Sampled	0.86400
LASSO	
Lambda. min	0.00080
Lambda.1se	0.00350

Supplementary Table S2. Baseline characteristics of study population

Variable		Overall, *N* = 158,957[1]	Non-lung cancer, *N* = 151,277[1]	Lung cancer, *N* = 7,680[1]	*P*-value[2]
Age (IQR)		57 (45, 67)	57 (45, 67)	62 (53, 69)	< 0.001
Gender					< 0.001
	Female	79,857 (50.24%)	76,746 (50.73%)	3,111 (40.51%)	
	Male	79,100 (49.76%)	74,531 (49.27%)	4,569 (59.49%)	
BMI (IQR)		22.3 (20.4, 24.4)	22.3 (20.4, 24.4)	22.4 (20.5, 24.4)	0.070
hs-CRP (IQR)		4 (1, 13)	4 (1, 12)	4 (1, 14)	0.006
ESR (IQR)		13 (6, 26)	13 (6, 26)	14 (6, 30)	< 0.001
HbA 1c (IQR)		5.50 (5.00, 6.00)	5.50 (5.00, 6.00)	5.50 (4.90, 6.00)	0.037
SBP (IQR)		125 (120, 132)	125 (120, 132)	125 (120, 132)	< 0.001
DBP (IQR)		78 (71, 80)	78 (71, 80)	78 (72, 80)	0.024
TC (IQR)		4.79 (4.16, 5.46)	4.79 (4.16, 5.46)	4.82 (4.17, 5.48)	0.031
LDL-C (IQR)		2.68 (2.19, 3.21)	2.68 (2.19, 3.20)	2.71 (2.23, 3.23)	0.001
TG (IQR)		1.26 (0.91, 1.82)	1.26 (0.91, 1.82)	1.27 (0.93, 1.79)	0.678
HDL-C (IQR)		1.34 (1.15, 1.58)	1.35 (1.15, 1.58)	1.33 (1.14, 1.55)	< 0.001
Uric acid (IQR)		307 (248, 376)	307 (248, 376)	312 (249, 378)	0.101
Number of psychiatric department visits (IQR)		1 (0, 3)	1 (0, 3)	1 (0, 3)	0.078
Smoke					< 0.001
	Never	124,496 (78.32%)	118,890 (78.59%)	5,606 (72.99%)	
	Seldom	3,777 (2.38%)	3,459 (2.29%)	318 (4.14%)	
	Everyday	28,624 (18.01%)	26,971 (17.83%)	1,653 (21.52%)	
	Quit	2,060 (1.30%)	1,957 (1.29%)	103 (1.34%)	
Psychotropic drug prescriptions count (IQR)		0 (0, 0)	0 (0, 0)	0 (0, 0)	< 0.001
Aspirin prescriptions count (IQR)		0 (0, 0)	0 (0, 0)	0 (0, 0)	< 0.001
Antiinflammatory drugs prescriptions count (IQR)		2 (1, 6)	2 (1, 6)	1 (0, 4)	< 0.001
Statin drugs prescriptions count (IQR)		0 (0, 1)	0 (0, 1)	0 (0, 0)	< 0.001
Metformin prescriptions count (IQR)		0 (0, 0)	0 (0, 0)	0 (0, 0)	< 0.001
Immunosuppressant prescriptions count (IQR)		2 (0, 5)	2 (0, 5)	0 (0, 2)	< 0.001
Sex hormone prescriptions count (IQR)		0 (0, 0)	0 (0, 0)	0 (0, 0)	< 0.001
Antibiotic prescriptions count (IQR)		7 (2, 15)	7 (2, 15)	3 (0, 9)	< 0.001
ACE prescriptions count (IQR)		0 (0, 0)	0 (0, 0)	0 (0, 0)	< 0.001
PPI prescriptions count (IQR)		1 (0, 5)	2 (0, 6)	0 (0, 3)	< 0.001
CT_Multiple lung nodules or masses					0.298
	No	152,849 (96.16%)	145,447 (96.15%)	7,402 (96.38%)	
	Yes	6,108 (3.84%)	5,830 (3.85%)	278 (3.62%)	
CT_Lung consolidation					0.903
	No	158,197 (99.52%)	150,553 (99.52%)	7,644 (99.53%)	
	Yes	760 (0.48%)	724 (0.48%)	36 (0.47%)	
CT_Lung soft tissue density mass					< 0.001
	No	157,395 (99.02%)	149,911 (99.10%)	7,484 (97.45%)	
	Yes	1,562 (0.98%)	1,366 (0.90%)	196 (2.55%)	
CT_Irregular margins					< 0.001
	No	146,606 (92.23%)	139,407 (92.15%)	7,199 (93.74%)	
	Yes	12,351 (7.77%)	11,870 (7.85%)	481 (6.26%)	
CT_Spiculation sign					< 0.001

