Predicting the future risk of lung cancer: development, and internal and external validation of the CanPredict (lung) model in 19·67 million people and evaluation of model performance against seven other risk prediction models

Weiqi Liao, Carol A C Coupland, Judith Burchardt, David R Baldwin, collaborators of the DART initiative, Fergus V Gleeson, Julia Hippisley-Cox*

Summary

Background Lung cancer is the second most common cancer in incidence and the leading cause of cancer deaths worldwide. Meanwhile, lung cancer screening with low-dose CT can reduce mortality. The UK National Screening Committee recommended targeted lung cancer screening on Sept 29, 2022, and asked for more modelling work to be done to help refine the recommendation. This study aims to develop and validate a risk prediction model—the CanPredict (lung) model—for lung cancer screening in the UK and compare the model performance against seven other risk prediction models.

Methods For this retrospective, population-based, cohort study, we used linked electronic health records from two English primary care databases: QResearch (Jan 1, 2005–March 31, 2020) and Clinical Practice Research Datalink (CPRD) Gold (Jan 1, 2004–Jan 1, 2015). The primary study outcome was an incident diagnosis of lung cancer. We used a Cox proportional-hazards model in the derivation cohort (12·99 million individuals aged 25–84 years from the QResearch database) to develop the CanPredict (lung) model in men and women. We used discrimination measures (Harrell's C statistic, D statistic, and the explained variation in time to diagnosis of lung cancer [R^2_D]) and calibration plots to evaluate model performance by sex and ethnicity, using data from QResearch (4·14 million people for internal validation) and CPRD (2·54 million for external validation). Seven models for predicting lung cancer risk (Liverpool Lung Project [LLP]$_{v2}$, LLP$_{v3}$, Lung Cancer Risk Assessment Tool [LCRAT], Prostate, Lung, Colorectal, and Ovarian [PLCO]$_{M2012}$, PLCO$_{M2014}$, Pittsburgh, and Bach) were selected to compare their model performance with the CanPredict (lung) model using two approaches: (1) in ever-smokers aged 55–74 years (the population recommended for lung cancer screening in the UK), and (2) in the populations for each model determined by that model's eligibility criteria.

Findings There were 73 380 incident lung cancer cases in the QResearch derivation cohort, 22 838 cases in the QResearch internal validation cohort, and 16 145 cases in the CPRD external validation cohort during follow-up. The predictors in the final model included sociodemographic characteristics (age, sex, ethnicity, Townsend score), lifestyle factors (BMI, smoking and alcohol status), comorbidities, family history of lung cancer, and personal history of other cancers. Some predictors were different between the models for women and men, but model performance was similar between sexes. The CanPredict (lung) model showed excellent discrimination and calibration in both internal and external validation of the full model, by sex and ethnicity. The model explained 65% of the variation in time to diagnosis of lung cancer in both sexes in the QResearch validation cohort and 59% of the R^2_D in both sexes in the CPRD validation cohort. Harrell's C statistics were 0·90 in the QResearch (validation) cohort and 0·87 in the CPRD cohort, and the D statistics were 2·8 in the QResearch (validation) cohort and 2·4 in the CPRD cohort. Compared with seven other lung cancer prediction models, the CanPredict (lung) model had the best performance in discrimination, calibration, and net benefit across three prediction horizons (5, 6, and 10 years) in the two approaches. The CanPredict (lung) model also had higher sensitivity than the current UK recommended models (LLP$_{v2}$ and PLCO$_{M2012}$), as it identified more lung cancer cases than those models by screening the same amount of individuals at high risk.

Interpretation The CanPredict (lung) model was developed, and internally and externally validated, using data from 19·67 million people from two English primary care databases. Our model has potential utility for risk stratification of the UK primary care population and selection of individuals at high risk of lung cancer for targeted screening. If our model is recommended to be implemented in primary care, each individual's risk can be calculated using information in the primary care electronic health records, and people at high risk can be identified for the lung cancer screening programme.

Funding Innovate UK (UK Research and Innovation).

Copyright © 2023 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Lancet Respir Med 2023

Published Online April 5, 2023 https://doi.org/10.1016/S2213-2600(23)00050-4

See Online/Comment https://doi.org/10.1016/S2213-2600(23)00083-8

For the Chinese translation of the abstract see Online for appendix 1

*Collaborators of the DART initiative are listed in appendix 2

Nuffield Department of Primary Care Health Sciences (W Liao PhD, Prof C A C Coupland PhD, J Burchardt MRCGP, Prof J Hippisley-Cox MD FRCP) **and Department of Oncology** (Prof F V Gleeson FRCR), **University of Oxford, Oxford, UK; School of Medicine, University of Nottingham, Nottingham, UK** (Prof C A C Coupland, Prof D R Baldwin MD FRCP); **Department of Respiratory Medicine, Nottingham University Hospitals NHS Trust, Nottingham, UK** (Prof D R Baldwin)

Correspondence to: Prof Julia Hippisley-Cox, Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Observatory Quarter, Oxford OX2 6GG, UK **julia.hippisley-cox@phc.ox.ac.uk**

See Online for appendix 2

Research in context

Evidence before this study

Lung cancer screening with low-dose CT can reduce mortality. Using risk prediction models to select people at high risk of lung cancer for screening programmes is an efficient strategy at the population level, because it can avoid misusing resources spent on screening people at low risk. The Liverpool Lung Project (LLP_{v2}) and Prostate Lung Colorectal and Ovarian ($PLCO_{M2012}$) models have been recommended to calculate individuals' risk of lung cancer in the Targeted Lung Health Check Programme in England. However, a previous study showed that the models achieved only moderate discrimination and were not well calibrated when externally validated using Clinical Practice Research Datalink (CPRD) data in the English primary care population. In preparation for ethical approval and data extraction from electronic health record databases, we searched PubMed using the terms "lung cancer" AND "prediction model" in free text and Medical Subject Headings (MeSH) in "Title/Abstract" between Jan 1, 2000 and Dec 31, 2020, with no language restrictions, to understand the contemporary research on prediction models for lung cancer and prepare a long list of potential predictors for data specification. We updated the literature search before starting the analysis and modelling work (September, 2021), and when writing this paper (June, 2022) to see whether there were any new published studies.

Added value of this study

Developed and internally and externally validated using robust statistical methodologies in the QResearch and CPRD databases with a large sample size (19·67 million people in total from the primary care population), the CanPredict (lung) model shows excellent discrimination and calibration in the full model, by sex and ethnicity. Compared with the other seven models (LLP_{v2}, LLP_{v3}, Lung Cancer Risk Assessment Tool [LCRAT], $PLCO_{M2012}$, $PLCO_{M2014}$, Pittsburgh, and Bach) in ever-smokers aged 55–74 years and using the eligibility criteria for participants in each model, the CanPredict (lung) model had the best model performance for discrimination, calibration, and net benefit. It also had higher sensitivity than the current UK recommended models (LLP_{v2} and $PLCO_{M2012}$). The CanPredict (lung) model is an inclusive and flexible model, with a prediction horizon of 1–10 years for different ethnicities. It also allows sex-specific risk stratification. The CanPredict (lung) model could be applied to the UK primary care population for risk stratification and selection of eligible individuals for lung cancer screening using low-dose CT.

Implications of all the available evidence

Through a thorough evaluation of model performance and comparison of eight prediction models, we provide research evidence to assist policy makers in deciding which risk prediction models could be used for lung cancer screening in the UK. The CanPredict (lung) model can be implemented in primary care computer systems, which allow batch-mode processes that can facilitate the selection of eligible individuals at high risk of lung cancer for screening. Such implementation would greatly reduce human resources when the screening programme rolls out nationally. Patients diagnosed through a screening-detected route are more likely to be diagnosed at earlier stages, which could lead to better patient outcomes and reduced lung cancer mortality. Given the incidence and mortality of lung cancer at the population level, the improvement in lung cancer could substantially contribute to the UK Government's ambition that 75% of people with cancer will be diagnosed at early stages by 2028.

Introduction

Lung cancer is a global public health issue. With 2·2 million new cases and 1·8 million deaths globally per year, lung cancer is the second most common cancer in incidence and the leading cause of cancer deaths worldwide.[1] Research evidence from randomised clinical trials has shown that using low-dose CT for lung cancer screening reduces mortality.[2,3] The US Preventive Services Task Force recommended using low-dose CT for lung cancer screening in 2013,[4] and relaxed the eligibility criteria for screening in 2021, by lowering the age threshold from 55 to 50 years, and smoking exposure from 30 to 20 pack-years.[5] On Sept 29, 2022, the UK National Screening Committee recommended targeted lung cancer screening for people aged 55–74 years who are at high risk of lung cancer. However, the committee did not recommend which models should be used for risk estimation. Instead, the committee asked for more modelling work to be done to help: (1) refine the recommendation; (2) address implementation challenges; and (3) determine the optimum protocols and pathways for lung cancer screening across the UK. Therefore, we conducted this study to develop and validate a multivariable prognostic model—the CanPredict (lung) model—to predict the future risk of lung cancer in men and women for up to 10 years, and compared the model performance against the other risk prediction models for lung cancer. We hope to provide timely research evidence for the UK National Screening Committee for decision making.

