OWL: an optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations

Zoucheng Pan,[a,f] Ruyang Zhang,[a,f] Sipeng Shen,[a] Yunzhi Lin,[a] Longyao Zhang,[a] Xiang Wang,[a] Qian Ye,[a] Xuan Wang,[a] Jiajin Chen,[a] Yang Zhao,[a] David C. Christiani,[b,c] Yi Li,[d] Feng Chen,[a,**] and Yongyue Wei[a,e,*]

[a]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, 211166, China
[b]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA
[c]Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, 02114, USA
[d]Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, USA
[e]Peking University Center for Public Health and Epidemic Preparedness & Response, Xueyuan Road, Haidian District, Beijing 100191, China

eBioMedicine
2023;88: 104443

Published Online xxx
https://doi.org/10.1016/j.ebiom.2023.104443

Summary

Background A reliable risk prediction model is critically important for identifying individuals with high risk of developing lung cancer as candidates for low-dose chest computed tomography (LDCT) screening. Leveraging a cutting-edge machine learning technique that accommodates a wide list of questionnaire-based predictors, we sought to optimize and validate a lung cancer prediction model.

Methods We developed an Optimized early Warning model for Lung cancer risk (OWL) using the XGBoost algorithm with 323,344 participants from the England area in UK Biobank (training set), and independently validated it with 93,227 participants from UKB Scotland and Wales area (validation set 1), as well as 70,605 and 66,231 participants in the Prostate, Lung, Colorectal, and Ovarian cancer screening trial (PLCO) control and intervention subpopulations, respectively (validation sets 2 & 3) and 23,138 and 18,669 participants in the United States National Lung Screening Trial (NLST) control and intervention subpopulations, respectively (validation sets 4 & 5). By comparing with three competitive prediction models, i.e., PLCO modified 2012 ($\text{PLCO}_{\text{m2012}}$), PLCO modified 2014 ($\text{PLCO}_{\text{all2014}}$), and the Liverpool Lung cancer Project risk model version 3 (LLPv3), we assessed the discrimination of OWL by the area under receiver operating characteristic curve (AUC) at the designed time point. We further evaluated the calibration using relative improvement in the ratio of expected to observed lung cancer cases (RI_{EO}), and illustrated the clinical utility by the decision curve analysis.

Findings For general population, with validation set 1, OWL (AUC = 0.855, 95% CI: 0.829–0.880) presented a better discriminative capability than $\text{PLCO}_{\text{all2014}}$ (AUC = 0.821, 95% CI: 0.794–0.848) ($p < 0.001$); with validation sets 2 & 3, AUC of OWL was comparable to $\text{PLCO}_{\text{all2014}}$ ($\text{AUC}_{\text{PLCOall2014}}$-$\text{AUC}_{\text{OWL}} < 1\%$). For ever-smokers, OWL outperformed $\text{PLCO}_{\text{m2012}}$ and $\text{PLCO}_{\text{all2014}}$ among ever-smokers in validation set 1 ($\text{AUC}_{\text{OWL}} = 0.842$, 95% CI: 0.814–0.871; $\text{AUC}_{\text{PLCOm2012}} = 0.792$, 95% CI: 0.760–0.823; $\text{AUC}_{\text{PLCOall2014}} = 0.791$, 95% CI: 0.760–0.822, all $p < 0.001$). OWL remained comparable to $\text{PLCO}_{\text{m2012}}$ and $\text{PLCO}_{\text{all2014}}$ in discrimination (AUC difference from −0.014 to 0.008) among the ever-smokers in validation sets 2 to 5. In all the validation sets, OWL outperformed LLPv3 among the general population and the ever-smokers. Of note, OWL showed significantly better calibration than $\text{PLCO}_{\text{m2012}}$, $\text{PLCO}_{\text{all2014}}$ (RI_{EO} from 43.1% to 92.3%, all $p < 0.001$), and LLPv3 (RI_{EO} from 41.4% to 98.7%, all $p < 0.001$) in most cases. For clinical utility, OWL exhibited significant improvement in average net benefits (*NB*) over $\text{PLCO}_{\text{all2014}}$ in validation set 1 (*NB* improvement: 32, $p < 0.001$); among ever smokers of validation set 1, OWL (average *NB* = 289) retained significant improvement over $\text{PLCO}_{\text{m2012}}$ (average *NB* = 213) ($p < 0.001$). OWL had equivalent *NB*s with $\text{PLCO}_{\text{m2012}}$ and $\text{PLCO}_{\text{all2014}}$ in PLCO and NLST populations, while outperforming LLPv3 in the three populations.

Interpretation OWL, with a high degree of predictive accuracy and robustness, is a general framework with scientific justifications and clinical utility that can aid in screening individuals with high risks of lung cancer.

*Corresponding author. Medical Science & Technology Building West 601, 38 Xueyuan Road, Haidian District, Beijing, 100191, China.
**Corresponding author. SPH Building Room 412, 101 Longmian Avenue, Nanjing, Jiangsu, 211166, China.
E-mail addresses: ywei@pku.edu.cn (Y. Wei), fengchen@njmu.edu.cn (F. Chen).
[f]These authors contributed equally to this work.

Funding National Natural Science Foundation of China, the US NIH.

Keywords: Lung cancer; Risk prediction; UK Biobank; External validation; Machine learning

Research in context

Evidence before this study

We systematically searched PubMed and Web of Science for research articles published in English before Dec 31, 2021, with the search terms “lung cancer”, “risk”, “prediction” and “model”. Among 8,257 preliminarily retrieved studies, 20 studies were eligible for this study and were included; in which, 20 lung cancer risk prediction models were identified that have been developed and some of them achieved acceptable performance, suggesting the usefulness of model-based screening strategies as an alternative of or supplement to criteria-based screening strategies. Besides, an expert panel recommended combination of criteria- and qualified risk prediction model-based strategies, to maximize the opportunity of identifying high risk individuals for low-dose chest computed tomography (LDCT). To better identify target populations for LDCT screening, we address the urgent need of developing a lung cancer risk prediction model that presents a high degree of accuracy and robustness.

Added value of this study

Through a systematic review, 15 well-established questionnaire-based risk factors were included in the model. We constructed an **O**ptimized early **W**arning model for **L**ung cancer risk (OWL) using the eXtreme Gradient Boosting (XGBoost) method based on the participants from the England area in UKB, and independently validated it with the participants from the Scotland and Wales areas of UKB, the Prostate, Lung, Colorectal and Ovarian cancer screening trial (PLCO), and the United States National Lung Screening Trial (NLST). The OWL model exhibited equivalent accuracy and robustness in lung cancer risk prediction to the PLCO modified 2012 model ($\text{PLCO}_{\text{m2012}}$) among the ever-smokers in the PLCO and NLST populations, while presenting better discrimination than $\text{PLCO}_{\text{m2012}}$, PLCO modified 2014 model ($\text{PLCO}_{\text{all2014}}$), and Liverpool lung project risk score version 3 model (LLPv3) among the Scotland and Wales populations in UKB (UKB external validation set). Notably, OWL showed slightly better calibration and improved the absolute lung cancer risk prediction than all competitive models. Moreover, OWL had higher clinical utility than all competitive models among validation set 1, meanwhile showing comparable clinical utility to $\text{PLCO}_{\text{m2012}}$ among the ever-smokers in PLCO and NLST populations.

Implications of all the available evidence

OWL is a reliable lung cancer risk prediction model with sufficient accuracy and robustness, which provides a feasible mean of identifying individuals at high risk of lung cancer as candidates for further LDCT screening.

Introduction

Lung cancer is the leading cause of cancer-related mortality, with an estimated over 2.2 million newly diagnosed cases and 1.8 million deaths in 2020 alone worldwide.[1] Its age-standardized five-year net survival is low, ranging from 10% to 20% in most countries.[2] Screening individuals with high risks of lung cancer has emerged as an effective means to reduce cancer morbidity and mortality by way of detecting early-stage cases or those predisposed to lung cancer, leading to more effective treatment and intervention strategies and, therefore, increased overall survival.

