

Selection of Eligible Participants for Screening for Lung Cancer using Primary Care Data

Emma L O'Dowd^{1,5} Consultant Respiratory Physician, Honorary Assistant Professor

Kevin Ten Haaf² Post-doctoral researcher

Stephen W. Duffy³ Professor of Cancer Screening

William Hamilton⁴ Professor of Primary care Diagnostics

Richard Hubbard⁵ British Lung Foundation Professor of Epidemiology

John K Field⁶ Professor of Molecular Oncology

Matthew E Callister⁷ Consultant respiratory physician, Honorary Professor of Medicine

Samuel M Janes⁸ Professor of Respiratory Medicine

Harry de Koning² Professor of Public Health and Screening

Jannette Rawlinson⁹ Patient Advocate

David R Baldwin^{1,5*} Consultant Respiratory Physician, Honorary Professor of Medicine

*author for correspondence

1 Nottingham University Hospitals NHS Trust, Nottingham UK

2 Erasmus MC, Rotterdam, Netherlands

3 Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK

4 University of Exeter Medical School, Exeter, UK

5 University of Nottingham, Nottingham, UK

6 Department of Molecular and Clinical Cancer Medicine, Institute of Systems, Molecular & Integrative Biology, University of Liverpool, UK

7 The Leeds Teaching Hospitals NHS Trust, Leeds, UK

8 Lungs for Living Resarch Centre, Division of Medicine, University College London, London, UIK

9. Patient Advocate committees of the British Thoracic Oncology Group, National Cancer Research Inistitute Lung subgroup and European Lung Foundation, Sandwell, UK

Abstract

Screening for lung cancer with low radiation dose CT is effective if offered to people at increased risk of the disease. Target populations can be defined by multivariable models that predict risk. Two externally validated models, the Liverpool Lung Project_{v2} (LLP_{v2}) and the Prostate Lung Colorectal and Ovarian_{m2012} PLCO_{m2012} are recommended to select participants for the National Health Service targeted Lung Health Check programme in England. Currently, direct contact with potential participant is required to obtain data for the models. A way to reduce the number of ineligible people contacted might be to apply the models directly to digital primary care data but how the models perform in this setting is not known.

Method

The Clinical Practice Research Datalink (CPRD), a computerised, longitudinal primary care database, was used to test both of the models. Lung cancer occurrence over 5 or 6 years was measured in a cohort aged 50 to 80 years of ever smokers and compared with the predicted risk over 5 years (LLP_{v2}) and 6 years (PLCO_{m2012}).

Results

Over 5 and 6 years there were 7123 and 7876 lung cancers respectively from a cohort of 834986. After recalibration of the model intercept, LLP_{v2} produced a c-statistic of 0.700 (0.694 to 0.710). Calibration plots showed an under-prediction of lung cancer cases at the lowest risk scores, followed by an over prediction (overall mean predicted risk of lung cancer 4.61%, actual risk of 0.9%). After imputing missing BMI and smoking intensity values, PLCO_{m2012} produced a c-statistic of 0.679 (0.673 – 0.685) in CPRD data. PLCO_{m2012} also over-estimated risk (mean predicted risk: 3.76%). Setting thresholds at <1% for LLP and < 0.15% for PLCO, 15.6 and 11.4% of ever smokers who developed lung cancer would not have been contacted for screening eligibility assessment, however this also avoids contacting 46% and 34.5% of the cohort.

Conclusion

Two validated multivariable models show only moderate discrimination when applied to routinely collected primary care data restricted to ever smokers aged 50 to 80 years, which may be explained by the quality and completeness of the data. However, they may be used as a way to substantially reduce the total number of people who are contacted as part of a screening programme accepting that 10-15% of ever-smoking people aged 50 to 80 years who develop lung cancer may not have the chance to be evaluated for screening eligibility. This will be an important consideration in health economic analyses. Further work is needed to establish to what extent newer models can improve on performance in primary care data.