For the **recommendations from the UK National Screening Committee** see https://view-health-screening-recommendations.service.gov.uk/lung-cancer/

For more on the **QResearch database** see https://www.qresearch.org/

Methods

Study design, study population, and data source

For this three-stage population-based cohort study, we used electronic health records (EHRs) from the QResearch database (version 45) to develop and internally validate the model, and then used the Clinical Practice Research Datalink (CPRD) Gold database for external validation of the model. The QResearch database is one of the largest health-care databases in England, and a Trusted Research Environment accredited by Health Data Research UK. Detailed information on the CPRD database has previously been published.[6] We included adult patients aged 25–84 years who were registered with

general practices and who contributed to the QResearch database between Jan 1, 2005 and March 31, 2020, and excluded those with a diagnosis of lung cancer before cohort entry. The broad age range covers most of the adult primary care population and provides great flexibility to select people in different age groups to evaluate model performance or perform subgroup analyses. To ensure the data were complete, the included individuals needed to have been registered in a general practice for at least 12 months, and the general practices needed to have contributed to the QResearch database for at least 12 months before the cohort entry date. For external validation using the CPRD Gold database, we used the same sampling criteria except the study period was from Jan 1, 2004 to Jan 1, 2015, due to data availability.

This project was approved by the QResearch Scientific Committee on March 8, 2021. QResearch is a research ethics-approved database, confirmed by the East Midlands–Derby Research Ethics Committee (research ethics reference: 18/EM/0400; project reference: OX37 DART). Research ethics for using the CPRD Gold data for this study was approved by CPRD's Research Data Governance (reference: 13_079R) on Jan 13, 2022.

We published a comprehensive research protocol and statistical analysis plan[7] for this project before conducting the analyses. We used the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement to guide the conduct and reporting of this study.[8,9] A TRIPOD checklist for this Article is in appendix 3 (pp 5–6).

Outcomes

The primary study outcome was an incident diagnosis of lung cancer recorded on one or more of the four linked data sources in QResearch—primary care and secondary care (ie, hospital episode statistics [HES]) databases, cancer registry (previously from Public Health England, now part of NHS Digital), and death registry (from the Office for National Statistics [ONS]). We used the earliest date on any of the four records as the date of lung cancer diagnosis. We used Read or SNOMED-CT codes to identify events from the general practice record, and ICD-10 codes to identify events from HES, and cancer and death registries. Cancer registry data were not available in our CPRD dataset; the primary outcome was from one or more of the three linked data sources: CPRD, HES, and ONS.

Statistical analysis

Stage 1: development and internal validation of the CanPredict (lung) model using QResearch

We used established methodologies to develop and validate risk prediction algorithms.[10,11] Three-quarters of general practices were randomly selected for the derivation dataset and the remaining quarter for the validation dataset. We shortlisted potential predictors through literature review and clinical input, and also considered the available information from the QResearch database. Multiple imputation with chained equations was used to replace missing values for ethnicity, BMI, and alcohol and smoking status, with five imputations for the derivation and validation datasets. We used the imputed values in our main analyses and Rubin's rules to combine the results across the imputed datasets.[12] Fractional polynomials[13] were used to model non-linear associations between the continuous variables (ie, age, BMI, and Townsend deprivation scores) and the outcome. Cox proportional-hazards models were used to estimate the coefficients for each risk factor for men and women separately, using robust variance estimates to allow for patient clustering within general practices. People who died from causes other than lung cancer were censored on the date of death in Cox regression; people who left their general practices were censored on the date they left. We initially fitted a full model with all potential predictors for men and women, tested for interaction terms, and retained variables with a hazard ratio (HR) of less than 0·91 or more than 1·10 (for binary or categorical variables) for clinical significance; statistical significance was set at the 0·01 level. Further explanation of how the CanPredict (lung) model works is in appendix 3 (pp 1–3).

Stage 2: comparison of the CanPredict (lung) model with seven other models for lung cancer prediction in the QResearch validation dataset

A systematic review of risk prediction models for lung cancer screening[14] and empirical studies comparing some mainstream prediction models have been published in recent years.[15,16] We referred to these studies and included models with prediction horizons of at least 5 years, as the purpose of screening is to detect cancer early in asymptomatic populations. The sojourn time for lung cancer progressing from a preclinical stage (detectable by screening tests) to clinical stages is 3–6 years, and is longer in women than in men.[17] Therefore, prediction models designed for a longer period are more suitable for screening than those designed for a shorter period (eg, 1 or 2 years) for diagnostic purposes.

See Online for appendix 3

We included seven models for comparison with our model: the Liverpool Lung Project (LLP_{v2} and LLP_{v3}) models,[18,19] the Lung Cancer Risk Assessment Tool (LCRAT),[20] the Prostate, Lung, Colorectal and Ovarian ($PLCO_{M2012}$ and $PLCO_{M2014}$) models,[21,22] the Pittsburgh model,[23] and the Bach model[24] (table 1). Most models were developed in the US population, and the LLP used regional data from northwest England. The rationale for including the seven models and how we handled EHRs for the variables in these models is in appendix 3 (pp 1–3).

Evaluation of model performance

We calculated the absolute predicted risks using each model for individuals in the validation datasets. We used Harrell's C statistic,[25] D statistic,[26] and the explained variation in time to diagnosis of lung cancer (R^2_D)[27] to

	CanPredict (lung) model	LLP_{v2} and LLP_{v3}[18,19]	LCRAT[20]	$PLCO_{M2012}$ and $PLCO_{M2014}$[21,22]	Pittsburgh[23]	Bach[24]
Country	England	England	USA	USA	USA	USA
Prediction horizon	Up to 10 years	5 years	5 years	6 years	6 years	Up to 10 years
Age range, years	25–84	40–84	55–74	55–74	50–79	45–69
Includes never-smokers (in addition to ever-smokers)	Yes	Yes	No	Yes (2014 only)	No	No
Predictors						
Age	Yes	Yes	Yes	Yes	Yes	Yes
Sex	Yes	Yes	Yes	..	..	Yes
Education	..	..	Yes	Yes	..	..
Race or ethnicity	Yes	..	Yes	Yes	..	..
Socioeconomic status (Townsend score)	Yes	..	..	..	..	..
Smoking status	Yes	..	..	Yes	Yes	..
Smoking duration, years	..	Yes	Yes	Yes	Yes	Yes
Smoking intensity, cigarettes per day	Yes	..	Yes	Yes	Yes	Yes
Years quit smoking	..	..	Yes	Yes	..	Yes
Pack-years of smoking	..	..	Yes	..	..	..
Alcohol	Yes	..	..	..	..	..
BMI	Yes	..	Yes	Yes	..	..
COPD	Yes	Yes	Yes	Yes	..	..
Personal history of cancer	Yes	Yes	..	Yes	..	..
Family history of lung cancer	Yes	Yes	Yes	Yes	..	..
History of pneumonia	Yes	Yes	..	..	..	..
Asbestos exposure	Yes	Yes	..	..	..	Yes
Asthma	Yes	..	..	..	..	..
Venous thromboembolism	Yes	..	..	..	..	..

Age, BMI, and Townsend scores were included as continuous variables in the CanPredict (lung) model; fractional polynomials were used to fit their non-linear association with the outcome (two fractional polynomial terms). COPD includes bronchitis and emphysema in the CanPredict model. COPD=chronic obstructive pulmonary disease. LLP=Liverpool Lung Project. LCRAT=Lung Cancer Risk Assessment Tool. PLCO=Prostate, Lung, Colorectal, and Ovarian Cancer Screening programme.

Table 1: **Summary of predictors in the CanPredict (lung) model and the other seven prediction models in this study**

evaluate the discrimination of models. These statistics were combined across the imputed datasets using Rubin's rules.[12] Harrell's C statistic is similar to the area under the receiver operating characteristic curve (AUC) but takes account of the censored nature of cohort data. Higher values of Harrell's C statistic and the D statistic indicate better discrimination. Higher R^2_D values indicate that a greater proportion of variation in time to lung cancer diagnosis is explained by the model. To assess calibration, we used the pmcalplot package in Stata[28] to compare the observed risks with the mean predicted risks in 20ths of the predicted risk (predicted risk probabilities were sorted ascendingly and then split [the sorted predicted risk] into 20 equal parts) by sex for each algorithm. Decision curve analysis[29] was used to evaluate the net benefit of the prediction models (clinical usefulness) by sex and prediction horizon.