	No	158,558 (99.75%)	150,969 (99.80%)	7,589 (98.82%)	
	Yes	399 (0.25%)	308 (0.20%)	91 (1.18%)	
CT_Lobulation					< 0.001
	No	158,708 (99.84%)	151,094 (99.88%)	7,614 (99.14%)	
	Yes	249 (0.16%)	183 (0.12%)	66 (0.86%)	
CT_Pulmonary cavity					0.749
	No	157,939 (99.36%)	150,306 (99.36%)	7,633 (99.39%)	
	Yes	1,018 (0.64%)	971 (0.64%)	47 (0.61%)	
CT_Lung space-occupying lesions					< 0.001
	No	157,830 (99.29%)	150,414 (99.43%)	7,416 (96.56%)	
	Yes	1,127 (0.71%)	863 (0.57%)	264 (3.44%)	
CT_Bronchial tube wall irregularity/thickening/thickening/abnormality					0.470
	No	157,317 (98.97%)	149,710 (98.96%)	7,607 (99.05%)	
	Yes	1,640 (1.03%)	1,567 (1.04%)	73 (0.95%)	
CT_Bronchial tube wall truncation/stenosis/obstruction					< 0.001
	No	158,556 (99.75%)	150,955 (99.79%)	7,601 (98.97%)	
	Yes	401 (0.25%)	322 (0.21%)	79 (1.03%)	
CT_Air bronchogram					0.456
	No	158,737 (99.86%)	151,070 (99.86%)	7,667 (99.83%)	
	Yes	220 (0.14%)	207 (0.14%)	13 (0.17%)	
CT_Atelectasis					0.757
	No	156,928 (98.72%)	149,349 (98.73%)	7,579 (98.68%)	
	Yes	2,029 (1.28%)	1,928 (1.27%)	101 (1.32%)	
CT_Pulmonary fibrous foci					< 0.001
	No	139,416 (87.71%)	132,345 (87.49%)	7,071 (92.07%)	
	Yes	19,541 (12.29%)	18,932 (12.51%)	609 (7.93%)	
CT_Multiple pulmonary bullae					0.553
	No	156,786 (98.63%)	149,205 (98.63%)	7,581 (98.71%)	
	Yes	2,171 (1.37%)	2,072 (1.37%)	99 (1.29%)	
CT_Pleural thickening/fibrosis/adhesions					0.518
	No	133,320 (83.87%)	126,899 (83.89%)	6,421 (83.61%)	
	Yes	25,637 (16.13%)	24,378 (16.11%)	1,259 (16.39%)	
CT_Pulmonary lymph node enlargement					< 0.001
	No	156,190 (98.26%)	148,741 (98.32%)	7,449 (96.99%)	
	Yes	2,767 (1.74%)	2,536 (1.68%)	231 (3.01%)	
CT_Pleural effusion					< 0.001
	No	157,601 (99.15%)	150,117 (99.23%)	7,484 (97.45%)	
	Yes	1,356 (0.85%)	1,160 (0.77%)	196 (2.55%)	
CT_Calcification					< 0.001
	No	144,854 (91.13%)	137,687 (91.02%)	7,167 (93.32%)	
	Yes	14,103 (8.87%)	13,590 (8.98%)	513 (6.68%)	
Maximum lung nodule size (IQR)		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.028
CT scans in the past five years (IQR)		0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)	< 0.001
Duration of allergic diseases history (IQR)		0 (0, 27)	0 (0, 28)	0 (0, 0)	< 0.001
Duration of autoimmune diseases history (IQR)		0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.003