INTRODUCTION

Randomised controlled trials have shown that screening with low dose computed tomography (LDCT) reduces lung cancer mortality.(1-3) Many countries are therefore planning implementation but questions remain around how to identify the population most likely to benefit. Most lung cancer screening trials used age and smoking pack year criteria to select participants.However, since the publication of the National Lung Screening Trial (NLST), further analyses have demonstrated that substantial variations in risk exist within trial populations.(4) Risk prediction models have been suggested to select eligible participants at high risk of lung cancer and have been shown to be more sensitive and specific compared with using age and smoking history alone.(5-10) This may be due in part to these models incorporating more detailed smoking data and considering other risk factors such as chronic obstructive pulmonary disease (COPD) or asbestos exposure. The United Kingdom Lung Screening Trial (UKLS) used a multi-factorial risk prediction model (Liverpool Lung Project Version 2 – LLP_{v2}) to select patients. (11, 12) The results of the UKLS showed that the proportion of patients in whom lung cancer was detected was similar in a single screening round to that achieved by three annual rounds in the NLST. However, the trade-off of selecting higher risk groups is that only a small proportion of the total population at risk of lung cancer is included and there is potential for those selected to be at greater risk of competing causes of death. To maximise the impact of a screening programme, models with better sensitivity and specificity are needed to ensure the greatest number of eligible people benefit whilst reducing the number of LDCTs required. As well as being accurate, models for use on a whole population need simple methods of data collection, or must use existing high-quality data.

In the United Kingdom (UK) primary care records have been used to identify ever smokers for further risk stratification are a way to limit the number of approaches that have to be make to cover the target population.(13, 14) Although, inevitably, a small proportion may be missed, a much larger proportion of ineligible people are not approached and with less inconvenience, worry and cost. UK pilots have used both the LLP_{v2}, at a threshold of either 2.5% or 5%, and the Prostate Lung Colorectal and Ovarian

(modified 2012) (PLCO_{m2012}) at a threshold of 1.51%. An earlier version of this model was used to select subjects for the Pan-Canadian Early Detection of Lung Cancer (PanCan) study.(7) The UK pilots found baseline cancer rates of 2-3% and the PanCan study 5%.(5) The National Health Service (NHS) England Lung Health Check targeted screening programme has therefore recommended using either a PLCO_{m2012} 6-year risk-threshold for lung cancer of 1.51% and/or an LLP_{v2} 5-year risk-threshold of 2.5% to define eligibility, but these risk models have not been validated or calibrated in primary care data. Previous external validations have compared models in well-defined data derived from screening trials.(15, 16) A recent “real-world” UK pilot screening programme found that the PLCO_{m2012} model performed much as expected, although the investigators found some degree of miscalibration.(17) However, their population had received screening, which may partly account for this miscalibration. It is therefore important to understand whether models used in selection can be used in routinely collected primary care data of non-screened individuals and to establish the most appropriate risk threshold for further evaluating screening eligibility.

METHODS

DATA SOURCE

We used data from the Clinical Practice Research Datalink (CPRD), a computerised, longitudinal primary care database, linked to a range of other health related data to provide a representative UK population health dataset. The data encompass 50 million patients, including 14 million who are currently registered. (18) All symptoms, medical diagnoses, prescriptions, investigations and results are entered into the computer system either during a consultation with a general practitioner (GP) or following communication from other healthcare providers.

PATIENT DATA

A general population cohort ≥ 40 years of age who were registered and contributing data for at least 12 months between 1st January 2000 and 31st December 2015 was extracted from CPRD. Patients ≥ 40 years of age who were diagnosed with lung cancer during this timeframe were identified from this cohort. To ensure that these were incident rather than prevalent cases, we excluded patients who registered less than 12 months prior to their diagnosis date. Data for English patients were linked to Cancer Registry data which provided additional information, including lung cancer pathological subtype and stage at diagnosis.