The United States (US) National Lung Screening Trial (NLST) showed that low-dose chest computed tomography (LDCT) screening reduced lung cancer-related mortality by 20%, and all-cause mortality by 6.7%.[3] Hence, LDCT screening for lung cancer has become a standard of care in the US,[4] with its benefits verified by various randomized trials.[5,6] Identifying suitable subpopulation at high risk of lung cancer for LDCT screening has become essential to maximize the cost-effectiveness of screening programs.[7,8] The Centers for Medicare and Medicaid Services (CMS) and the US Preventive Services Task Force (USPSTF) have recommended offering annual LDCT screening to asymptomatic individuals meeting these criteria based on age and smoking (Supplementary Fig. S1A and B).[9] These criteria, though simple and useful, may miss opportunities to capture more individuals with high risk.[10,11]

Substantial efforts have been devoted to developing and validating lung cancer risk prediction models (Supplementary Table S1).[12] In particular, the Prostate, Lung, Colorectal and Ovarian cancer screening trial (PLCO) modified 2012 ($\text{PLCO}_{\text{m2012}}$), the PLCO modified 2014 ($\text{PLCO}_{\text{all2014}}$), and the Liverpool lung cancer project risk model version 3 (LLPv3) have exhibited a considerable discriminative ability,[13–16] suggesting the usefulness of model-based screening strategies as an alternative of criteria-based strategies. An expert panel recommended combinations of criteria- and model-based strategies,[17] and further applications of verified risk prediction models to criteria-negative (CN) subpopulations (Supplementary Fig. S1C) in order to maximize the opportunity of identifying high-risk individuals from population suitable for LDCT screening.[10] This motivated an urgent need of developing robust and accurate prediction models, which benefit from incorporating reliable and relevant risk factors as well as effective modeling techniques, such as eXtreme Gradient Boosting (XGBoost). XGBoost has surfaced as a powerful machine learning algorithm that can capture non-linear and interaction effects of predictors.[18,19]

This study was to develop an **O**ptimized early **W**arning model for **L**ung cancer risk (OWL) by using XGBoost technique. To identity the reliable predictors as many as possible, we focused on the prior well-established lung cancer risk prediction models[20] and the lung cancer risk factors recently identified by causal inference,[21] and obtained a broad list of reliable predictors. We then used XGBoost to develop model using the participants from the England area in the UK Biobank (UKB) and externally validated it using the populations from the UKB Scotland and Wales areas, PLCO, and NLST.

Methods

Study populations

UKB population for model development and validation

UKB (https://www.ukbiobank.ac.uk/) is a large-scale prospective cohort of over 500,000 participants aged between 37 and 73 years at the time of recruitment between 2006 and 2010. Follow up for cancer incidence and death was conducted via cancer and death registries; participants were followed up from the date of baseline attendance until the date of diagnosis of an invasive primary lung cancer, death, or loss of follow up, whichever occurred first. The data was approved by UKB under the approval number of 57471, and the data extracted time is 2020-08-04. Inclusion criteria of this study were: (i) eligible for the original study; (ii) not diagnosed of lung cancer before participation; (ii) not diagnosed of other cancers before participation. A total of 416,671 participants were eligible for the study, of which 323,344 participants from England were used to develop the model (training set), and 93,227 participants from Scotland and Wales were used to externally validate the model (validation set 1) (Supplementary Fig. S2A).

PLCO and NLST populations for model validation

PLCO (approval number: PLCO-731) and NLST (approval number: NLST-755) are two large-scale population screening trials. PLCO recruited approximately 155,000 participants aged 55–74 years between 1993 and 2001. NLST recruited 53,452 participants aged 55–74 years with at least 30 pack-years of smoking and no more than 15 years since smoking cessation. Inclusion criteria of the study were: (i) eligible for the original study; (ii) not diagnosed of lung cancer before participation; (ii) not diagnosed of other cancers before participation; (iv) no lung nodules screened at baseline. A total of 136,836 and 41,807 participants in PLCO and NLST, respectively, were eligible for analysis (Supplementary Fig. S2B and C).

Statistical analysis

We applied a three-steps analytical strategy, i.e., model optimization and development, model evaluation and comparison, and model improvement; see the workflow in Fig. 1.

Step 1: model optimization and development

Inclusion of risk factors. A systematic review, following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Supplementary Fig. S3), up to December 31st, 2021 revealed a total of 20 lung cancer risk prediction models and 65 risk factors (Supplementary Tables S1 and S2).[22] The Prediction model **R**isk **O**f **B**ias (ROB) Assessment Tool was used to assess model quality, or the potential for inflated model performance estimates due to shortcomings in design/analysis.[23] This tool evaluated the levels of ROB in five domains: participant selection, definition and measurement of predictors, definition and measurement of outcome, sample size and participant flow, and statistical analysis. Each domain was rated as high, low, or unclear ROB. Overall judgement of risk of bias was derived from the judgement on all domains: low risk if all domains had low risk of bias, high risk if any domain had high risk of bias, otherwise unclear risk. The results of quality assessment were detailed in Supplementary Table S3. In addition, genetic predispositions were considered by using a polygenic risk score (PRS) constructed among European ancestry population.[24,25] Among the 65 risk factors, 15 questionnaire-based risk factors were available in UKB, PLCO, and NLST, and were used to develop OWL (Supplementary Table S2). Missing values are handled by XGBoost automatically by classifying the instance into default direction when the feature needed is missing (Supplementary Method Section).[18]

Optimization in prediction algorithm and validation. We applied the XGBoost algorithm by using the R package *xgboost*. XGBoost is a fast learning framework that uses gradient boosted decision trees to optimize the loss function.[26] The optimal tuning parameters needed for OWL were derived from the training set (UKB England area) by using a grid-search algorithm with 5-fold cross-validation (Supplementary Table S4).[27] The Shapley Additive explanation (SHAP) value was utilized to explain the XGBoost model.[28] Internal validation of model discrimination was evaluated by the concordance index (C-index) via 5-fold cross-validation. Given the possible heterogeneity

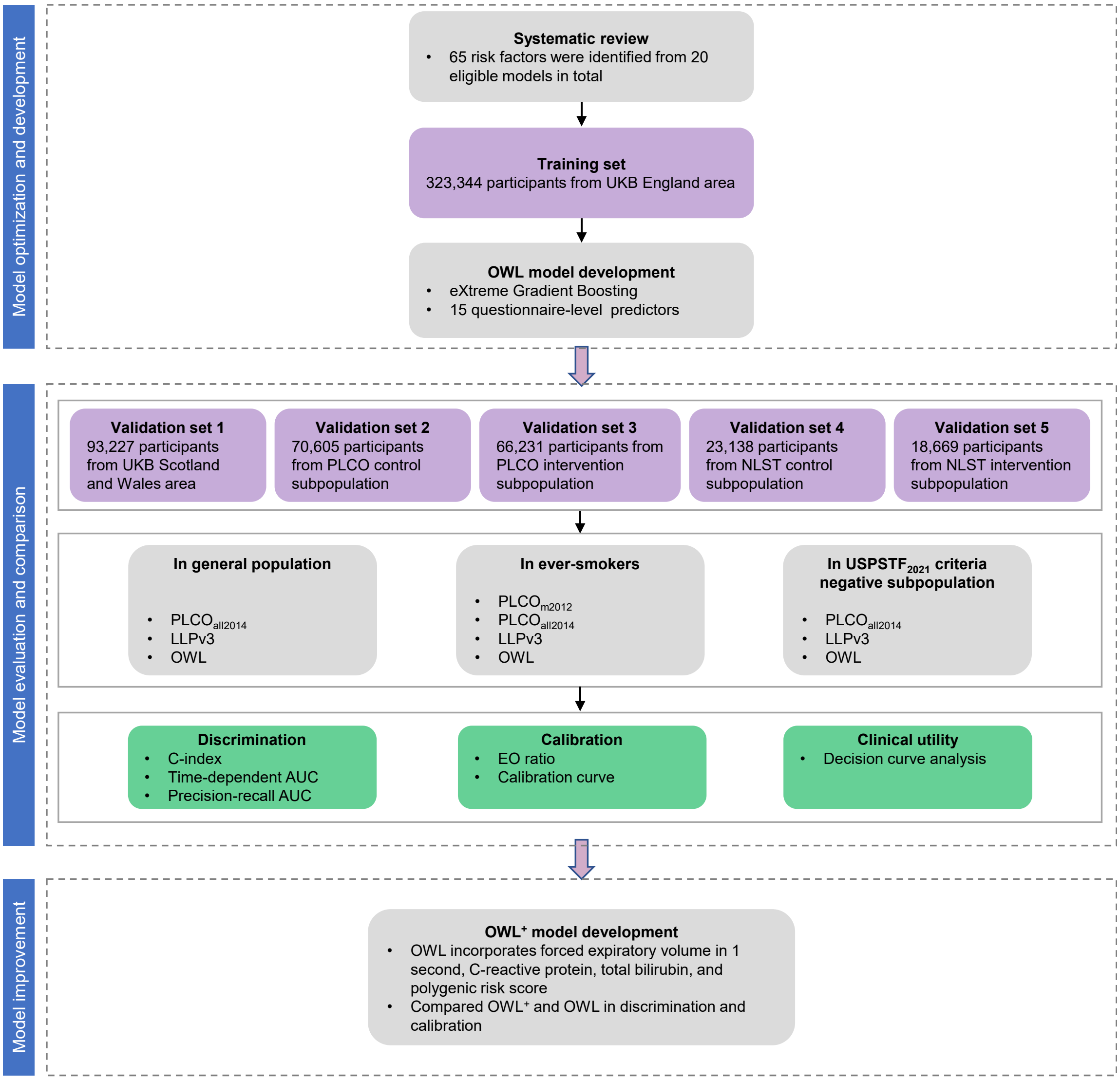
Model optimization and development
Systematic review
• 65 risk factors were identified from 20 eligible models in total
Training set
323,344 participants from UKB England area
OWL model development
• eXtreme Gradient Boosting
• 15 questionnaire-level predictors
Model evaluation and comparison
Validation set 1
93,227 participants from UKB Scotland and Wales area
Validation set 2
70,605 participants from PLCO control subpopulation
Validation set 3
66,231 participants from PLCO intervention subpopulation
Validation set 4
23,138 participants from NLST control subpopulation
Validation set 5
18,669 participants from NLST intervention subpopulation
In general population
• $PLCO_{all2014}$
• LLPv3
• OWL
In ever-smokers
• $PLCO_{m2012}$
• $PLCO_{all2014}$
• LLPv3
• OWL
In $USPSTF_{2021}$ criteria negative subpopulation
• $PLCO_{all2014}$
• LLPv3
• OWL
Discrimination
• C-index
• Time-dependent AUC
• Precision-recall AUC
Calibration
• EO ratio
• Calibration curve
Clinical utility
• Decision curve analysis
Model improvement
OWL^+ model development
• OWL incorporates forced expiratory volume in 1 second, C-reactive protein, total bilirubin, and polygenic risk score
• Compared OWL^+ and OWL in discrimination and calibration