Duration of asthma history (IQR)	0 (0, 19)	0 (0, 20)	0 (0, 0)	< 0.001
Duration of diabetes history (IQR)	0 (0, 0)	0 (0, 0)	0 (0, 0)	< 0.001
Duration of hypertension history (IQR)	0 (0, 0)	0 (0, 0)	0 (0, 0)	< 0.001
Duration of hyperlipidemia history (IQR)	0 (0, 42)	0 (0, 42)	0 (0, 30)	0.003
Duration of gout history (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	< 0.001
Insomnia history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	< 0.001
Constipation history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	< 0.001
Chest discomfort history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	< 0.001
Chest pain history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	< 0.001
Hematemesis history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.063
Dyspnea history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.812
Obsolete TB				< 0.001
No	143,548 (90.31%)	136,894 (90.49%)	6,654 (86.64%)	
Yes	15,409 (9.69%)	14,383 (9.51%)	1,026 (13.36%)	
Group				0.386
Training Cohort	111,269 (70.00%)	105,927 (70.02%)	5,342 (69.56%)	
Validation Cohort	47,688 (30.00%)	45,350 (29.98%)	2,338 (30.44%)	

Note. Some percentages do not add up to 100 because of rounding. IQR: Interquartile range; CT: Under CT imaging; BMI: Body Mass Index; hs-CRP: High-sensitivity CRP; ESR: Erythrocyte sedimentation rate; HbA 1c: Glycated hemoglobin; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; TC: Triglyceride; LDL-C: Low-density lipoprotein cholesterol; TG: Total cholesterol; HDL-C: High-density lipoprotein cholesterol; ACE: Angiotensin converting enzyme; PPI: Protein-protein interactions; TB: Tuberculosis. [1] Median (IQR); n (%). [2] Wilcoxon rank sum test; Pearson's Chi-squared test.

Supplementary Table S3. Sociodemographic and clinicopathologic characteristics of training and validation cohort with lung cancer

Variable	Overall, *N* = 158,957[1]	Training Cohort, *N* = 111,269[1]	Validation Cohort, *N* = 47,688[1]	*P*-value[2]
Age (IQR)	57 (45, 67)	57 (45, 67)	57 (45, 67)	0.989
Gender				0.857
Female	79,857 (50.24%)	55,883 (50.22%)	23,974 (50.27%)	
Male	79,100 (49.76%)	55,386 (49.78%)	23,714 (49.73%)	
BMI (IQR)	22.3 (20.4, 24.4)	22.3 (20.4, 24.3)	22.3 (20.4, 24.4)	0.110
hs-CRP (IQR)	4 (1, 13)	4 (1, 13)	4 (1, 13)	0.710
ESR (IQR)	13 (6, 26)	13 (6, 26)	13 (6, 26)	0.891
HbA 1c (IQR)	5.50 (5.00, 6.00)	5.50 (5.00, 6.00)	5.50 (5.00, 6.00)	0.400
SBP (IQR)	125 (120, 132)	125 (120, 132)	125 (120, 132)	0.675
DBP (IQR)	78 (71, 80)	78 (71, 80)	78 (70, 80)	0.598
TC (IQR)	4.79 (4.16, 5.46)	4.79 (4.16, 5.46)	4.78 (4.15, 5.45)	0.133
IDL-C (IQR)	2.68 (2.19, 3.21)	2.69 (2.20, 3.21)	2.68 (2.19, 3.20)	0.169
TG (IQR)	1.26 (0.91, 1.82)	1.27 (0.91, 1.82)	1.26 (0.91, 1.81)	0.553
HDL-C (IQR)	1.34 (1.15, 1.58)	1.34 (1.15, 1.58)	1.35 (1.15, 1.58)	0.708
Uric acid (IQR)	307 (248, 376)	307 (248, 376)	307 (248, 376)	0.596
Number of psychiatric department visits (IQR)	1 (0, 3)	1 (0, 3)	1 (0, 3)	0.799
Smoke				0.066
Never	124,496 (78.32%)	87,097 (78.28%)	37,399 (78.42%)	
Seldom	3,777 (2.38%)	2,613 (2.35%)	1,164 (2.44%)	
Everyday	28,624 (18.01%)	20,156 (18.11%)	8,468 (17.76%)	
Quit	2,060 (1.30%)	1,403 (1.26%)	657 (1.38%)	
Psychotropic drug prescriptions count (IQR)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0.369
Aspirin prescriptions count (IQR)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0.025
Antiinflammatory drugs prescriptions count (IQR)	2 (1, 6)	2 (1, 6)	2 (1, 6)	0.104
Statin drugs prescriptions count (IQR)	0 (0, 1)	0 (0, 1)	0 (0, 1)	0.169
Metformin prescriptions count (IQR)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0.426
Immunosuppressant prescriptions count (IQR)	2 (0, 5)	2 (0, 5)	2 (0, 5)	0.624
Sex hormone prescriptions count (IQR)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0.939
Antibiotic prescriptions count (IQR)	7 (2, 15)	7 (2, 15)	7 (2, 15)	0.448
ACE prescriptions count (IQR)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0.561
PPI prescriptions count (IQR)	1 (0, 5)	1 (0, 5)	1 (0, 5)	0.294
CT_Multiple lung nodules or masses				0.660
No	152,849 (96.16%)	106,978 (96.14%)	45,871 (96.19%)	
Yes	6,108 (3.84%)	4,291 (3.86%)	1,817 (3.81%)	
CT_Lung consolidation				0.383