For more on **lung cancer screening** see https://view-health-screening-recommendations.service.gov.uk/lung-cancer/ and https://www.cancerresearchuk.org/health-professional/screening/lung-cancer-screening

The published protocol stated we would use the Brier score, but we did not use it in this study for the following two reasons. The Brier score includes components of both discrimination and calibration, but we have specific statistical methods to evaluate the two aspects of the model separately. In addition, the Brier score should be interpreted carefully; a lower Brier score does not necessarily imply higher calibration. A scaled Brier score reflects the overall performance of the model, similar to an R^2-type assessment, and we already used R^2_D in this study.

We first evaluated the discrimination and calibration of the CanPredict (lung) model in the whole QResearch validation cohort, by sex and ethnicity. Then, we evaluated the discrimination, calibration, and net benefits of all eight prediction models using two approaches. The first approach was to evaluate the models in current and ex-smokers (called ever-smokers in this Article) aged 55–74 years, which is the population for the Targeted Lung Health Check programme in England and recommended for lung cancer screening in the UK . The second approach was to compare the CanPredict (lung) model with each of the seven other models using those models' eligibility criteria for study participants and their prediction horizons. The CanPredict (lung) model was developed in a study population with a wide age range and included all smoking statuses, with a prediction horizon from 1 to 10 years. Therefore, the CanPredict (lung) model could adapt to the eligibility criteria of each model and allow for the comparison to be made (appendix 3 pp 1–3).

	Primary care population (total n= 19 671 498)			Incident lung cancer cases (total n=112 363)		
	Derivation (QResearch)	Validation (QResearch)	Validation (CPRD)	Derivation (QResearch)	Validation (QResearch)	Validation (CPRD)
Sample size	12 991 042	4 137 199	2 543 257	73 380	22 838	16 145
Sex						
Male	6 476 207 (49·9%)	2 059 175 (49·8%)	1 255 755 (49·4%)	41 003 (55·9%)	12 768 (55·9%)	9324 (57·8%)
Female	6 514 835 (50·1%)	2 078 024 (50·2%)	1 287 502 (50·6%)	32 377 (44·1%)	10 070 (44·1%)	6821 (42·2%)
Age, mean (SD)	45·0 (15·6)	45·2 (15·6)	48·5 (15·3)	65·8 (10·4)	66·0 (10·4)	66·7 (10·1)
Age groups						
25–29 years	2 307 317 (17·8%)	718 752 (17·4%)	240 013 (9·4%)	122 (0·2%)	24 (0·1%)	6 (0·04%)
30–34 years	2 002 174 (15·4%)	628 639 (15·2%)	308 991 (12·2%)	261 (0·4%)	80 (0·4%)	39 (0·2%)
35–39 years	1 618 962 (12·5%)	517 697 (12·5%)	330 466 (13·0%)	652 (0·9%)	189 (0·8%)	95 (0·6%)
40–44 years	1 358 535 (10·5%)	436 425 (10·5%)	305 698 (12·0%)	1421 (1·9%)	479 (2·1%)	264 (1·6%)
45–49 years	1 159 196 (8·9%)	369 609 (8·9%)	258 045 (10·2%)	2908 (4·0%)	895 (3·9%)	487 (3·0%)
50–54 years	996 141 (7·7%)	318 840 (7·7%)	226 532 (8·9%)	5212 (7·1%)	1562 (6·8%)	1038 (6·4%)
55–59 years	913 302 (7·0%)	294 035 (7·1%)	228 239 (9·0%)	8956 (12·2%)	2681 (11·7%)	1896 (11·7%)
60–64 years	753 825 (5·8%)	242 618 (5·9%)	184 555 (7·3%)	11 276 (15·4%)	3466 (15·2%)	2440 (15·1%)
65–69 years	637 454 (4·9%)	204 891 (5·0%)	152 012 (6·0%)	12 966 (17·7%)	4101 (18·0%)	2872 (17·8%)
70–74 years	524 814 (4·0%)	169 382 (4·1%)	128 486 (5·1%)	12 710 (17·3%)	3959 (17·3%)	2900 (18·0%)
75–79 years	431 758 (3·3%)	140 891 (3·4%)	108 115 (4·3%)	10 927 (14·9%)	3452 (15·1%)	2589 (16·0%)
80–84 years	287 564 (2·2%)	95 420 (2·3%)	72 105 (2·8%)	5969 (8·1%)	1950 (8·5%)	1519 (9·4%)
Townsend score, mean (SD)	0·6 (3·2)	0·4 (3·2)	−0·6 (3·2)	0·4 (3·1)	0·3 (3·1)	−0·1 (3·3)
Townsend quintile						
Quintile 1 (most affluent)	2 830 215 (21·8%)	928 894 (22·5%)	576 717 (22·7%)	16 164 (22·0%)	5128 (22·5%)	2999 (18·6%)
Quintile 2	2 651 740 (20·4%)	869 971 (21·0%)	569 982 (22·4%)	16 110 (22·0%)	4939 (21·6%)	3374 (20·9%)
Quintile 3	2 513 467 (19·3%)	823 887 (19·9%)	532 660 (20·9%)	15 511 (21·1%)	4976 (21·8%)	3501 (21·7%)
Quintile 4	2 447 443 (18·8%)	777 626 (18·8%)	505 016 (19·9%)	14 042 (19·1%)	4212 (18·4%)	3446 (21·3%)
Quintile 5 (most deprived)	2 548 177 (19·6%)	736 821 (17·8%)	358 882 (14·1%)	11 553 (15·7%)	3583 (15·7%)	2825 (17·5%)
Ethnicity						
Recorded	9 381 066 (72·2%)	2 963 779 (71·6%)	1 199 302 (47·2%)	51 054 (69·6%)	16 200 (70·9%)	5624 (34·8%)
White	7 481 059 (57·6%)	2 365 041 (57·2%)	1 052 950 (41·4%)	48 418 (66·0%)	15 405 (67·5%)	5490 (34·0%)
Indian	337 885 (2·6%)	121 854 (2·9%)	31 043 (1·2%)	359 (0·5%)	121 (0·5%)	25 (0·2%)
Pakistani	198 742 (1·5%)	61 207 (1·5%)	12 372 (0·5%)	273 (0·4%)	69 (0·3%)	11 (0·1%)
Bangladeshi	138 942 (1·1%)	35 258 (0·9%)	3937 (0·2%)	344 (0·5%)	72 (0·3%)	9 (0·1%)
Other Asian	220 406 (1·7%)	74 806 (1·8%)	21 375 (0·8%)	227 (0·3%)	84 (0·4%)	12 (0·1%)
Caribbean	138 224 (1·1%)	42 853 (1·0%)	11 085 (0·4%)	559 (0·8%)	183 (0·8%)	24 (0·1%)
Black African	323 946 (2·5%)	99 629 (2·4%)	26 127 (1·0%)	226 (0·3%)	59 (0·3%)	11 (0·1%)
Chinese	104 082 (0·8%)	27 714 (0·7%)	6675 (0·3%)	108 (0·1%)	47 (0·2%)	12 (0·1%)
Other	437 780 (3·4%)	135 417 (3·3%)	33 738 (1·3%)	540 (0·7%)	160 (0·7%)	30 (0·2%)
BMI						
Recorded	10 797 197 (83·1%)	3 459 976 (83·6%)	1 872 186 (73·6%)	66 190 (90·2%)	20 841 (91·3%)	11 528 (71·4%)
Mean (SD)	26·5 (5·3)	26·4 (5·2)	26·0 (4·7)	26·4 (5·1)	26·4 (5·0)	25·8 (4·3)
Smoking status						
Recorded	12 119 036 (93·3%)	3 868 125 (93·5%)	2 445 833 (96·2%)	70 739 (96·4%)	22 171 (97·1%)	16 030 (99·3%)
Non-smoker	6 851 842 (52·7%)	2 208 409 (53·4%)	1 162 279 (45·7%)	14 352 (19·6%)	4859 (21·3%)	2446 (15·2%)
Ex-smoker	2 389 705 (18·4%)	756 862 (18·3%)	412 494 (16·2%)	21 909 (29·9%)	6671 (29·2%)	3513 (21·8%)
Light smoker (1–9 cigarettes/day)	2 150 683 (16·6%)	673 258 (16·3%)	389 840 (15·3%)	24 810 (33·8%)	7688 (33·7%)	2893 (17·9%)
Moderate smoker (10–19 cigarettes/day)	465 255 (3·6%)	146 132 (3·5%)	293 130 (11·5%)	5150 (7·0%)	1547 (6·8%)	3782 (23·4%)
Heavy smoker (≥20 cigarettes/day)	261 551 (2·0%)	83 464 (2·0%)	188 090 (7·4%)	4518 (6·2%)	1406 (6·2%)	3396 (21·0%)
Alcohol status						
Recorded	10 531 942 (81·1%)	3 438 730 (83·1%)	2 134 853 (83·9%)	65 272 (89·0%)	20 649 (90·4%)	14 132 (87·5%)
Non-drinker	6 591 000 (50·7%)	2 163 345 (52·3%)	381 956 (15·0%)	41 780 (56·9%)	13 126 (57·5%)	2568 (15·9%)
Trivial (<1 unit/day)	2 013 473 (15·5%)	658 992 (15·9%)	966 644 (38·0%)	10 364 (14·1%)	3260 (14·3%)	5758 (35·7%)
Light (1–2 units/day)	996 128 (7·7%)	312 150 (7·5%)	578 481 (22·7%)	5340 (7·3%)	1733 (7·6%)	3830 (23·7%)