***Fig. 1:* Study design and workflow.** UKB = UK Biobank; PLCO = the Prostate, Lung, Colorectal and Ovarian cancer screening trial; NLST = the National Lung Screening Trial; $PLCO_{m2012}$ = the PLCO modified 2012; $PLCO_{all2014}$ = the PLCO modified 2014; OWL = the Optimized early Warning model for Lung cancer risk; LLPv3 = the Liverpool lung cancer project risk model version 3; C-index = Concordance index; Time-dependent AUC = Area under receiver operating characteristic curve; Precision-recall AUC = Area under Precision-recall curve; EO ratio = Expected to observed lung cancer case ratio.

between groups, the PLCO and NLST overall populations were further divided into control (without lung screening) and intervention (with lung screening) subpopulations, respectively.[29] In summary, OWL was independently validated in validation set 1 (UKB Scotland and Wales area), validation sets 2 & 3 (PLCO control and intervention subpopulations), and validation sets 4 & 5 (NLST control and intervention subpopulations).

Estimate absolute risk and threshold for screening. As the output of XGBoost model have the interpretation of relative risks, we estimated the baseline hazard of lung cancer incidence and further calculated absolute risk, i.e., the probability of lung cancer (Supplementary Method Section). Because most participants in the training set were censored after 8-year of follow-up (Supplementary Fig. S4), we calculated the absolute risk of lung cancer for each individual within 8-year follow-up post recruitment and established various thresholds for screening (Supplementary Method Section).[13,30]

Step 2: model evaluation and comparison

OWL was developed for the general population, while $PLCO_{m2012}$ and $PLCO_{all2014}$, with similar performance, were respectively designed to predict 6-year lung cancer risks for ever-smokers and general population,[13,16] and LLPv3 was designed to predict 5-year lung cancer risks for the general population.[15] As such, among the general population in validation sets 1 to 3, we first compared OWL with $\text{PLCO}_{\text{all2014}}$ at 6-year risk prediction and LLPv3 at 5-year risk prediction. Next, we compared OWL with $\text{PLCO}_{\text{m2012}}$ and $\text{PLCO}_{\text{all2014}}$ at 6-year risk prediction, and with LLPv3 at 5-year risk prediction among ever-smokers in validation sets 1 to 5. As both $\text{PLCO}_{\text{m2012}}$ and $\text{PLCO}_{\text{all2014}}$ were developed in validation set 2 and a comparison with them in this dataset may not be fair, we elected not to compare OWL with them in this set. Different models were compared in discrimination, calibration, and clinical utility, as described below.

Discrimination. Model discriminative ability was measured by the time-dependent area under receiver operating characteristic curve (AUC) and area under precision–recall curve (PRAUC).[31–34] The AUCs were compared using permutation test with 2000 permutations.[35–37] Moreover, C-index was also used to assess the discriminative power of OWL. We further divided participants based on the quantiles of the OWL scores and plotted Kaplan–Meier (KM) curves for each group. The separation of these curves was formally tested by the log-rank test.

Calibration. Model calibration was assessed by the ratio of the expected to the observed lung cancer cases (EO ratio),[29,38,39] and by plotting the mean predicted absolute risk against the mean observed lung cancer risk within each risk group (calibration curve). In addition, we constructed the relative improvement in EO ratio (RI_{EO}) to evaluate the relative improvement of calibration (Supplementary Method Section).

Clinical utility. The clinical utility of prediction models was evaluated by decision curve analysis (DCA),[40–42] and quantitatively assessed using the average net benefit (*NB*) derived from DCA (Supplementary Method Section).

Step 3: model improvement

To evaluate the contribution of physiological [forced expiratory volume in 1 s (FEV1)], laboratory [C-reactive protein (CRP), total bilirubin], and genetic (PRS) indicators for lung cancer risk prediction, we constructed a PLG score, which integrated the **P**hysiological, **L**aboratory, and **G**enetic indicators, and assessed the interplay of the PLG score and OWL score in impacting the risk of lung cancer (Supplementary Method Section). Further, we improved the OWL model by adding FEV1, CRP, total bilirubin, and PRS. The new model is termed the OWL^{+} model.

All the statistical analyses were performed using R version 3.6.3 (The R Foundation for Statistical Computing, Vienna, Austria). A two-sided $p < 0.05$ was considered statistically significant, unless otherwise stated.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all of the data and the final responsibility to submit for publication.

Results

Participant characteristics

The UKB population had a median of 10.7 years of follow-up. In the training set (UKB England area), 1,655 of 323,344 were newly diagnosed with lung cancer after recruitment; in validation set 1 (UKB Scotland and Wales area), there were 564 out of 93,227 newly diagnosed with lung cancer. In PLCO population, 3,042 out of 136,836 were new lung cancers during follow-up (median follow-up: 12.1 years), and in NLST, 1,132 out of 41,807 were new lung cancers during follow-up (median follow-up: 6.6 years). Demographic and clinical characteristics of these populations were detailed in Table 1 and Supplementary Table S5; the participants in UKB were younger than the PLCO and NLST participants (mean age: 56.11 ± 8.13 in UKB, 62.55 ± 5.35 in PLCO, 61.22 ± 4.96 in NLST; $p_{\text{anova}} < 0.001$); the UKB population had fewer ever-smokers (44.6% in UKB vs. 53.9% PLCO, $p < 0.001$) and lower cumulative smoking intensity (smoking pack-years among ever-smokers: 23.30 ± 18.62 in UKB vs. 35.79 ± 29.23 in PLCO, $p < 0.001$) than the PLCO population; all of the participants in NLST were heavily smokers and had higher smoking pack-years than those in PLCO (56.06 ± 23.85 in NLST, $p_{\text{NLST vs. PLCO}} < 0.001$).