LUNG CANCER PREDICTION MODELS

This study externally validates two lung cancer prediction models: the LLP_{v2} and the PLCO_{m2012}. (5, 8, 19) CPRD data were used to identify and categorise the required variables to fit the models and derive a risk score for 5-year (LLP_{v2}) and 6-year (PLCO_{m2012}) risk score for lung cancer respectively. Personal history of pneumonia, COPD, smoking status, any cancer and family history of cancer were identified using medical code lists. Asbestos exposure is not routinely available in CPRD and to avoid bias by assuming that all patients were not exposed to asbestos, we searched CPRD for medical codes indicating 'asbestosis'. (20) Data on ethnicity and education were not available so we assumed all patients to be white and have basic education (i.e. assuming normal secondary school completion in

the UK approximates to completing high school in the US). Additional CPRD files were used to extract data on body mass index (BMI). LLP_{v2} incorporates age at lung cancer diagnosis for a first degree relative, however details on the type of cancer in family members and age at diagnosis are not routinely collected in CPRD and therefore any lung cancer in a first degree relative was considered to be early onset (age <60). Models were also assessed with this variable excluded.

SMOKING DATA IN CPRD

Unlike trial data which record detailed individual smoking data at the time of a risk assessment, smoking data in CPRD are recorded whenever the person visits the GP. The GP records the details using medical and Read codes to indicate a patient's smoking status; current, ex or never, and the intensity of smoking in categories. These categories are defined as: 1) very heavy smoker 40+ cig/day, 2) heavy smoker 20-40 cig/day, 3) moderate smoker 10-19 cig/day, 4) light smoker 1-9 cig/day, 5) trivial smoker <1 cig/day and 6) smoker quantity unknown. In the PLCO_{m2012} model, smoking intensity is fitted as a continuous variable (7) so in order to fit the model we had to convert the categorical variable to a specific number of cigarettes smoked per day. Therefore we assumed very heavy smokers to have smoked 40 cig/day, heavy smokers 20 cig/day, moderate smokers 10 cig/day, light smokers 5 cig/day and trivial smokers to have smoked 2 cig/day. Patients with missing smoking data throughout their follow up were considered to be never smokers.

Only 10% of the population who were categorised as ever smokers had a documented age of starting smoking. Based on published literature, we assumed the age at which people started smoking to be 18 years, which also coincides with the legal age to buy cigarettes in UK from 2007 (21-23). Sixty eight per cent of ex-smokers had a date of smoking cessation recorded in the additional CPRD files. Median day difference between smoking cessation date and the risk assessment date was calculated and substituted for 32% of the ex-smokers with missing date of smoking cessation.

DATA SETUP

LLP_{v2} and PLCO_{m2012} predict 5 and 6 year lung cancer incidence respectively. We calculated 5- and 6-year risk scores for all CPRD patients registered on 1st January 2009. We looked at 5 year incidence of having lung cancer for LLP_{v2} model, i.e. until 31st December 2013; and 6 year incidence of having lung cancer for PLCO_{m2012} model, i.e. until 31st December 2014. Lung cancer screening is unlikely to be offered to people aged below 50 years or above 80 years based on current modelling: therefore we excluded people aged <50 years or >80 years at the point of the risk assessment (1st January 2009). Similarly lung cancer screening is unlikely to be offered to never smokers and so only ever0-smokers were included in the cohort. This resulted in 842,109 individuals in our CPRD cohort.