No	158,197 (99.52%)	110,748 (99.53%)	47,449 (99.50%)	
Yes	760 (0.48%)	521 (0.47%)	239 (0.50%)	
CT_Lung soft tissue density mass				0.210
No	157,395 (99.02%)	110,153 (99.00%)	47,242 (99.06%)	
Yes	1,562 (0.98%)	1,116 (1.00%)	446 (0.94%)	
CT_Irregular margins				0.876
No	146,606 (92.23%)	102,631 (92.24%)	43,975 (92.21%)	
Yes	12,351 (7.77%)	8,638 (7.76%)	3,713 (7.79%)	
CT_Spiculation sign				0.533
No	158,558 (99.75%)	110,984 (99.74%)	47,574 (99.76%)	
Yes	399 (0.25%)	285 (0.26%)	114 (0.24%)	
CT_Lobulation				0.967
No	158,708 (99.84%)	111,095 (99.84%)	47,613 (99.84%)	
Yes	249 (0.16%)	174 (0.16%)	75 (0.16%)	
CT_Pulmonary cavity				0.555
No	157,939 (99.36%)	110,565 (99.37%)	47,374 (99.34%)	
Yes	1,018 (0.64%)	704 (0.63%)	314 (0.66%)	
CT_Lung space-occupying lesions				0.789
No	157,830 (99.29%)	110,476 (99.29%)	47,354 (99.30%)	
Yes	1,127 (0.71%)	793 (0.71%)	334 (0.70%)	
CT_Bronchial tube wall irregularity/thickening/thickening/abnormality				0.255
No	157,317 (98.97%)	110,100 (98.95%)	47,217 (99.01%)	
Yes	1,640 (1.03%)	1,169 (1.05%)	471 (0.99%)	
CT_Bronchial tube wall truncation/stenosis/obstruction				0.768
No	158,556 (99.75%)	110,991 (99.75%)	47,565 (99.74%)	
Yes	401 (0.25%)	278 (0.25%)	123 (0.26%)	
CT_Air bronchogram				0.462
No	158,737 (99.86%)	111,120 (99.87%)	47,617 (99.85%)	
Yes	220 (0.14%)	149 (0.13%)	71 (0.15%)	
CT_Atelectasis				0.933
No	156,928 (98.72%)	109,847 (98.72%)	47,081 (98.73%)	
Yes	2,029 (1.28%)	1,422 (1.28%)	607 (1.27%)	
CT_Pulmonary fibrous foci				0.337
No	139,416 (87.71%)	97,648 (87.76%)	41,768 (87.59%)	
Yes	19,541 (12.29%)	13,621 (12.24%)	5,920 (12.41%)	
CT_Multiple pulmonary bullae				0.839
No	156,786 (98.63%)	109,745 (98.63%)	47,041 (98.64%)	
Yes	2,171 (1.37%)	1,524 (1.37%)	647 (1.36%)	
CT_Pleural thickening/fibrosis/adhesions				0.192
No	133,320 (83.87%)	93,411 (83.95%)	39,909 (83.69%)	
Yes	25,637 (16.13%)	17,858 (16.05%)	7,779 (16.31%)	
CT_Pulmonary lymph node enlargement				0.111