Development of OWL

As 15 questionnaire-based and well-established predictors were identified from the previously established prediction models, we chose them to develop OWL without further selection; the features and the tuning parameters in OWL were detailed in Supplementary Table S6. The optimal hyper-parameters for OWL were: eta = 0.03, gamma = 0.2, subsample = 0.8, colsample_bytree = 1, min_child_weight = 10, nroud = 259. We further estimated predictors' contribution to OWL by SHAP value. For example, the longer the duration of smoking was associated with higher SHAP, indicating an increased risk of lung cancer (Supplementary Fig. S5). Moreover, baseline survival within 8-year follow-up was shown in Supplementary Table S7. To facilitate further validation and application, we have launched an interactive web-tool for implementation of OWL (http://47.97.212.52/#/OWL), and uploaded model source code of OWL model to Github (https://github.com/WeiLab4Research/OWL.git). Year-specific thresholds for

	UKB			PLCO			NLST		
	England area (N = 323,344)	Scotland and Wales area (N = 93,227)	Total (N = 416,571)	Control (N = 70,605)	Intervention (N = 66,231)	Total (N = 136,836)	Control (N = 23,138)	Intervention (N = 18,669)	Total (N = 41,807)
Age (years)[a]	56.12 ± 8.15	56.07 ± 8.04	56.11 ± 8.13	62.61 ± 5.36	62.50 ± 5.33	62.55 ± 5.35	61.28 ± 4.98	61.15 ± 4.94	61.22 ± 4.96
Gender (%)									
Female	168,750 (52.2)	41,986 (51.3)	216,718 (52.0)	35,266 (49.9)	32,914 (49.7)	68,180 (49.8)	9,221 (39.9)	7,342 (39.3)	16,563 (39.6)
Male	154,594 (47.8)	39,791 (48.7)	199,852 (48.0)	35,339 (50.1)	33,317 (50.3)	68,656 (50.2)	13,917 (60.1)	11,327 (60.7)	25,244 (60.4)
BMI (kg/m^2)[a]	27.34 ± 4.77	27.67 ± 4.74	27.41 ± 4.77	27.28 ± 4.90	27.34 ± 4.91	27.29 ± 4.83	27.97 ± 5.07	28.00 ± 5.07	27.99 ± 5.07
Missing	2,151 (0.7)	429 (0.5)	2,580 (0.6)	1,334 (1.9)	792 (1.2)	1,700 (1.4)	192 (0.8)	119 (0.6)	311 (0.7)
Race[a]									
White	301,767 (93.3)	91,530 (98.2)	393,297 (94.4)	62,333 (88.3)	58,438 (88.2)	120,771 (88.3)	20,988 (90.7)	16,800 (90.0)	37,788 (90.4)
Black	6,458 (2.0)	228 (0.2)	6,686 (1.6)	3,638 (5.2)	3,366 (5.1)	7,004 (5.1)	993 (4.3)	920 (4.9)	1,913 (4.6)
Hispanic	0 (0.0)	0 (0.0)	0 (0.0)	1,343 (1.9)	1,272 (1.9)	2,615 (1.9)	0 (0.0)	0 (0.0)	0 (0.0)
Asian	9,479 (2.9)	807 (0.9)	10,286 (2.5)	2,677 (3.8)	2,574 (3.9)	5,251 (3.8)	483 (2.1)	424 (2.3)	907 (2.2)
American Indian or Alaskan Native	0 (0.0)	0 (0.0)	0 (0.0)	173 (0.2)	169 (0.3)	342 (0.2)	85 (0.4)	68 (0.4)	153 (0.4)
Native Hawaiian or Pacific Islander	0 (0.0)	0 (0.0)	0 (0.0)	409 (0.6)	377 (0.6)	786 (0.6)	92 (0.4)	71 (0.4)	163 (0.4)
Missing	5,640 (1.7)	662 (0.7)	6,302 (1.5)	32 (0.0)	35 (0.1)	67 (0.0)	497 (2.1)	386 (2.1)	883 (2.1)
Smoke status (%)[a]									
Never	177,011 (54.7)	52,103 (55.9)	229,114 (55.0)	32,576 (46.1)	30,940 (46.7)	63,516 (46.4)	0 (0.0)	0 (0.0)	0 (0.0)
Former	110,296 (34.1)	30,621 (32.8)	140,917 (33.8)	30,470 (43.2)	28,344 (42.8)	58,814 (43.0)	12,028 (52.0)	9,770 (52.3)	21,798 (52.1)
Current	34,008 (10.5)	10,119 (10.9)	44,127 (10.6)	7,545 (10.7)	6,930 (10.5)	14,475 (10.6)	11,110 (48.0)	8,899 (47.7)	20,009 (47.9)
Missing	2,029 (0.7)	384 (0.4)	2,413 (0.6)	14 (0.0)	17 (0.0)	31 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Age at begin smoking (years)[b]	17.43 ± 4.33	17.32 ± 4.25	17.41 ± 4.32	18.59 ± 5.00	18.62 ± 5.11	18.61 ± 5.05	16.71 ± 3.73	16.69 ± 3.73	16.68 ± 3.70
Duration of smoking (years)[a,b]	26.88 ± 12.87	27.88 ± 12.74	27.11 ± 12.85	27.70 ± 13.79	27.37 ± 13.80	27.54 ± 13.79	39.67 ± 7.32	39.49 ± 7.29	39.54 ± 7.31
Average number of cigarettes per day[a,b]	18.00 ± 10.07	18.85 ± 10.40	18.20 ± 10.15	24.96 ± 14.71	24.60 ± 14.43	24.79 ± 14.58	28.44 ± 11.55	28.49 ± 11.42	28.69 ± 11.50
Pack year[b]	22.83 ± 18.38	24.89 ± 19.34	23.30 ± 18.62	36.24 ± 29.55	35.31 ± 28.87	35.79 ± 29.23	55.73 ± 23.79	55.58 ± 23.71	56.06 ± 23.85
Duration since quitting smoking (years)[a,b]	19.10 ± 11.82	18.42 ± 11.65	18.95 ± 11.78	20.22 ± 12.03	20.37 ± 11.96	20.29 ± 12.00	7.32 ± 4.76	7.30 ± 4.77	7.31 ± 4.76
Education level (%)[a]									
Less than high-school graduate	84,142 (26.0)	23,313 (25.0)	107,455 (25.8)	5,187 (7.3)	4,849 (7.3)	10,036 (7.3)	1,352 (5.8)	1,105 (5.9)	2,457 (5.9)
High-school graduate	35,738 (11.1)	9,754 (10.5)	45,492 (10.9)	16,265 (23.0)	15,145 (22.9)	31,410 (23.0)	5,537 (23.9)	4,235 (22.7)	9,772 (23.4)
Some training after high school	15,673 (4.8)	4,789 (5.2)	20,462 (4.9)	8,968 (12.7)	8,230 (12.4)	17,198 (12.6)	3,193 (13.9)	2,591 (13.9)	5,784 (13.8)
Some college	0 (0.0)	0 (0.0)	0 (0.0)	15,236 (21.7)	14,441 (21.8)	29,677 (21.7)	5,296 (22.9)	4,371 (23.4)	9,667 (23.1)
College graduate	125,359 (38.8)	34,537 (37.0)	159,887 (38.4)	11,879 (16.8)	11,313 (17.1)	23,192 (16.9)	3,907 (16.9)	3,219 (17.2)	7,126 (17.0)
Postgraduate or professional degree	0 (0.0)	0 (0.0)	0 (0.0)	12,800 (18.1)	12,152 (18.3)	24,952 (18.2)	3,709 (16.0)	3,052 (16.4)	6,761 (16.2)
Missing	62,441 (19.3)	20,834 (22.3)	83,275 (20.0)	270 (0.4)	101 (0.2)	371 (0.3)	144 (0.6)	96 (0.5)	240 (0.6)
Family history of lung cancer (%)[a,c]									
No	288,671 (89.3)	82,051 (88.0)	370,722 (89.0)	60,912 (86.3)	57,054 (86.1)	117,966 (86.2)	17,739 (76.7)	14,373 (77.0)	32,112 (76.8)
Yes	27,470 (8.5)	9,453 (10.1)	36,923 (8.9)	7,314 (10.4)	6,865 (10.4)	14,179 (10.4)	4,998 (21.6)	4,008 (21.5)	9,006 (21.5)
Missing	7,203 (2.2)	1,723 (1.8)	8,926 (2.1)	2,379 (3.3)	2,312 (3.5)	4,691 (3.4)	401 (1.7)	288 (1.5)	689 (1.6)
Diabetes (%)									
No	306,628 (94.8)	88,617 (95.1)	395,245 (94.9)	64,629 (91.6)	60,837 (91.9)	125,466 (91.7)	20,760 (89.7)	16,747 (89.7)	37,507 (89.7)
Yes	16,716 (5.2)	4,610 (4.9)	21,326 (5.1)	5,398 (7.6)	5,127 (7.7)	10,525 (7.7)	2,244 (9.7)	1,844 (9.9)	4,088 (9.8)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	578 (0.8)	267 (0.4)	845 (0.6)	134 (0.6)	78 (0.4)	212 (0.5)
Chronic bronchitis (%)									
No	322,475 (99.7)	93,046 (99.8)	415,521 (99.7)	66,674 (94.4)	62,849 (94.9)	129,523 (94.7)	20,816 (90.0)	16,812 (90.1)	37,628 (90.0)
Yes	869 (0.3)	181 (0.2)	1,050 (0.3)	3,311 (4.7)	3,082 (4.6)	6,393 (4.7)	2,151 (9.3)	1,752 (9.4)	3,903 (9.3)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	620 (0.9)	300 (0.5)	920 (0.6)	171 (0.7)	105 (0.6)	276 (0.7)