(Table 2 continues on next page)

	Primary care population (total n= 19 671 498)			Incident lung cancer cases (total n=112 363)		
	Derivation (QResearch)	Validation (QResearch)	Validation (CPRD)	Derivation (QResearch)	Validation (QResearch)	Validation (CPRD)
(Continued from previous page)						
Moderate (3–6 units/day)	801 199 (6·2%)	263 899 (6·4%)	170 114 (6·7%)	6467 (8·8%)	2059 (9·0%)	1567 (9·7%)
Heavy (7–9 units/day)	63 064 (0·5%)	20 415 (0·5%)	19 397 (0·8%)	805 (1·1%)	314 (1·4%)	208 (1·3%)
Very heavy (>9 units/day)	55 131 (0·4%)	16 032 (0·4%)	18 261 (0·7%)	395 (0·5%)	118 (0·5%)	201 (1·2%)
Amount not recorded	11 947 (0·1%)	3897 (0·1%)	..	121 (0·2%)	39 (0·2%)	..
Comorbidities						
COPD	182 746 (1·4%)	57 823 (1·4%)	37 676 (1·5%)	9686 (13·2%)	2995 (13·1%)	1982 (12·3%)
Asthma	1 295 982 (10·0%)	415 159 (10·0%)	257 244 (10·1%)	7976 (10·9%)	2493 (10·9%)	1730 (10·7%)
Pneumonia	185 790 (1·4%)	58 429 (1·4%)	43 817 (1·7%)	2347 (3·2%)	735 (3·2%)	602 (3·7%)
Asbestos exposure or asbestosis	12 714 (0·1%)	4034 (0·1%)	2706 (0·1%)	377 (0·5%)	127 (0·6%)	71 (0·4%)
Venous thromboembolism	144 720 (1·1%)	46 325 (1·1%)	33 912 (1·3%)	1943 (2·6%)	639 (2·8%)	453 (2·8%)
Cancers recorded at baseline						
Blood cancer	37 301 (0·3%)	11 969 (0·3%)	6973 (0·3%)	640 (0·9%)	197 (0·9%)	119 (0·7%)
Breast cancer	87 142 (0·7%)	28 764 (0·7%)	20 957 (0·8%)	1470 (2·0%)	495 (2·2%)	312 (1·9%)
Cervical cancer	9026 (0·1%)	2944 (0·1%)	2014 (0·1%)	184 (0·3%)	66 (0·3%)	40 (0·2%)
Colorectal cancer	30 258 (0·2%)	9746 (0·2%)	6641 (0·3%)	687 (0·9%)	240 (1·0%)	156 (1·0%)
Gastric-oesophageal cancer	5460 (<0·1%)	1828 (<0·1%)	1105 (<0·1%)	95 (0·1%)	40 (0·2%)	24 (0·1%)
Oral cancer	6004 (<0·1%)	1951 (<0·1%)	1176 (<0·1%)	211 (0·3%)	81 (0·4%)	41 (0·3%)
Ovarian cancer	7397 (0·1%)	2369 (0·1%)	1635 (0·1%)	85 (0·1%)	23 (0·1%)	17 (0·1%)
Renal cancer	20 290 (0·2%)	6391 (0·2%)	4006 (0·2%)	623 (0·8%)	208 (0·9%)	142 (0·9%)
Uterine cancer	7938 (0·1%)	2571 (0·1%)	1605 (0·1%)	119 (0·2%)	46 (0·2%)	24 (0·1%)
Family history of lung cancer	83 557 (0·6%)	28 929 (0·7%)	7351 (0·3%)	730 (1·0%)	286 (1·3%)	76 (0·5%)

Data are n (%), unless otherwise stated. COPD=chronic obstructive pulmonary disease. CPRD=Clinical Practice Research Datalink.

Table 2: **Baseline demographic and clinical characteristics of the derivation, and internal and external validation cohorts for the CanPredict (lung) model using the QResearch and CPRD Gold databases**

Stage 3: external validation using CPRD Gold

We then used a non-overlapping CPRD dataset to externally validate the CanPredict (lung) model for people aged 25–84 years. We also compared the performance of the CanPredict (lung) model with the other seven models in ever-smokers aged 55–74 years using the CPRD data. We used the same methods described above for these two sets of analyses.

All analyses were conducted using Stata (17.0) and R (4.1.0).

Comparison of different risk stratification approaches

We compared the performance of our risk prediction model with the models recommended in the Targeted Lung Health Check for risk stratification (LLP_{v2} and $PLCO_{M2012}$) in both QResearch and CPRD validation cohorts. The sensitivity values for different strategies were calculated,[30] which are the number and proportion of people diagnosed with lung cancer within the prediction horizon (5 years for LLP_{v2} and 6 years for $PLCO_{M2012}$) among the same proportion of individuals identified as high risk.

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.

Results

Table 2 shows the baseline demographic and clinical characteristics of the study population and the incident lung cancer cases in the derivation, and internal and external validation cohorts for the CanPredict (lung) model using the QResearch and CPRD databases. Of 12 991 042 individuals aged 25–84 years in the derivation cohort, 73 380 people developed incident lung cancer (0·56%) during follow-up. Of 4 137 199 individuals in the QResearch validation cohort, 22 838 developed incident lung cancer (0·55%). In the CPRD validation cohort, 16 145 (0·63%) of 2 543 257 developed lung cancer. The median follow-up time for the whole QResearch cohort (17·13 million people) was 4·2 years (IQR 1·7–8·5); individuals in the CPRD cohort had a longer median follow-up time of 6·4 years (IQR 2·9–10·4). The mean age at lung cancer diagnosis was similar in the two cohorts (around the age of 66–67 years). The proportion of the primary care population aged 25–34 years was higher in the QResearch database than in the CPRD database, especially for those aged 25–29 years. For other age bands from 35–39 to 80–84 years, the proportions were slightly higher in the CPRD cohort than in the QResearch cohort. There was a higher proportion of men than women in the incident lung cancer cases (55·9% in QResearch and 57·8% in CPRD). Around 70% of people

had their ethnicity recorded in QResearch, but only 47·2% of the primary care population and 34·8% of people with lung cancer had ethnicity information recorded in CPRD. The completeness of BMI was also lower in CPRD (73·6%) than in QResearch (83·1–83·6%). Smoking status was well recorded in both QResearch (93·3–93·5%) and CPRD (96·2%). Both the primary care population and patients diagnosed with lung cancer in QResearch had higher proportions of non-smokers, ex-smokers, and light smokers (1–9 cigarettes per day) than those in CPRD. There were higher proportions of moderate smokers (10–19 cigarettes per day) and heavy smokers (≥20 cigarettes per day) in the primary care population and patients diagnosed with lung cancer in CPRD than in QResearch. Other characteristics were broadly similar between QResearch and CPRD. The proportions with comorbidities, a family history of lung cancer, and any previous cancer at baseline were much higher in incident lung cancer cases than in the two primary care cohorts.