STATISTICAL ANALYSIS AND MULTIPLE IMPUTATION

All data management and statistical analysis were performed using STATA version 16 (StataCorp) and the study was conducted and reported in line with the Transparent Reporting of a multivariate prediction model for Individual Prediction or Diagnosis (TRIPOD) guidelines (24). Occurrence of lung cancer was treated as a binary outcome at 5 years for LLP and 6 years for PLCO_{m2012}. Distributions of demographic variables between lung cancer patients and non-lung cancer patients were evaluated. All patients actively participating in CPRD on 1st January 2009 were used to assess the performance of LLP (n=842,109). Multiple imputation by chained equation (MICE), to replace missing data on BMI (10%) and smoking quantity (28%), was performed based on all candidate predictors before applying the PLCO_{m2012} model. We created 5 imputed datasets for our cohort and combined them using Rubin's rule to obtain final model estimates.(25) On the basis of the most conservative figure of 4,115 lung cancer events (exclusion of all possible duplications) during the six year post risk assessment and eleven risk predictors in PLCO_{m2012}, we had a sample size of 374 lung cancer diagnoses per predictor, well above the minimum requirement of 10 events per predictor suggested by Peduzzi et al (26). We compared the demographics of the CPRD derived dataset with that of the original development sets for the LLP and PLCO_{m2012}.(7, 19)

We assessed the performance of the models in terms of discrimination and calibration plots. (6, 27). The area under the receiver operating curve (AUC) was used to assess discrimination, ranging from 0.5 indicating no discrimination to 1.0 indicating perfect discrimination. 'Pmcalplot' package on STATA was used to plot observed and predicted risk probabilities. We also assessed the performance of models by risk-thresholds and calculated the number needed to screen to identify 1 lung cancer patient based on those risk-threshold figures. For the LLP_{v2} model the risk-quartiles were set at risks: <1%, 1% to <2.5%, 2.5% to <5% and 5% or greater, while for PLCO_{m2012}, our cohort was divided into risk tertiles of risks: <0.15%, 0.15% to 1.5% and greater than 1.5%. The values of the considered risk thresholds for the two models differ, as the models differ in absolute risk estimates due to differences in risk-levels between their development datasets. For comparison we also calculated AUCs for each model using data from NLST and PLCO.

ETHICAL APPROVAL

Approval for use of data for this project was granted by the CPRD Independent Scientific Advisory Committee (ISAC) (Protocol numbers 18_223 and 20_014R).

RESULTS

STUDY PARTICIPANTS AND COMPARISON WITH ORIGINAL MODELS

We analysed data on all (n=5,997,270) people actively contributing to CPRD between 1st January 2000 and 31st December 2015. Lung cancer incidence was 85.8 per 100,000 person-years. The overall incidence was higher for men compared with women (98.5 vs 73.4 per 100,000 person-years). People aged between 50 and 80 years who were ever smokers were selected to form the validation population for LLP_{v2} and PLCO_{m2012} models. This comprised 842,109 participants. An overview of the demographics / model characteristics of the LLP_{v2} and PLCO_{m2012} cohorts and the CPRD cohort is presented in Tables 1a and 1b. Complete information for all risk factors was available for 100% of the population for LLP_{v2} validation but only 66% had complete information for PLCO_{m2012}, mainly due to missing data on BMI and smoking intensity.

CPRD COHORT AND ORIGINAL LLP COHORT DEMOGRAPHIC COMPARISON

The median age of the lung cancer patients was higher in the CPRD cohort compared with LLP (69 years vs 66 years respectively) and the median age of controls was lower in CPRD compared with LLP (61 years vs 66 years).(19) The CPRD group also had a higher proportion of female lung cancer cases compared with LLP (43% vs 38%). Even using a more inclusive definition of family history of any cancer and asbestosis as a surrogate for asbestos exposure, patients identified with those risk factors in CPRD were limited (family history of any cancer in CPRD = 0.1% vs LLP cohort = 21% and asbestosis in CPRD = <1% vs asbestos exposure in LLP cohort = 35% in cases). The full comparison is provided in Table 1a.