(Table 1 continues on next page)

	UKB			PLCO			NLST		
	England area (N = 323,344)	Scotland and Wales area (N = 93,227)	Total (N = 416,571)	Control (N = 70,605)	Intervention (N = 66,231)	Total (N = 136,836)	Control (N = 23,138)	Intervention (N = 18,669)	Total (N = 41,807)
(Continued from previous page)									
Emphysema (%)									
No	322,857 (99.8)	93,089 (99.9)	415,946 (99.8)	68,260 (96.7)	64,378 (97.2)	132,638 (96.9)	21,248 (91.8)	17,238 (92.3)	3,8486 (92.1)
Yes	487 (0.2)	138 (0.1)	625 (0.2)	1,763 (2.5)	1,586 (2.4)	3,349 (2.5)	1,712 (7.4)	1,323 (7.1)	3,035 (7.3)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	582 (0.8)	267 (0.4)	849 (0.6)	178 (0.8)	108 (0.6)	286 (0.7)
COPD (%)[a]									
No	317,658 (98.2)	91,544 (98.2)	409,202 (98.2)	65,420 (92.7)	61,723 (93.2)	127,143 (92.9)	19,034 (82.3)	1,5432 (82.7)	34,466 (82.4)
Yes	5,686 (1.8)	1,683 (1.8)	7,369 (1.8)	4,556 (6.5)	4,190 (6.3)	8,746 (6.4)	3,907 (16.9)	3,108 (16.6)	7,015 (16.8)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	629 (0.8)	318 (0.5)	947 (0.7)	197 (0.9)	129 (0.7)	326 (0.8)

UKB = UK Biobank; PLCO = the Prostate, Lung, Colorectal and Ovarian cancer screening trial; NLST = the National Lung Screening Trial. COPD = chronic bronchitis, emphysema, or chronic obstructive pulmonary disease. Continuous variables were expressed in mean ± standard deviation, and categorical variables in number and percentage (%). [a]Age, BMI, race, smoke status, duration of smoking, number of cigarettes per day, duration since quitting smoking, education level, family history of lung cancer, and COPD were included in PLCO$_{m2012.}$ [b]Age at begin smoking, duration of smoking, average number of cigarettes per day, and pack year were summarized among ever-smokers. Duration since quitting smoking were summarized among former smokers. [c]Either parents, siblings, or children were diagnosed of lung cancer before baseline survey.

***Table 1:* Distribution of predictors of OWL in the UKB, PLCO, and NLST populations.**

risk discrimination were estimated (Supplementary Table S8) for different levels of prespecified sensitivities. Taking 8-years prediction as an example, to include 80% of lung cancers, a threshold of 0.303% would yield a specificity of 74% and a positive predictive value of 1.5%.

Comparison of models' discrimination

For general population

In the training set, the OWL model yielded a C-index of 0.86 (95% CI: 0.85–0.87). Internal 5-fold cross-validation resulted in a C-index of 0.84. External validation of OWL yielded similar results, i.e., a C-index of 0.86 (95% CI: 0.84–0.87), 0.84 (95% CI: 0.83–0.85) and 0.84 (95% CI: 0.83–0.85) in validation sets 1–3, respectively. The time-dependent AUCs of OWL remained around 0.85 within 8 years follow-up across validation sets 1–3 (Fig. 2A), and the OWL score could differentiate individuals' risk significantly (Supplementary Fig. S6).

In validation set 1, OWL exhibited the best discrimination ability compared to PLCO$_{all2014}$ (6-year prediction, AUC$_{OWL}$ = 0.855, 95% CI: 0.829–0.880 vs. AUC$_{PLCOall2014}$ = 0.821, 95% CI: 0.794–0.848; $p < 0.001$) and LLPv3 (5-year prediction, AUC$_{OWL}$ = 0.850, 95% CI: 0.822–0.878 vs. AUC$_{LLPv3}$ = 0.832, 95% CI: 0.803–0.861; p = 0.001). In validation sets 2 & 3, OWL showed discriminative ability comparable to PLCO$_{all2014}$ (6-year prediction, AUC$_{OWL}$ = 0.861, 95% CI: 0.841–0.881 vs. AUC$_{PLCOall2014}$ = 0.869, 95% CI: 0.850–0.889, AUC$_{PLCOall2014\text{-}AUCOWL}$ < 1%), and better discrimination than LLPv3 (Fig. 2A, Supplementary Table S9). The results of comparison of PRAUC were consistent with those of AUC (Supplementary Table S10).

For ever-smokers

In validation set 1, OWL achieved the highest discrimination (AUC$_{5\text{-year}}$ = 0.838, 95% CI: 0.806–0.870; AUC$_{6\text{-year}}$ = 0.842, 95% CI: 0.814–0.871), followed by PLCO$_{m2012}$ (AUC$_{6\text{-year}}$ = 0.792, 95% CI: 0.760–0.823), PLCO$_{all2014}$ (AUC$_{6\text{-year}}$ = 0.791, 95% CI: 0.760–0.823), and LLPv3 (AUC$_{5\text{-year}}$ = 0.804, 95% CI: 0.770–0.838) (Fig. 2B, Supplementary Table S11). In validation sets 2 to 5 (PLCO and NLST populations), OWL was comparable to PLCO$_{m2012}$ and PLCO$_{all2014}$ (6-years AUC difference from −0.014 to 0.008) and outperformed LLPv3 (Fig. 2B). The comparison of PRAUC yielded similar results with those of AUC (Supplementary Table S12).

Comparison of models' calibration

For general population

OWL had a better calibration within 8-year follow-up compared to the comparative models (Fig. 3A, Supplementary Table S13, Supplementary Fig. S7). In validation set 1, OWL reached the best calibration (5-year prediction, EO ratio = 1.002, 95% CI: 0.899–1.133; 6-year prediction, EO ratio = 1.037, 95% CI: 0.939–1.159), followed by LLPv3 (EO ratio = 0.846, 95% CI: 0.759–0.957; RI$_{EO}$ of OWL vs. LLPv3 = 98.7%, $p < 0.001$), while PLCO$_{all2014}$ had the moderate bias in general population (EO ratio = 1.303, 95% CI: 1.179–1.455; RI$_{EO}$ of OWL vs. PLCO$_{all2014}$ = 87.8%, $p < 0.001$). In validation sets 2 & 3, OWL was better calibrated than PLCO$_{all2014}$ (RI$_{EO}$ = 48.5% in validation set 3) and LLPv3 (RI$_{EO}$ = 41.4% and 69.2% in validation sets 2 & 3, respectively).

For ever-smokers

Among the ever-smokers in all validation sets, OWL achieved a better calibration compared to PLCO$_{m2012}$, PLCO$_{all2014}$ (RI$_{EO}$ from 43.1% to 92.3%, all $p < 0.001$), and LLPv3 (RI$_{EO}$ from 44.9% to 90.4%, all $p < 0.001$) (Fig. 3B, Supplementary Fig. S8, Table S14), with exception of the comparison with PLCO$_{m2012}$ and PLCO$_{all2014}$ in validation set 3 and LLPv3 in validation set 5.