No	156,190 (98.26%)	109,294 (98.23%)	46,896 (98.34%)	
Yes	2,767 (1.74%)	1,975 (1.77%)	792 (1.66%)	
CT_Pleural effusion				0.600
No	157,601 (99.15%)	110,311 (99.14%)	47,290 (99.17%)	
Yes	1,356 (0.85%)	958 (0.86%)	398 (0.83%)	
CT_Calcification				0.878
No	144,854 (91.13%)	101,389 (91.12%)	43,465 (91.14%)	
Yes	14,103 (8.87%)	9,880 (8.88%)	4,223 (8.86%)	
Maximum lung nodule size (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.547
CT scans in the past five years (IQR)	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.00 (0.00, 1.00)	0.879
Duration of allergic diseases history (IQR)	0.00 (0.00, 27.00)	0.00 (0.00, 27.00)	0.00 (0.00, 27.00)	0.602
Duration of autoimmune diseases history (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.302
Duration of asthma history (IQR)	0.00 (0.00, 19.00)	0.00 (0.00, 19.00)	0.00 (0.00, 19.00)	0.969
Duration of diabetes history (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.428
Duration of hypertension history (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.509
Duration of hyperlipidemia history (IQR)	0 (0, 42)	0 (0, 42)	0 (0, 42)	0.607
Duration of gout history (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.553
Insomnia history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.152
Constipation history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.390
Chest discomfort history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.849
Chest pain history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.814
Hematemesis history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.241
Dyspnea history in the past 5 years (IQR)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.938
Malignant				0.386
Non-lung cancer	151,277 (95.17%)	105,927 (95.20%)	45,350 (95.10%)	
Lung cancer	7,680 (4.83%)	5,342 (4.80%)	2,338 (4.90%)	
Obsolete TB				0.138
No	143,548 (90.31%)	100,563 (90.38%)	42,985 (90.14%)	
Yes	15,409 (9.69%)	10,706 (9.62%)	4,703 (9.86%)	

Note. Some percentages do not add up to 100 because of rounding. [1] Median (IQR); n (%). [2] Wilcoxon rank sum test; Pearson's Chi-squared test.

Supplementary Table S4. The univariate logistic regression analysis

Characteristics	*OR* (95% *CI*)	*P*-value
Age	1.0208(1.019–1.0225)	< 0.001
Gender		
Female	Ref	
Male	1.5703(1.4845–1.6609)	< 0.001
BMI	1.0046(0.996–1.0133)	0.2968
hs-CRP	1.0032(1.0022–1.0041)	< 0.001
ESR	1.0037(1.0025–1.0049)	< 0.001
HbA 1c	0.9875(0.9761–0.9991)	0.0347
SBP	1.0031(1.0011–1.0051)	0.0024
DBP	1.0033(1.0004–1.0062)	0.0263
TC	1.007(0.9803–1.0345)	0.6105
LDL-C	1.0401(1.0054–1.0759)	0.0230
TG	0.9795(0.9525–1.0073)	0.1470
HDL-C	0.8954(0.8277–0.9685)	0.0058
Uric acid	1.0002(0.9999–1.0004)	0.2315
Number of psychiatric department visits	0.9955(0.9922–0.9987)	0.0067
Smoke		
Never	Ref	
Seldom	1.9199(1.6633–2.2162)	< 0.001
Everyday	1.3504(1.2631–1.4438)	< 0.001
Quit	1.1473(0.9017–1.4597)	0.2637
Psychotropic drug prescriptions count	0.9919(0.9873–0.9965)	0.0006
Aspirin prescriptions count	0.979(0.9705–0.9876)	< 0.001
Antiinflammatory drugs prescriptions count	0.9639(0.9589–0.969)	< 0.001
Statin drugs prescriptions count	0.9831(0.9782–0.988)	< 0.001
Metformin prescriptions count	0.9939(0.9893–0.9986)	0.0102
Immunosuppressant prescriptions count	0.9336(0.9262–0.941)	< 0.001
Sex hormone prescriptions count	0.8784(0.8373–0.9215)	< 0.001
Antibiotic prescriptions count	0.963(0.9598–0.9661)	< 0.001
ACE prescriptions count	0.9949(0.9887–1.0012)	0.1149
PPI prescriptions count	0.9732(0.9691–0.9773)	< 0.001
CT_Multiple lung nodules or masses		
No	Ref	
Yes	0.8636(0.742–1.0052)	0.0584
CT_Lung consolidation		
No	Ref	
Yes	0.8322(0.5374–1.2887)	0.4103
CT_Lung soft tissue density mass		
No	Ref	
Yes	2.8456(2.376–3.4082)	< 0.001
CT_Irregular margins		
No	Ref	
Yes	0.7762(0.693–0.8695)	< 0.001