CPRD COHORT AND PLCO_{m2012} COHORT DEMOGRAPHIC COMPARISON

The PLCO_{m2012} model used information on 80,672 participants from the PLCO study who were ever smokers. Lung cancer cases in CPRD were older than both the NLST and PLCO groups (median age 69 years vs 64 years and 65 years respectively).(7) Ten per cent of the BMI data was missing in CPRD patients (Table 1b), but the median BMI was the same for cases and controls for all of the cohorts. CPRD patients also had a low recording of family history or previous personal history of malignant

cancer compared with NLST and PLCO cohorts, reflecting the poor recording of family level data in primary care records. Twenty seven per cent of NLST participants had a self-reported history of COPD, which was more than both CPRD and PLCO cohorts (22% and 20% respectively). Data on ethnicity and education status were not routinely recorded at the time of data extraction in CPRD data, so they are not provided in Table 1b. Only 19% of non-lung cancer participants in PLCO were current smokers compared with 43% in CPRD and 47/ 48% in the NLST CT and CXR arms. A detailed comparison of smoking data between the cohorts is provided in Table 1b.

COMPARISON OF RISK PREDICTION MODEL PERFORMANCE IN CPRD

LLP_{v2}

In CPRD, 7123 lung cancer events took place in 5 years between 1st January 2009 and 31st December 2013. The original LLP_{v2} model, which included never smokers, produced a c-statistic of 0.70 in 10-fold cross validation.(19) After recalibration of the model intercept, the external validation of LLP_{v2} produced a c-statistic of 0.700 (0.694 to 0.710) in CPRD data (Table 2). The calibration plot of the recalibrated model is shown in Figure 1. There was an under-prediction of lung cancer cases at the lowest risk scores, followed by an over prediction. The overall mean predicted risk of lung cancer patients in the CPRD cohort was 4.61%. This compares with the actual risk of 0.9%.

Table 3 shows the patient features, proportion of lung cancer patients identified and number of individuals needed to screen to detect one lung cancer patient using a variety of risk categories. Lung cancer patients had a higher mean predicted risk score compared with non-lung cancer cases in each category. Approximately 71% of the lung cancer patients had a predicted risk score of >2.5%. The number of individuals needed to screen to detect 1 cancer (NNS) ranged from 322 in individuals with a risk of <1% to 54 in individuals with a risk >5%. A risk threshold of >5% included 43.7% of lung cancers and 20% of the total cohort. The corresponding figures for >2.5% and >1% were 70.8% of cancers and 40.8% of the cohort, and 84.5% of cancers and 57.5% of the cohort. Setting a risk threshold of 1% gives a NNS of 80, but would still miss 15.6% of the lung cancer cases, chiefly those with a younger median

age (56 years) and shorter duration of smoking (all ≤ 40 years duration). However, 43.5% of the cohort would not need to be screened

PLCO_{m2012}

We identified 7,876 lung cancer events that took place in the 6 years following PLCO_{m2012} risk assessment on 1st January 2009. After imputing missing BMI and smoking intensity values PLCO_{m2012} produced a c-statistic of 0.679 (0.673 – 0.685) in CPRD data. Furthermore, even following recalibration of the model intercept (Figure 2) there was still poor calibration of PLCO_{m2012} in CPRD data. The overall mean predicted risk for lung cancer patients by PLCO_{m2012} model in the cohort was 3.76%. Similar to LLP_{v2}, PLCO_{m2012} under-predicted lung cancer cases at the lowest risk scores, followed by over prediction.

Table 4 shows the patient features, proportion of lung cancer patients identified and number of individuals needed to screen to detect one lung cancer patient using a variety of risk thresholds. Using imputed data, if a PLCO_{m2012} risk-threshold of $>1.51\%$ were applied to the CPRD population, it would detect 48.6% of the total lung cancer cases, with a NNS of 52 (23.5% of the total cohort selected). Setting the risk threshold to 0.15% increases the NNS to 88 (72.8% total cohort). This misses 11.41% of lung cancer cases, predominantly those with the lowest smoking intensity or where smoking data were incorrectly recorded in CPRD (Table 4). However, 31.8% of people would not need to be screened. Those in the highest risk threshold group ($>1.5\%$) had a higher median age (71 years vs 66 years for $<0.15\%$) and were more likely to be current smokers (63% vs 36% respectively). Proportionally more in the highest risk threshold also had a diagnosis of COPD (29% vs 14%).