The predictors in the final CanPredict (lung) model included age, sex, ethnicity, Townsend score (a proxy for an individual's level of socioeconomic deprivation), smoking status and intensity, alcohol status, BMI, chronic obstructive pulmonary disease (COPD), asthma, pneumonia, venous thromboembolism, asbestos exposure, family history of lung cancer, personal history of cancer, and an interaction between age and smoking status (appendix 3 p 14). The adjusted hazard ratios (HRs) for the predictors are in table 3. Compared with being White, other ethnicities were less likely to be diagnosed with lung cancer (HR<1), except that the HRs for Bangladeshi and Chinese women were not significantly different from White women. The HR increased with current smoking status and smoking intensity (cigarettes per day) and generally displayed a dose–response association in both sexes. Two smoking-related conditions, COPD and previous oral cancer, had an HR of more than 2 in both sexes. Asbestos exposure, previous colorectal and gastric-oesophageal cancers, and alcohol status were all significant predictors in men, but not significant in women. Four previous gynaecological cancers (breast, cervical, ovarian, and uterine) were significant predictors in the women. The non-linear associations between the continuous variables (ie, age, BMI, Townsend score, and interaction between age and smoking status) and the outcome and the interaction terms are shown in appendix 3 (pp 13–14).

The descriptive statistics for the predicted risks for each of the eight models in the two approaches (ie, in ever-smoker patients aged 55–74 years and by the eligibility criteria of each model) are in appendix 3 (pp 9–10). The results for the full CanPredict (lung) model in people aged 25–84 years and in ever-smokers aged 55–74 (approach 1) are reported in the main text and appendix 3 (pp 19–27), and the results for approach 2 are fully reported in appendix 3 (pp 11 and 28–35).

	Men	Women
Ethnic groups		
White	Reference category	··
Indian	0·50 (0·44–0·58)	0·68 (0·58–0·79)
Pakistani	0·60 (0·52–0·69)	0·72 (0·55–0·93)
Bangladeshi	0·86 (0·75–0·98)	1·04 (0·83–1·32)
Other Asian	0·56 (0·46–0·69)	0·75 (0·59–0·96)
Caribbean	0·69 (0·63–0·75)	0·66 (0·58–0·76)
Black African	0·53 (0·43–0·64)	0·69 (0·56–0·85)
Chinese	0·59 (0·46–0·75)	1·19 (0·93–1·51)
Other ethnicities	0·72 (0·65–0·79)	0·79 (0·67–0·92)
Smoking status		
Non-smoker	Reference category	··
Ex-smoker	2·02 (1·82–2·24)	2·23 (2·03–2·44)
Light smoker (1–9 cigarettes per day)	6·59 (6·11–7·11)	5·89 (5·50–6·31)
Moderate smoker (10–19 cigarettes per day)	5·55 (4·94–6·23)	6·28 (5·66–6·95)
Heavy smoker (≥20 cigarettes per day)	9·25 (8·35–10·2)	9·90 (8·88–11·0)
Comorbidities*		
COPD	2·38 (2·30–2·45)	2·42 (2·34–2·51)
Asthma	0·94 (0·91–0·97)	1·12 (1·08–1·16)
Asbestos exposure	1·33 (1·20–1·48)	NS
Pneumonia	1·21 (1·15–1·28)	1·26 (1·18–1·34)
Venous thromboembolism	1·10 (1·03–1·18)	1·16 (1·09–1·24)
Cancer history*		
Family history of lung cancer	1·44 (1·30–1·59)	1·21 (1·09–1·35)
Previous oral cancer	2·55 (2·18–2·98)	2·30 (1·76–3·01)
Previous blood cancer	1·68 (1·52–1·86)	1·90 (1·68–2·14)
Previous renal cancer	1·36 (1·24–1·49)	1·59 (1·36–1·86)
Previous colorectal cancer	1·44 (1·31–1·58)	NS
Previous gastric-oesophageal cancer	1·27 (1·01–1·61)	NS
Previous breast cancer	NA	1·56 (1·48–1·65)
Previous cervical cancer	NA	1·55 (1·34–1·79)
Previous ovarian cancer	NA	1·27 (1·03–1·58)
Previous uterine cancer	NA	1·30 (1·08–1·55)
Alcohol status		
Non-drinker	Reference category	··
Trivial (<1 unit per day)	0·94 (0·91–0·97)	NS
Light (1–2 units per day)	0·94 (0·91–0·98)	NS
Moderate (3–6 units per day)	0·99 (0·97–1·03)	NS
Heavy (7–9 units per day)	1·13 (1·05–1·22)	NS
Very heavy (>9 units per day)	1·11 (1·00–1·24)	NS

The hazard ratios are adjusted for the fractional polynomial terms for age, BMI, Townsend score, and the interaction between age and smoking status; they are presented along with 95% CIs. COPD=chronic obstructive pulmonary disease. NA=not applicable. NS=non-significant. *Reference categories are the absence of the particular condition or cancer history.

Table 3: **Hazard ratios for the predictors in the CanPredict (lung) model**

In evaluating the discrimination of models, in the QResearch validation cohort, the CanPredict (lung) model explained 65% of the variation in time to diagnosis of lung cancer (R^2_D) in men and women aged 25–84 years.

	Harrell's C statistic		D statistic		Explained variation in time to lung cancer diagnosis*	
	Women	Men	Women	Men	Women	Men
Full validation cohort (individuals aged 25–84 years)						
CanPredict (lung) model	0·897 (0·893–0·900)	0·904 (0·901–0·906)	2·810 (2·770–2·850)	2·790 (2·760–2·830)	0·654 (0·648–0·660)	0·650 (0·645–0·656)
The Targeted Lung Health Check criteria (ever-smoker patients aged 55–74 years, approach 1)						
Prediction horizon: 5 years						
CanPredict (lung) model	0·727 (0·715–0·739)	0·735 (0·726–0·745)	1·925 (1·862–1·989)	1·709 (1·655–1·762)	0·469 (0·453–0·486)	0·411 (0·396–0·426)
LLP_{v2}	0·647 (0·635–0·659)	0·655 (0·645–0·665)	1·560 (1·498–1·622)	1·285 (1·232–1·338)	0·367 (0·349–0·386)	0·283 (0·266–0·299)
LLP_{v3}	0·660 (0·648–0·672)	0·662 (0·652–0·672)	1·634 (1·572–1·697)	1·337 (1·284–1·390)	0·389 (0·371–0·407)	0·299 (0·282–0·316)
LCRAT	0·642 (0·629–0·655)	0·657 (0·646–0·667)	1·573 (1·509–1·636)	1·329 (1·275–1·383)	0·371 (0·352–0·390)	0·297 (0·280–0·313)
Prediction horizon: 6 years						
CanPredict (lung) model	0·726 (0·715–0·737)	0·735 (0·726–0·743)	1·910 (1·851–1·968)	1·717 (1·668–1·767)	0·465 (0·450–0·481)	0·413 (0·399–0·427)
$PLCO_{M2012}$	0·531 (0·517–0·544)	0·545 (0·534–0·557)	1·281 (1·226–1·335)	0·984 (0·936–1·032)	0·281 (0·264–0·299)	0·188 (0·173–0·203)
$PLCO_{M2014}$	0·526 (0·513–0·540)	0·540 (0·529–0·551)	0·734 (0·676–0·792)	0·652 (0·602–0·701)	0·114 (0·098–0·130)	0·092 (0·079–0·105)
Pittsburgh	0·643 (0·631–0·654)	0·657 (0·648–0·666)	1·577 (1·518–1·636)	1·318 (1·231–1·406)	0·373 (0·355–0·390)	0·293 (0·266–0·321)
Prediction horizon: 10 years						
CanPredict (lung) model	0·724 (0·715–0·733)	0·731 (0·723–0·738)	1·896 (1·847–1·945)	1·717 (1·675–1·759)	0·462 (0·449–0·475)	0·413 (0·401–0·425)
Bach	0·575 (0·566–0·585)	0·586 (0·577–0·595)	1·345 (1·298–1·392)	1·112 (1·071–1·153)	0·302 (0·287–0·316)	0·228 (0·215–0·241)

Higher values of Harrell's C statistic, D statistic, and R^2_D indicate better discrimination. LCRAT=Lung Cancer Risk Assessment Tool. LLP=Liverpool Lung Project. PLCO=Prostate, Lung, Colorectal and Ovarian. R^2_D=explained variation in time to lung cancer diagnosis. *Referred to as R^2_D in the text.

Table 4: **Discrimination statistics of the CanPredict (lung) model in the full validation cohort (QResearch) and all eight prediction models using the Targeted Lung Health Check criteria**

The D statistic was 2·8 and Harrell's C statistic was 0·9 in both sexes (indicating excellent discrimination; table 4). The model also has excellent discrimination statistics, stratified by ethnicity and sex (appendix 3 p 8). For ever-smokers aged 55–74 years (approach 1), the CanPredict (lung) model had higher values of Harrell's C statistic, D statistic, and compared with all other seven prediction models in both sexes. Apart from the CanPredict (lung) model, LLP_{v3} had slightly higher values in the three discrimination statistics than LLP_{v2} and LCRAT in the 5-year prediction horizon. The Pittsburgh predictor consistently had the highest values in the three discrimination statistics in the 6-year horizon, followed by $PLCO_{M2012}$; $PLCO_{M2014}$ had the smallest values. A common pattern in the 6-year horizon category was that men had higher Harrell's C statistics than women, and women had higher D statistics and R^2_D values than men across all models.