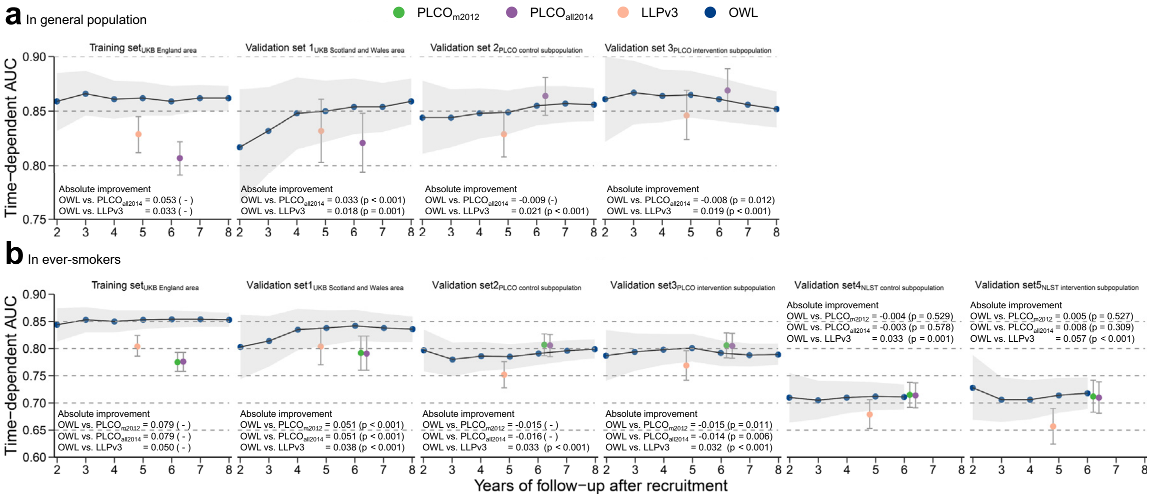

***Fig. 2:* Comparisons of model discrimination.** (a) Comparisons of OWL, PLCO$_{all2014}$, and LLPv3 in general population. (b) Comparison of OWL, PLCO$_{m2012}$, PLCO$_{all2014}$, and LLPv3 in ever-smokers. Because OWL was developed in training set, no comparison with OWL was performed in this set. Because PLCO$_{m2012}$ and PLCO$_{all2014}$ were developed in PLCO control subpopulation, OWL was not compared with PLCO$_{m2012}$ and PLCO$_{all2014}$ in this set. UKB = UK Biobank; PLCO = the Prostate, Lung, Colorectal and Ovarian cancer screening trial; NLST = the National Lung Screening Trial; PLCO$_{m2012}$ = the PLCO modified 2012; PLCO$_{all2014}$ = the PLCO modified 2014; LLPv3 = the Liverpool lung cancer project risk model version 3; OWL = the Optimized early Warning model for Lung cancer risk; AUC = Area under receiver operating characteristic curve. - = p value for the model comparison is not applicable due to the current set is the training set of either one of the comparing models.

Comparison of models' clinical utility

For general population

OWL presented better clinical utility than the competing models (Fig. 4). With 6 years of follow-up, for example, OWL yielded an average *NB* of 126 (per 100,000 people), significantly higher than PLCO$_{all2014}$ (average *NB* = 94) in validation set 1 (improvement in average *NB* = 32, $p < 0.001$). In validation set 3, OWL exhibited a slightly lower, but non-significant, *NB* than PLCO$_{all2014}$ (average $NB_{PLCOall2014}$ = 419 vs. NB_{OWL} = 402, p = 0.108). In validation sets 1–3, OWL outperformed LLPv3 in *NB* (Fig. 4).

For ever-smokers

Similar results hold for the comparisons among the ever-smokers (Supplementary Fig. S9). In validation set 1, compared to PLCO$_{m2012}$ and PLCO$_{all2014}$, OWL showed a considerable improvement of *NB* (average $NB_{PLCOm2012}$ = 213 and $NB_{PLCOall2014}$ = 212 vs. NB_{OWL} = 284, both p values < 0.001). In validation sets 3–5, OWL exhibited clinical utility equivalent to PLCO$_{m2012}$ and PLCO$_{all2014}$ (average *NB* difference from −34 to 47, all p > 0.05). Among all the validation sets, OWL showed significantly improved average *NB* values (average *NB* difference from 45 to 139, all p < 0.001) compared to LLPv3.

Performance of models in the USPSTF$_{2021}$ criteria-negative subpopulation

By applying PLCO$_{all2014}$, LLPv3, and OWL to the USPSTF$_{2021}$ criteria-negative (CN) subpopulation (including never-smokers), all could identify a part of lung cancer cases among those who were failed to be detected by USPSTF$_{2021}$ criteria (Supplementary Table S15). Among CN subpopulations, OWL exhibited adequate prediction accuracy and yielded a C-index of 0.79 (95% CI 0.75–0.82), 0.76 (95% CI: 0.73–0.79), and 0.77 (95% CI: 0.74–0.79) in validation sets 1–3, respectively. OWL outperformed PLCO$_{all2014}$ and LLPv3 in validation set 1 in both discrimination, calibration and clinical utility, while presented discrimination, calibration and clinical utility comparable to PLCO$_{all2014}$ and LLPv3 among the CN subpopulations in validation sets 2–3 (Supplementary Tables S15 and S16, Fig. S10).

Contribution of physiological, laboratory, and genetic indicators

We further assessed the contribution of physiological, laboratory, and genetic indicators for lung cancer risk prediction. We observed the PLG score could further significantly distinguish the risk of lung cancer within each stratum categorized by OWL (Supplementary Fig. S11). For example, among the individuals with high-risks graded by OWL, individuals in the highest tertile group defined by the PLG score (with an 8-year cumulative incidence of 1.8%) had 2.6 times of risk of lung cancer as those in the lowest tertile group defined by the PLG score (with an 8-year cumulative incidence of 0.7%) (RR = 2.6, 95% CI: 2.2–2.9). Further, the OWL^{+} model, by adding physiological, laboratory and genetic indicators to the OWL model, showed a significantly improved risk discrimination in lung cancer (Supplementary Table S17).

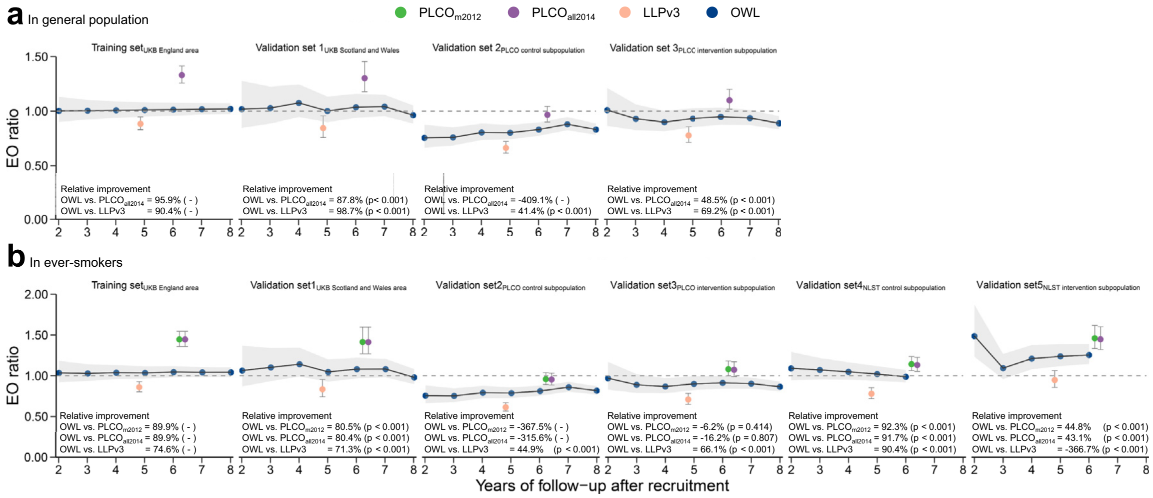

***Fig. 3:* Comparison of model calibration.** (a) Comparisons of OWL, $PLCO_{all2014}$, and LLPv3 in general population. (b) Comparison of OWL, $PLCO_{m2012}$, $PLCO_{all2014}$, and LLPv3 in ever-smokers. Because OWL was developed in training set, no comparison with OWL was performed in this set. Because $PLCO_{m2012}$ and $PLCO_{all2014}$ were developed in PLCO control subpopulation, OWL was not compared with $PLCO_{m2012}$ and $PLCO_{all2014}$ in this set. UKB = UK Biobank; PLCO = the Prostate, Lung, Colorectal and Ovarian cancer screening trial; $PLCO_{m2012}$ = the PLCO modified 2012; $PLCO_{all2014}$ = the PLCO modified 2014; LLPv3 = the Liverpool lung cancer project risk model version 3; OWL = the Optimized early Warning model for Lung cancer risk; EO ratio = Expected to observed lung cancer case ratio. - = p value for the model comparison is not applicable due to the current set is the training set of either one of the comparing models.