CT_Spiculation sign		
No	Ref	
Yes	5.566(4.1951–7.3849)	< 0.001
CT_Lobulation		
No	Ref	
Yes	7.6134(5.4538–10.6281)	< 0.001
CT_Pulmonary cavity		
No	Ref	
Yes	0.7898(0.5371–1.1612)	0.2301
CT_Lung space-occupying lesions		
No	Ref	
Yes	6.0796(5.1401–7.1907)	< 0.001
CT_Bronchial tube wall irregularity/thickening/thickening/abnormality		
No	Ref	
Yes	0.8106(0.6029–1.0899)	0.1645
CT_Bronchial tube wall truncation/stenosis/obstruction		
No	Ref	
Yes	4.2729(3.1348–5.8243)	< 0.001
CT_Air bronchogram		
No	Ref	
Yes	1.2752(0.6496–2.5034)	0.4800
CT_Atelectasis		
No	Ref	
Yes	0.9492(0.739–1.2192)	0.6831
CT_Pulmonary fibrous foci		
No	Ref	
Yes	0.6042(0.5462–0.6684)	< 0.001
CT_Multiple pulmonary bullae		
No	Ref	
Yes	1.0267(0.8127–1.2971)	0.825
CT_Pleural thickening/fibrosis/adhesions		
No	Ref	
Yes	0.9951(0.9232–1.0726)	0.8979
CT_Pulmonary lymph node enlargement		
No	Ref	
Yes	1.8709(1.5935–2.1967)	< 0.001
CT_Pleural effusion		
No	Ref	
Yes	3.5176(2.937–4.2129)	< 0.001
CT_Calcification		
No	Ref	
Yes	0.6957(0.6225–0.7776)	< 0.001
Maximum lung nodule size	1.0097(1.0068–1.0126)	< 0.001
CT scans in the past five years	0.7343(0.7077–0.7619)	< 0.001

Duration of allergic diseases history	0.9846(0.9831–0.986)	< 0.001
Duration of autoimmune diseases history	0.9784(0.962–0.995)	0.0108
Duration of asthma history	0.9743(0.9723–0.9763)	< 0.001
Duration of diabetes history	0.9967(0.9954–0.998)	< 0.001
Duration of hypertension history	0.9981(0.9971–0.9991)	0.0001
Duration of hyperlipidemia history	0.9951(0.994–0.9961)	< 0.001
Duration of gout history	0.9948(0.9927–0.9969)	< 0.001
Insomnia history in the past 5 years	0.992(0.9906–0.9934)	< 0.001
Constipation history in the past 5 years	0.9901(0.9881–0.992)	< 0.001
Chest discomfort history in the past 5 years	0.9915(0.9893–0.9938)	< 0.001
Chest pain history in the past 5 years	0.9851(0.982–0.9883)	< 0.001
Hematemesis history in the past 5 years	0.9975(0.9926–1.0024)	0.3113
Dyspnea history in the past 5 years	0.9924(0.9471–1.0398)	0.7485
Obsolete TB		
No	Ref	
Yes	1.4685(1.3533–1.5935)	< 0.001