The full CanPredict (lung) model showed good calibration in both sexes in the QResearch validation cohort and by ethnicity (appendix 3 pp 15–18). Compared with the White population, the number of lung cancer cases in ethnic minorities was small. Therefore, the confidence intervals for some risk bands were wide in the calibration plots. The majority of ethnic minorities were at low risk (observed and predicted risk <1%), especially women. For ever-smokers aged 55–74 years, apart from the CanPredict (lung) model, other prediction models were poorly calibrated. In the 5-year horizon (appendix 3 p 19), there was overestimation (ie, predicted risks were greater than observed risks) in the LLP_{v2}, LLP_{v3}, and LCRAT models in both sexes, especially in LLP_{v2} and at higher risk bands. LLP_{v3} was better calibrated than LLP_{v2}. In the 6-year horizon (appendix 3 p 20), the $PLCO_{M2012}$ model severely underestimated (ie, predicted risks were lower than observed risks) at low-risk bands (which had high proportions of light smokers and patients with increased socioeconomic deprivation) and overestimated at high-risk bands. $PLCO_{M2014}$ underestimated risk across all 20 bands of the predicted risk in both sexes, whereas the Pittsburgh predictor overestimated the risk. The Bach model was poorly calibrated in both sexes, underestimated at low-risk bands, and overestimated at high-risk bands (appendix 3 p 21). Calibration plots restricted to predicted risks of 5% or less for better visibility at low-risk bands are shown in appendix 3 (pp 22–24).

The net benefit (decision curve analysis) for the prediction models in ever-smokers aged 55–74 years by sex over the three prediction horizons in the QResearch validation cohort are in appendix 3 (pp 25–27). The CanPredict (lung) model had the highest net benefit, compared with the other prediction models and strategies considering either no individuals or all individuals for intervention across a range of risk thresholds.

The CanPredict (lung) model showed excellent discrimination in the external CPRD validation cohort for the full model (primary care population aged 25–84 years), although the statistics for discrimination were smaller than those in the QResearch validation cohort. The CanPredict (lung) model explained 59% of the R^2_D; the D statistic was around 2·4, and Harrell's C statistic was 0·87 in women and 0·88 in men (table 5). The model for men performed slightly better than the model for women in CPRD.

	Harrell's C statistic		D statistic		Explained variation in time to lung cancer diagnosis*	
	Women	Men	Women	Men	Women	Men
Full validation cohort (individuals aged 25–84 years)						
CanPredict (lung) model	0·869 (0·865–0·873)	0·878 (0·875–0·881)	2·44 (2·40–2·48)	2·47 (2·43–2·50)	0·587 (0·579–0·595)	0·592 (0·585–0·599)
The Targeted Lung Health Check criteria (ever-smoker patients aged 55–74 years, approach 1)						
Prediction horizon: 5 years						
CanPredict (lung) model	0·694 (0·682–0·707)	0·700 (0·690–0·710)	1·196 (1·121–1·272)	1·207 (1·147–1·268)	0·255 (0·231–0·279)	0·258 (0·239–0·277)
LLP_{v2}	0·628 (0·616–0·640)	0·645 (0·635–0·654)	0·705 (0·631–0·779)	0·785 (0·723–0·847)	0·106 (0·086–0·126)	0·128 (0·111–0·146)
LLP_{v3}	0·634 (0·622–0·646)	0·651 (0·641–0·661)	0·766 (0·692–0·840)	0·858 (0·795–0·921)	0·123 (0·102–0·144)	0·149 (0·131–0·168)
LCRAT	0·699 (0·687–0·711)	0·693 (0·684–0·703)	1·182 (1·108–1·256)	1·115 (1·056–1·174)	0·250 (0·227–0·274)	0·229 (0·210–0·248)
Prediction horizon: 6 years						
CanPredict (lung) model	0·693 (0·681–0·704)	0·698 (0·688–0·707)	1·187 (1·117–1·258)	1·188 (1·132–1·245)	0·252 (0·230–0·274)	0·252 (0·234–0·270)
$PLCO_{M2012}$	0·679 (0·668–0·691)	0·649 (0·639–0·659)	1·064 (0·996–1·132)	0·888 (0·833–0·943)	0·213 (0·192–0·234)	0·158 (0·142–0·175)
$PLCO_{M2014}$	0·684 (0·673–0·696)	0·653 (0·643–0·663)	1·119 (1·052–1·187)	0·937 (0·882–0·992)	0·230 (0·209–0·252)	0·173 (0·156–0·190)
Pittsburgh	0·653 (0·641–0·664)	0·658 (0·649–0·667)	0·957 (0·886–1·028)	0·935 (0·878–0·992)	0·179 (0·158–0·201)	0·173 (0·155–0·190)
Prediction horizon: 10 years						
CanPredict (lung) model	0·687 (0·677–0·696)	0·695 (0·687–0·703)	1·151 (1·094–1·209)	1·176 (1·129–1·222)	0·240 (0·222–0·259)	0·248 (0·233–0·263)
Bach	0·682 (0·673–0·691)	0·664 (0·657–0·672)	1·050 (0·996–1·105)	0·929 (0·885–0·973)	0·208 (0·191–0·226)	0·171 (0·157–0·184)

Higher values of Harrell's C statistic, D statistic, and R^2_D indicate better discrimination. LCRAT=Lung Cancer Risk Assessment Tool. LLP=Liverpool Lung Project. PLCO=Prostate, Lung, Colorectal and Ovarian. R^2_D=explained variation in time to lung cancer diagnosis. *Referred to as R^2_D in the text.

Table 5: **Discrimination statistics of the CanPredict (lung) model in the external validation cohort (CPRD Gold) and eight prediction models using the Targeted Lung Health Check criteria**

For ever-smokers aged 55–74 years, the CanPredict (lung) model had higher values of Harrell's C statistic, the D statistic, and R^2_D in both sexes, compared with the other seven prediction models (table 5), except that Harrell's C statistic was slightly lower in the CanPredict (lung) model than in the LCRAT in the model for women. Both LLP_{v2} and $PLCO_{M2012}$ were at the bottom of the discrimination statistics in each prediction horizon in external validation.

Most values in the three discrimination measures for the eight models were smaller in the CPRD validation cohort than those in the QResearch validation cohort, with some exceptions. Harrell's C statistic was significantly higher in $PLCO_{M2012}$, $PLCO_{M2014}$, and Bach (non-overlapping 95% CI), and slightly higher in LCRAT and Pittsburgh (with some overlapping 95% CI between QResearch and CPRD); the D statistic and R^2_D were significantly higher in $PLCO_{M2014}$ in the CPRD validation cohort than in the QResearch cohort.

The calibration plots for external validation using CPRD data are in appendix 3 (pp 36–39). The full CanPredict (lung model; individuals aged 25–84 years) was well calibrated at low-risk bands and overestimated the risk for the top 10% of people with predicted risk of more than 5% in both sexes (appendix 3 p 36). For ever-smokers aged 55–74 years, in the 5-year horizon (appendix 3 p 37), the calibration in the four models (ie, CanPredict, LLPv2, LLPv3, and LCRAT) was better for men than for women. Overestimation was observed in all four models in both sexes, and LLP_{v2} had the worst calibration. In the 6-year horizon (appendix 3 p 38), the CanPredict (lung) model overestimated people with predicted risk of more than 2% in both sexes. The $PLCO_{M2012}$ underestimated the risks in the majority of women with a predicted risk less than 2% and men with a predicted risk less than 4%. $PLCO_{M2014}$ underestimated risks across all risk bands in both sexes, whereas the Pittsburgh predictor overestimated in all risk bands. In the 10-year horizon (appendix 3 p 39), both the CanPredict (lung) model and the Bach model overestimated the risk, but the Bach model had poorer calibration than the CanPredict (lung) model in both sexes.

The net benefit (decision curve analysis) for the eight prediction models in ever-smokers aged 55–74 years by sex in the CPRD validation cohort are in appendix 3 (pp 40–42). The CanPredict (lung) model had the highest net benefit across a range of risk thresholds in the three prediction horizons.