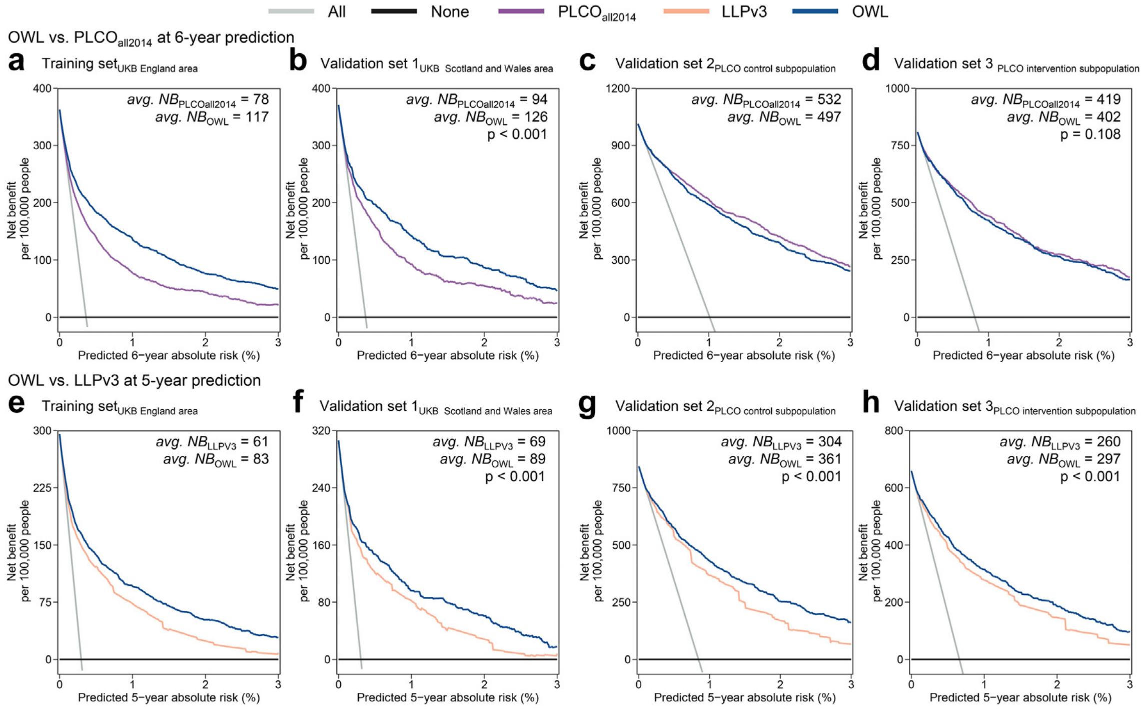

***Fig. 4:* Decision curve analysis of models in general population.** (a-d) Comparison OWL and $PLCO_{all2014}$ in training set, and validation set 1 to 3. (e-h) Comparison OWL and LLPv3 in training set, and validation set 1 to 3. Because OWL was developed in UKB training set, no comparison with OWL was performed in this set. Because $PLCO_{all2014}$ was developed in PLCO control subpopulation, OWL was not compared with $PLCO_{all2014}$ in this set. UKB = UK Biobank; PLCO = the Prostate, Lung, Colorectal and Ovarian cancer screening trial; $PLCO_{all2014}$ = the PLCO modified 2014; LLPv3 = the Liverpool lung cancer project risk model version 3; OWL = the Optimized early Warning model for Lung cancer risk; *NB* = Net benefit.

Discussion

We have developed a systematic approach for developing a useful risk prediction model to facilitate identification of target populations for LDCT screening. That is, by leveraging prior lung cancer risk studies via a systematic review, we identified relevant clinical and genetic predictors; we optimized the prediction model by utilizing XGBoost, a powerful machine learning algorithm, and extensively validated the model in five external validation sets from three UKB, PLCO and NLST populations. The proposed OWL model, with satisfactory predictive accuracy and robustness, could be a feasible tool to select subpopulation with high risks of lung cancer for further LDCT screening.

We focused on comparing OWL with $\text{PLCO}_{\text{m2012}}$, $\text{PLCO}_{\text{all2014}}$ and LLPv3, due to their good performance.[43,44] Recently published interim analysis of a prospective cohort study indicated that $\text{PLCO}_{\text{m2012}}$ was more efficient than the USPSTF_{2013} criteria for selecting individuals to enroll into lung cancer screening progrmmes.[10] In our comprehensive comparative analysis, OWL appears to present better discriminative ability than $\text{PLCO}_{\text{m2012}}$, $\text{PLCO}_{\text{all2014}}$, and LLPv3 in the general population.[45] Meanwhile, OWL had a comparable discriminative power with $\text{PLCO}_{\text{m2012}}$ model even among the PLCO population, and had a efficiency similar to $\text{PLCO}_{\text{m2012}}$ in the NLST population, which is composed of heavy smokers, indicating plausible generalizability and scalability of the proposed OWL model. Notably, OWL showed better calibration than $\text{PLCO}_{\text{m2012}}$ in the ever-smokers of PLCO and NLST, and should be considered for more round validations. The UKB, a large-scale prospective cohort for general population, enables OWL to be designed to 'catch more cases' in the longer follow-up and extend the period of prediction to 8-year, and thus provides a fundamental advantage of OWL compared with the competing models that are designed to estimate the risk at either 6-year or 5-year. On the other hand, OWL is a well-calibrated survival model with ability to predict the risk of lung cancer at any time point in the next 8-years, which may be closer to the needs of practical application than $\text{PLCO}_{\text{m2012}}$ and LLPv3.

OWL provided a general framework for identifying high-risk individuals among criteria-negative individuals.[46] In the UKB England area, for example, there were 511 USPSTF_{2021} criteria-negative individuals that developed lung cancer within 8 years of follow-up, which highlighted the fact that lung cancer is a complex disease driven by multiple factors, including environment, clinical, and genetic factors.[47] As a result, USPSTF_{2021} criteria-negative individual could be with high risks of lung cancer if harboring risk factors beyond age and smoking information. Identifying these risk factors and profiling these USPSTF_{2021} criteria-negative patients is critically important, and our proposed OWL model address this urgent call.

We have also delved into the quantitative assessment of the clinical utility of OWL by using decision curve analysis, i.e., the net benefit. The net benefit reflecting the number of net true positives per 100,000 persons assessed for lung cancer risk.[41,48] Considering of 33 million people at risk in the US, OWL, with an average *NB* of 123, is equivalent to a strategy that led to LDCT screen in an expected 41 thousand people at risk, with all screened positive for lung cancer. However, the $\text{PLCO}_{\text{all2014}}$ model is equivalent to a lower benefit strategy that led to LDCT screen in an expected 27 thousand people at risk, and all were screened positive.

OWL has the same ease of use as the competitive models. In terms of the number of variables the model need, OWL requires 15 variables, $\text{PLCO}_{\text{m2012}}$ and $\text{PLCO}_{\text{all2014}}$ require 11 variables, and LLPv3 requires 7 variables, while the variables used in the above models are all easily obtained from questionnaire. In addition, although the OWL model cannot be expressed by mathematical formulas, we provide an easy-to-use platform for general users.

Notably, there was a significant difference in cumulative smoking intensity between ever-smokers of UKB, PLCO, and NLST: the mean pack-year of UKB ever-smokers was the lowest, followed by PLCO ever-smokers, and the NLST ever-smokers was the highest, suggesting higher concentration of high-risk persons in PLCO and NLST. It was acknowledged that high discrimination is easier to obtain in population where people are heterogeneous with regard to risk.[13] Therefore, the discrimination of OWL among ever-smokers of PLCO and NLST populations was lower than that observed in the ever-smokers of UKB.