Supplementary Table S5. Multivariable logistic regression model of study variables in the training dataset

Characteristics	***OR* (95% *CI*)**	***P*-value**
Age	1.034(1.032–1.036)	< 0.001
Gender		
Female	Ref	
Male	1.340(1.256–1.429)	< 0.001
Smoke		
Never	Ref	
Seldom	1.412(1.212–1.645)	< 0.001
Everyday	1.097(1.017–1.184)	0.016
Quit	0.993(0.773–1.275)	0.956
Antiinflammatory drugs prescriptions count	0.988(0.983–0.994)	< 0.001
Antibiotic prescriptions count	0.985(0.981–0.988)	< 0.001
CT scans in the past five years	0.634(0.608–0.660)	< 0.001
Duration of allergic diseases history	0.992(0.990–0.993)	< 0.001
Duration of asthma history	0.972(0.970–0.974)	< 0.001
Duration of diabetes history	0.997(0.995–0.998)	< 0.001
Duration of hypertension history	0.997(0.996–0.998)	< 0.001
Duration of hyperlipidemia history	0.996(0.995–0.997)	< 0.001
Insomnia history in the past 5 years	0.998(0.996–0.999)	0.005
Constipation history in the past 5 years	0.995(0.993–0.997)	< 0.001
Chest pain history in the past 5 years	0.991(0.988–0.994)	< 0.001
CT_Lung soft tissue density mass		
No	Ref	
Yes	1.506(1.201–1.889)	< 0.001
CT_Irregular margins		
No	Ref	
Yes	0.686(0.607–0.775)	< 0.001
CT_Spiculation sign		
No	Ref	
Yes	2.931(2.047–4.199)	< 0.001
CT_Lobulation		
No	Ref	
Yes	2.209(1.417–3.445)	< 0.001
CT_Lung space-occupying lesions		
No	Ref	
Yes	3.886(3.085–4.895)	< 0.001
CT_Bronchial tube wall truncation/stenosis/obstruction		
No	Ref	
Yes	2.725(1.870–3.970)	< 0.001
CT_Pulmonary fibrous foci		
No	Ref	
Yes	0.608(0.548–0.674)	< 0.001
CT_Pleural effusion		
No	Ref	

Yes	1.782(1.412–2.248)	< 0.001
CT_Calcification		
No	Ref	
Yes	0.584(0.520–0.656)	< 0.001
Obsolete TB		
No	Ref	
Yes	1.228(1.125–1.340)	< 0.001

.

Supplementary Table S6. Performance of the machine learning models for lung cancer risk prediction in high-risk population

	Algorithm	Sensitivity	Specificity	Accuracy	AUROC
Training Cohort	**DT**	0.346 (0.334, 0.359)	0.968 (0.967, 0.969)	0.939 (0.937, 0.940)	0.747(0.741–0.753)
	LightGBM	0.749 (0.737, 0.760)	0.816 (0.814, 0.818)	0.813 (0.810, 0.815)	0.864(0.859–0.869)
	XGBoost	**0.808 (0.797, 0.818)**	0.782 (0.780, 0.785)	0.784 (0.781, 0.786)	**0.880(0.875–0.884)**
	Logistic	0.671 (0.658, 0.684)	0.763 (0.761, 0.766)	0.759 (0.756, 0.761)	0.761(0.754–0.769)
Validation Cohort	**DT**	0.345 (0.326, 0.365)	0.968 (0.966, 0.969)	0.937 (0.935, 0.939)	0.735(0.725–0.745)
	LightGBM	0.754 (0.736, 0.771)	0.798 (0.795, 0.802)	0.796 (0.792, 0.800)	0.853(0.845–0.861)
	XGBoost	**0.790 (0.773, 0.807)**	0.776 (0.772, 0.780)	0.777 (0.773, 0.781)	**0.860(0.852–0.868)**
	Logistic	0.666 (0.647, 0.685)	0.756 (0.752, 0.760)	0.752 (0.748, 0.756)	0.755(0.743–0.766)

Note. DT, decision tree.