When comparing the CanPredict (lung) model with each of the seven other models using those models' eligibility criteria for study participants and their prediction horizons (the second approach), the CanPredict (lung) model had higher values in all three discrimination measures across three prediction horizons (appendix 3 pp 11). As for calibration, $PLCO_{M2014}$ and the Pittsburgh predictor had the same patterns as reported previously in ever-smokers aged 55–74 years (the first approach), whereas LLP_{v2}, LLP_{v3}, and the Bach model had better calibration when applying their specific model eligibility criteria than they had with the first approach. However, the CanPredict (lung) model still had better calibration even using the eligibility criteria for each model, with the predicted risks closely matching the observed risk across all predicted risk bands (appendix 3 pp 28–31). The CanPredict (lung) model also

had the best net benefit (decision curve analysis; appendix 3 pp 32–35).

Different strategies for identifying individuals at high risk of lung cancer using the QResearch and CPRD validation cohorts were compared and are presented in appendix 3 (p 43). The CanPredict (lung) model had higher sensitivity than LLP_{v2} and $PLCO_{M2012}$, as it identified more lung cancer cases by screening the same amount of individuals at high risk.

Discussion

In this study, we used data from 19·67 million asymptomatic people from the primary care population to develop, and internally and externally validate, the CanPredict (lung) model to estimate the risk of men and women aged 25–84 years being diagnosed with lung cancer in the next 10 years using the QResearch and CPRD databases. The predictors include sociodemographic characteristics, lifestyle factors, comorbidities, family history of lung cancer, and personal history of other cancers. We compared the CanPredict (lung) model against seven other lung cancer prediction models and found that it had the best performance in discrimination, calibration, and net benefits across three prediction horizons (5, 6, and 10 years) among ever-smokers aged 55–74 years in both the QResearch and CPRD validation cohorts. We also compared the model performance using the eligibility criteria of each model and found that the CanPredict (lung) model performed better than each model.

The findings in the QResearch and CPRD validation cohorts were generally consistent. However, we observed some discrepancies in model performance, such as increased Harrell's C statistics in $PLCO_{M2012}$ and $PLCO_{M2014}$ in the CPRD validation cohort, despite the majority of discrimination statistics across models being decreased in this cohort. The heterogeneity in sociodemographic characteristics between the databases might provide some explanations for the differences in the results. Smoking and age are two known strong predictors of lung cancer. The primary care population and patients diagnosed with lung cancer in the CPRD cohort smoked more heavily than those in the QResearch cohort. The primary care population in CPRD was on average 3·4 years older than that in QResearch (48·5 *vs* 45·1 years). The higher proportion of young people in the QResearch database than in the CRPD database is likely to increase the discrimination measures in the full model, as people younger than 35 years are less likely to be diagnosed with lung cancer. The primary care population in the CPRD cohort was more affluent, and patients diagnosed with lung cancer were more socioeconomically deprived, than those in QResearch.

Lung cancer does occur in never smokers. We included never-smokers in our model, as we would like our model to be as inclusive as possible and the model can be used widely in the primary care population for early detection of lung cancer. Never smokers might have some comorbidities, genetic susceptibility, a personal history of cancer, or a family history of lung cancer as background risk factors. Ever-smokers aged 55–74 years were recommended for lung cancer screening in the UK by the National Screening Committee. We focused on a thorough assessment of the model performance of this population in this study to inform health policy. However, we recognise the need for an accurate estimation of lung cancer risk in never smokers, and whether never-smokers with a high risk should be recommended for lung cancer screening. We will design a separate study to address this important question in the future.

Missing data and loss to follow-up are two inevitable problems in longitudinal cohorts.[31] We used multiple imputation to replace the missing data in both model development and validation cohorts. Individuals lost to follow-up were censored in Cox regression, and the performance measures calculated in the validation data accounted for censoring. Missing data for ethnicity was a problem in the CPRD cohort. Besides White people, the sample size for other ethnicities diagnosed with lung cancer was 30 individuals or fewer (including both men and women and across all age groups). We only evaluated the performance of the full CanPredict (lung) model in the QResearch validation cohort, but not in the CPRD validation cohort, due to the small number of people from ethnic minorities in the CPRD cohort.

It is more important for prediction models to have accurate calibration at low-risk bands (eg, ≤3%) than at higher risk bands for screening, as people with predicted risks over the threshold will be eligible for screening anyway, whereas miscalibration at low-risk bands might result in missing people potentially with lung cancer by not screening them or wasting resources by screening people at low-risk. However, we have not recommended a risk threshold for our models for lung cancer screening, as the balance between benefits and harms, cost-effectiveness, availability of health resources, accessibility and health equality, and the potential impact of the screening programme at the population level, will all need to be taken into consideration when deciding a threshold for lung cancer screening. A separate study to evaluate cost-effectiveness and determine the risk thresholds for lung cancer screening is ongoing.

When externally validating the LLP_{v2} and $PLCO_{M2012}$ models using CPRD, O'Dowd and colleagues found that both models underestimated the risk in individuals at low predicted risk but overestimated the risk in individuals at higher risk.[32] Harrell's C statistics for the LLP_{v2} and $PLCO_{M2012}$ models in our two validation cohorts were lower than those in O'Dowd's study (reported as AUC in their study). Possible explanations included study sample selection, sample size, and study period, as well as the availability of information, population and geographical coverage between the QResearch and CPRD databases,

and different ways of handling the variables unavailable from the EHR database. Despite these differences, both studies reached the same conclusion that the LLP_{v2} and $PLCO_{M2012}$ models did not have satisfactory discrimination and were not well calibrated when externally validated using different English primary care datasets. Therefore, they might not be directly applicable to the English primary care population.

Robbins and colleagues evaluated the performance of several models (LLP_{v2}, LLP_{v3}, LCRAT, Lung Cancer Death Risk Assessment Tool [LCDRAT], $PLCO_{m2012}$, and Bach) in current and former smokers aged 40–80 years using three UK cohorts (UK Biobank, EPIC-UK, and the Generations Study) to define the eligibility for lung cancer screening.[16] We did not include LCDRAT in our study, as this model predicts lung cancer death. The AUC values for the models in two subgroups (people aged 40–74 years and 55–74 years; 0·73–0·82) were much higher in the three cohorts than in those in our study. For calibration, all models overestimated risk in all cohorts. The ratio of expected versus observed risk was between 1·20 (LLP_{v3}) and 2·25 (LLP_{v2}). We found both overestimation and underestimation appeared in our study samples. Robbins and colleagues emphasised the importance of validating prediction tools in specific countries—we agree with this point.

Ten Haaf and colleagues mentioned that little attention had been given to sex-specific risk stratification in current practice.[33] We developed and validated our models by sex and found that some of the significant predictors differed between men and women. Even for the same predictor, coefficients and HRs might differ by sex (table 3). Histological subtypes and preclinical duration of lung cancer might vary between sexes as well.[17,33] We also found that the three discrimination measures were often different between sexes in model evaluation. Given all these differences, it might be worthwhile to consider sex-specific risk stratification and screening intervals for lung cancer screening.

Informed by the existing research evidence and clinical expertise, we included as many relevant predictors as possible when we developed the CanPredict (lung) model. Using contemporaneous primary care EHRs to develop and validate risk prediction models is likely to have greater face validity (ie, that information from EHRs appears to measure what it is intended to measure), generalisability, and applicability to the UK primary care population than with study designs such as clinical trials or survey, or data from other countries. Our study benefits from a large sample size and a long duration of follow-up. We also used another non-overlapping data source (CPRD Gold) from geographically distinct general practices to externally validate the CanPredict (lung) model. The populations in the QResearch and CPRD databases are representative of the whole English primary care population,[6,10] in terms of age, sex, ethnicity, socioeconomic deprivation, smoking exposure, and geographical coverage. With local validation and recalibration, our models could also be used internationally.

We would like to consider our CanPredict (lung) model as an inclusive algorithm. It was developed and validated in a wide age range and included all ethnicities and smoking statuses. Our models can calculate an individual's risk of developing lung cancer during 1–10 years follow-up, which is more flexible than models with a fixed prediction horizon such as 5 years for LLP or 6 years for PLCO. The CanPredict (lung) model allows sex-specific risk stratification, which is a unique strength. It also outperformed other prediction models in ever-smokers aged 55–74 years, which means our model is robust and suitable for selecting eligible individuals for lung cancer screening. This shows the potential of our models in clinical application and improving population health.

Finally, we followed good research practice. We pre-registered the research protocol and statistical analysis plan in the public domain. We used robust and advanced statistical methods to develop and validate our models. Lung cancer is one of the most common cancers and this study has a large sample size and number of events per variable, so our model should be accurate and reliable. We thoroughly assessed the model performance of eight different prediction models in different approaches and data sources. We followed the TRIPOD guideline[8,9] to conduct and report this study.