Our study has several strengths. First, OWL was developed based on a large-scale prospective cohort, and extensively and externally validated in five large-scale validation sets. As such, OWL presented much accuracy, robustness, and generalizability, and extends the span of prediction to 8 years than competitive models. Second, we performed a systematic review of lung cancer risk prediction models and conducted a comprehensive comparison analysis with three competing models. Our results suggested the competitiveness of OWL. Finally, we developed an OWL^{+} model by incorporating physiological, laboratory, and genetic indicators causally associated with the risk of lung cancer, which exemplifies a prototype of next-generation precision prevention tools tailored by individuals' unique genetic and phenotypic information.[24]

We acknowledge limitations. First, the candidate predictors in our model were confined to those identified from prior studies, more predictors need to be investigated to improve the performance of OWL. Second, the FEV1, CRP, total bilirubin, and PRS were unavailable in PLCO (the majority of subjects were with missing genotypic data) and NLST, the OWL^{+} warrants more rounds of external validation. Third, our study sample largely consists of participants of European ancestry, so the generalizability to other ethnicities is unknow. However, OWL is worthy of further trans-ancestry validation in other large-scale populations. Finally, the lung cancer development process is dynamic, time-dependent predictors are needed to address the dynamic prediction capability.[49]

Conclusion

We proposed an optimized early warning model for lung cancer risk (OWL) by using an advanced machine learning technique, XGBoost, based on a large-scale UKB cohort, and independently validated it in five validation sets from the UKB, PLCO, and NLST populations. Our comparative study suggested that OWL performed well in predictive accuracy and robustness, exhibiting as a feasible means for sieving individuals with high risks of lung cancer for further LDCT screening.

Contributors

Y.W. and F.C. contributed to the study conception and design. Z.P., R.Z., and S.S. performed data management and curation. Z.P., R.Z., S.S., and J.C. performed data analysis; Z.P., Y.Lin., L.Z., Xiang.W., Q.Y., and Xuan.W. contributed to the systematic review, data extraction and curation. Z.P., R.Z., and Y.W. drafted the manuscript. F.C. Y.Li., Y.W., and D.C.C. critically reviewed the interpretation and revised the manuscript. All authors read and approved the final version of the manuscript.

Data sharing statement

Data request, subject to regulatory guidelines, should be made to the corresponding author at ywei@pku.edu.cn. The authors affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Declaration of interests

We declare no competing interests.

Acknowledgments

We thank all participants and research staffs who participated in UKB, PLCO, and NLST for providing the data. We also thank Liangmin Wei, Juanjuan Fan and Yue Jiang from Nanjing Medical University for providing valuable assistances in model construction. This study was funded by the National Natural Science Foundation of China (81973142 to YW, 82220108002 to FC and 82273737 to RZ); the NIH (CA209414, CA092824, and ES000002 to DCC, CA269398 and CA249096 to YLi) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). RZ was partially supported by the Qing Lan Project of the Higher Education Institutions of Jiangsu Province and the Outstanding Young Level Academic Leadership Training Program of Nanjing Medical University.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104443.

References

1 Sung H, Ferlay J, Siegel RL, 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):209–249.

2 Allemani C, Matsuda T, Di Carlo V, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*. 2018;391(10125):1023–1075.

3 Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395–409.

4 Moyer VA. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2014;160(5):330–338.

5 de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med*. 2020;382(6):503–513.

6 Field JK, Duffy SW, Baldwin DR, et al. The UK Lung Cancer Screening Trial: a pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer. *Health Technol Assess*. 2016;20(40):1–146.

7 Oudkerk M, Devaraj A, Vliegenthart R, et al. European position statement on lung cancer screening. *Lancet Oncol*. 2017;18(12):e754–e766.

8 Oudkerk M, Liu S, Heuvelmans MA, Walter JE, Field JK. Lung cancer LDCT screening and mortality reduction — evidence, pitfalls and future perspectives. *Nat Rev Clin Oncol*. 2021;18(3):135–151.

9 Krist AH, Davidson KW, Mangione CM, et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *JAMA*. 2021;325(10):962–970.

10 Tammemägi MC, Ruparel M, Tremblay A, et al. USPSTF2013 vs. PLCOm2012 lung cancer screening eligibility criteria (International Lung Screening Trial): interim analysis of a prospective cohort study. *Lancet Oncol*. 2021;23(1):138–148.

11 Ten Haaf K, Bastani M, Cao P, et al. A comparative modeling analysis of risk-based lung cancer screening strategies. *J Natl Cancer Inst*. 2020;112(5):466–479.

12 Toumazis I, Bastani M, Han SS, Plevritis SK. Risk-based lung cancer screening: a systematic review. *Lung Cancer*. 2020;147:154–186.

13 Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. *N Engl J Med*. 2013;368(8):728–736.

14 Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer*. 2008;98(2):270–276.

15 Field JK, Vulkan D, Davies MPA, Duffy SW, Gabe R. Liverpool Lung Project lung cancer risk stratification model: calibration and prospective validation. *Thorax*. 2021;76(2):161–168.

16 Tammemägi MC, Church TR, Hocking WG, 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.

17 Mazzone PJ, Silvestri GA, Souter LH, et al. Screening for lung cancer: CHEST guideline and expert panel report. *Chest*. 2021;160(5):e427–e494.

18 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794.

19 Hong W, Zhou X, Jin S, et al. A comparison of XGBoost, random forest, and nomograph for the prediction of disease severity in patients with COVID-19 pneumonia: implications of cytokine and immune cell profile. *Front Cell Infect Microbiol*. 2022;12:819267.

20 Lyu ZY, Tan FW, Lin CQ, et al. The development and validation of risk prediction model for lung cancer: a systematic review. *Zhonghua Yu Fang Yi Xue Za Zhi*. 2020;54(4):430–437.

21 Horsfall LJ, Burgess S, Hall I, Nazareth I. Genetically raised serum bilirubin levels and lung cancer: a cohort study and Mendelian randomisation using UK Biobank. *Thorax*. 2020;75(11):955–964.

22 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

23 Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–58.

24 Hung RJ, Warkentin MT, Brhane Y, et al. Assessing lung cancer absolute risk trajectory based on a polygenic risk model. *Cancer Res*. 2021;81(6):1607–1615.

25 Zhang R, Shen S, Wei Y, et al. A large-scale genome-wide gene-gene interaction study of lung cancer susceptibility in Europeans

with a trans-ethnic validation in asians. *J Thorac Oncol*. 2022;17(8):974–990.

26 Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367–378.

27 Verwaeren J, Van der Weeën P, De Baets B. A search grid for parameter optimization as a byproduct of model sensitivity analysis. *Appl Math Comput*. 2015;261:8–27.

28 Lundberg SM, Lee S-I, eds. *A unified approach to interpreting model predictions*. NIPS; 2017.

29 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): 2300–2311.

30 Rutter CM, Miglioretti DL. Estimating the accuracy of psychological scales using longitudinal data. *Biostatistics*. 2003;4(1):97–107.

31 Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56(2):337–344.

32 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.

33 Abe D, Inaji M, Hase T, et al. A prehospital triage system to detect traumatic intracranial hemorrhage using machine learning algorithms. *JAMA Netw Open*. 2022;5(6):e2216393.

34 He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263–1284.

35 Li T, Tong W, Roberts R, Liu Z, Thakkar S. Deep learning on high-throughput transcriptomics to predict drug-induced liver injury. *Front Bioeng Biotechnol*. 2020;8:562677.

36 Wilcox RR. Chapter 5 - comparing two groups. In: Wilcox RR, ed. *Introduction to robust estimation and hypothesis testing*. Fifth Edition. Academic Press; 2022:153–251.

37 Bandos AI, Rockette HE, Gur D. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Stat Med*. 2005;24(18):2873–2893.

38 Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460.

39 Robbins HA, Alcala K, Swerdlow AJ, et al. Comparative performance of lung cancer risk models to define lung screening eligibility in the United Kingdom. *Br J Cancer*. 2021;124(12):2026–2034.

40 Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. *JAMA*. 2015;313(4):409–410.

41 Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74(6):796–804.

42 Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. *Lancet Oncol*. 2015;16(4):e173–e180.

43 Katki HA, Kovalchik SA, Petito LC, et al. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. *Ann Intern Med*. 2018;169(1):10–19.

44 Ten Haaf K, Jeon J, Tammemägi MC, et al. Risk prediction models for selection of lung cancer screening candidates: a retrospective validation study. *PLoS Med*. 2017;14(4):e1002277.

45 Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol*. 2017;186(9):1026–1034.

46 Faselis C, Nations JA, Morgan CJ, et al. Assessment of lung cancer risk among smokers for whom annual screening is not recommended. *JAMA Oncol*. 2022;8(10):1428–1437.

47 Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. *Eur Respir J*. 2016;48(3):889–902.

48 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.

49 Chen W, Sin DD, FitzGerald JM, et al. An individualized prediction model for long-term lung function trajectory and risk of COPD in the general population. *Chest*. 2020;157(3):547–557.