Limitations included the fact that we used routinely collected EHRs to validate other lung cancer prediction models, but did not use information directly collected from patients in a screening setting. Not all relevant information on predictor variables in other prediction models is available from EHRs. Therefore, we needed to make reasonable assumptions and adapt the situation for the EHR in the English population when we calculated the risk scores for other prediction models. The unavailable information and the way we handled the variables (making similar assumptions to another study[32]) might inaccurately estimate the risk scores for individuals in some models, which could either overestimate or underestimate the risk. This might consequently influence the evaluation of model performance for some models. It might be possible for us to conduct a new study to compare the CanPredict (lung) model using data from primary care records with the LLP_{v2} and $PLCO_{M2012}$ models using data collected and the scores calculated from the Targeted Lung Health Check programme in the future, using the infrastructure and data bank of the DART project.

We intend to make the CanPredict (lung) model publicly available, subject to further funding for implementation and Medicines and Healthcare Products Regulatory Agency medical device compliance. The CanPredict (lung) model can also be implemented in primary care computer systems, which allow batch-mode processes that use existing information in EHRs at each practice to facilitate

the selection of eligible individuals at high risk for lung cancer screening. Compared with using questionnaires to collect information from participants and calculating risk scores to check their eligibility, the batch-mode process is more efficient, especially when the lung cancer screening programme rolls out at scale nationally. It can not only substantially reduce human resources and costs, but also save time and streamline the administrative process for better patient experience and increased patient satisfaction. Furthermore, it could facilitate patient–general practitioner discussion about the risks and benefits of lung cancer screening for individual patients. This could improve patients' awareness of their health status and risk level, which in turn might increase their willingness to participate in the screening programme or lead to behavioural changes such as considering smoking cessation.

Developed and internally and externally validated using primary care EHRs, the CanPredict (lung) model can estimate an individual adult's risk of lung cancer diagnosis for up to 10 years. It has the best performance among other prediction models for lung cancer in discrimination and calibration for both sexes across three prediction horizons (5, 6, and 10 years). It also has the highest net benefit. The CanPredict (lung) model is suitable for risk stratification of the English primary care population and for selecting individuals at high risk for targeted lung cancer screening in the UK.

Contributors

FVG and JH-C secured the funding for this study. FVG is the chief investigator of the DART project, and JH-C is the joint package lead (WP6—primary care, population health, and health economics). JH-C and WL contributed to the study conceptualisation. JH-C developed the clinical code lists and worked with WL on the data specification. WL led on ethical approval. JH-C and WL designed the statistical analysis plan. WL drafted the whole research protocol and statistical analysis plan, with methodological input from JH-C and CACC, and clinical and contextual input from JH-C, DRB, FVG, and JB. JH-C undertook the data extraction, data cleaning, developed the prediction model, and undertook the initial validation on QResearch with input from CACC. WL undertook analyses on the seven prediction models and validated in QResearch and CPRD datasets. WL and JH-C accessed and verified all the data in the study, wrote the codes, performed the analyses, and take responsibility for the integrity of the data and the accuracy of data analysis. CACC contributed substantially to the statistical methodology. WL and JH-C drafted the paper. All authors contributed to the interpretation of the results and revision of the manuscript, approved the final version of the manuscript, and had final responsibility for the decision to submit for publication. JH-C is the guarantor for the study.

Declaration of interests

JH-C is an unpaid director of QResearch, a not-for-profit organisation in a partnership between the University of Oxford and EMIS Health, who supply the QResearch database for this work. JH-C is also a founder and shareholder of ClinRisk, who produce open-source and closed-source software to implement clinical risk algorithms, and was its medical director until May 31, 2019. FVG is a shareholder of Optellums Ltd, an AI company that produces diagnostic algorithms for nodules on CT scans, mainly in lung cancer, and received honoraria from Roche. But these are unrelated to this study. DRB received honoraria from Astra Zeneca, Roche, Bristol Myers Squibb, and MSD, which is not related to this study. CACC received payment from previous consultancy with ClinRisk Ltd, which is outside of the current work. WL and JB have no interests to declare.

Data sharing

To guarantee the confidentiality of personal and health information of patients, only the named authors have had full access to the data during the study, in accordance with the relevant licence agreements. Information on access to the QResearch data is available on the QResearch website (www.qresearch.org).

Acknowledgments

This study is funded by Innovate UK (UK Research and Innovation, grant reference: 40255). We thank the two lay members from the Roy Castle Lung Cancer Foundation who reviewed our lay summary of the DART–QResearch Project for ethical approval and provided very helpful feedback. QResearch received funding from the NIHR Biomedical Research Centre (Oxford), John Fell Oxford University Press Research Fund, Cancer Research UK (Grant number C5255/A18085), through the Cancer Research UK Oxford Centre, the Oxford Wellcome Institutional Strategic Support Fund (204826/Z/16/Z). We acknowledge the contribution of the patients and the general practices who contribute to QResearch and Egton Medical Information Systems (EMIS) Health, the University of Nottingham, and the Chancellor, Masters, and Scholars of the University of Oxford for expertise in establishing, developing, and supporting the QResearch database. The Hospital Episode Statistics data used in this analysis are copyright (2022) to the Health and Social Care Information Centre and re-used with the permission of the Health and Social Care Information Centre and the University of Oxford (all rights reserved). This project involves data derived from patient-level information collected by the National Health Service (NHS), as part of the care and support of cancer patients. The data is collated, maintained, and quality assured by the National Cancer Registration and Analysis Service, which was part of Public Health England (PHE). Access to the data was facilitated by the PHE Office for Data Release, which becomes part of NHS Digital and subsequently NHS England. The death registration data are provided by the Office for National Statistics. NHS England and the Office of National Statistics bear no responsibility for the analysis or interpretation of the data. The views expressed in this manuscript are those of the authors.

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:** 209–49.

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

3 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:** 503–13.

4 Moyer VA. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *Ann Intern Med* 2014; **160:** 330–38.

5 Krist AH, Davidson KW, Mangione CM, et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *JAMA* 2021; **325:** 962–70.

6 Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44:** 827–36.

7 Liao W, Burchardt J, Coupland C, Gleeson F, Hippisley-Cox J, DART initiative. Development, validation, and evaluation of prediction models to identify individuals at high risk of lung cancer for screening in the English primary care population using the QResearch database: research protocol and statistical analysis plan. *medRxiv* 2022; published online Jan 7. https://doi.org/10.1101/2022.01.07.22268789 (preprint).

8 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015; **162:** 55–63.

9 Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162:** W1–73.

10 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017; **357:** j2099.

11 Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* 2015; **5:** e007825.

12 Rubin DB. Multiple imputation for non-response in surveys. New York: John Wiley, 1987.

13 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999; **28:** 964–74.

14 Toumazis I, Bastani M, Han SS, Plevritis SK. Risk-based lung cancer screening: a systematic review. *Lung Cancer* 2020; **147:** 154–86.

15 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:** e1002277.

16 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:** 2026–34.

17 Ten Haaf K, van Rosmalen J, de Koning HJ. Lung cancer detectability by test, histology, stage, and gender: estimates from the NLST and the PLCO trials. *Cancer Epidemiol Biomarkers Prev* 2015; **24:** 154–61.

18 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:** 270–76.

19 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:** 161–68.

20 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:** 2300–11.

21 Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. *N Engl J Med* 2013; **368:** 728–36.

22 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:** e1001764.

23 Wilson DO, Weissfeld J. A simple model for predicting lung cancer occurrence in a lung cancer screening program: the Pittsburgh Predictor. *Lung Cancer* 2015; **89:** 31–37.

24 Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003; **95:** 470–78.

25 Newson RB. Comparing the predictive powers of survival models using Harrell's C or Somers' D. *Stata J* 2010; **10:** 339–58.

26 Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23:** 723–48.

27 Royston P. Explained variation for survival models. *Stata J* 2006; **6:** 1–14.

28 Ensor J, Snell KI, Martin EC. PMCALPLOT: Stata module to produce calibration plot of prediction model performance. Statistical Software Components S458486. Chestnut Hill, MA: Boston College Department of Economics, 2018.

29 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019; **3:** 18.

30 Hippisley-Cox J, Coupland C. Predicting the risk of prostate cancer in asymptomatic men: a cohort study to develop and validate a novel algorithm. *Br J Gen Pract* 2021; **71:** e364–71.

31 Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; **24:** 198–208.

32 O'Dowd EL, Ten Haaf K, Kaur J, et al. Selection of eligible participants for screening for lung cancer using primary care data. *Thorax* 2022; **77:** 882–90.

33 Ten Haaf K, van der Aalst CM, de Koning HJ, Kaaks R, Tammemägi MC. Personalising lung cancer screening: an overview of risk-stratification opportunities and challenges. *Int J Cancer* 2021; **149:** 250–63.