DISCUSSION

MAIN FINDINGS

This is the first study to validate and recalibrate the LLP_{v2} and PLCO_{m2012} models using primary care data. After restricting the primary care data to include only ever smokers aged 50 to 80 years, our work showed that discrimination was only moderate for both models and, following recalibration of the model intercept, there was still poor calibration of the PLCO_{m2012} model in CPRD. Both models showed under-prediction at low risk followed by an over-prediction of those at highest risk. The detail required to use these models in practice is considerable (particularly with regards to smoking data) and would require a face to face or telephone consultation in order to augment the data already held in primary care records. Using both models at the current suggested risk thresholds (>1.51% for PLCO_{m2012} and >2.5% or >5% for LLP_{v2}) missed 51%, 29% and 56% of lung cancer cases respectively. This concerned largely those with younger median age and lower smoking duration for LLP. Those who were missed by PLCO_{m2012} at lower thresholds were less likely to have COPD and were more likely to be ex-smokers with lower smoking intensity. Both models showed higher AUCs in PLCO data, reflecting the wider spread of lung cancer risk (inclusion of never smokers).

STRENGTHS AND LIMITATIONS

This is the largest external validation of the LLP_{v2} and PLCO_{m2012} risk models in the literature and, to our knowledge, the first using primary care data. Data in CPRD are prospectively recorded at the time of consultation in primary care which minimises reporting and recall bias, however, the information relies on accurate coding and timely data entry in primary care. To minimise errors related to this, we only used data entered by practices after the practice met the CPRD data quality and completeness standard. This study has tested the risk models in a context outside of their intended use. Both models should be populated with data collected from a potential participant in screening. Instead the study shows how the models perform in routinely collected primary care data. Therefore, one of the key limitations is the lack of detailed smoking data in CPRD. One could argue that this places the PLCO_{m2012} model at a disadvantage in validation, as the primary risk factors which drive the model are age and

detailed smoking history. However, one of the objectives of this piece of work was not only to validate and calibrate the models in primary care data, but also to look at the feasibility of running these models in routinely collected data to select patients for entry into screening. Smoking data in primary care are recorded at the time of registration at a surgery using a questionnaire or during face to face consultations. Many practices record smoking status at regular intervals as part of the Quality Outcome Framework.(28) Studies looking at the validity of smoking records in electronic primary care data have shown that it is in line with that obtained from population surveys such as the Health Survey for England in terms of proportion of people who are current or ex-smokers in age categories. (29)Those with no recorded smoking status are likely to be never smokers or smokers who quit before the age of 30 years, so it is unlikely that we would be excluding or misclassifying a substantial number of eligible smokers by labelling these people as never smokers.(30-32) As smoking intensity is grouped into categories in CPRD we had to assign each participant a specific number of cigarettes smoked per day, which was largely in multiples of 5 or 10 (apart from trivial smokers). Work by Shiffman has shown that even when a contemporaneous smoking history is taken from a person it is prone to digit bias.(33) In his study, two thirds of participants asked about daily smoking consumption recorded smoking quantity in multiples of 10, suggesting that our approach is not unreasonable. The key drivers in these risk models are age, sex and smoking, so missing data on other predictor variables is less likely to impact the performance of the risk models. Thus, although it is easy to criticise data completeness and accuracy in routinely collected primary care data, the reality is that it is often better than assumed. Therefore, using these data has to be balanced against the extra cost of directly-acquired data which itself may be subject to incompleteness and bias.

Recently the LLP_{v2} risk model and a recalibrated version (LLP_{v3}) have been validated and calibrated using questionnaire data from the 75,958 UKLS individuals who responded to the first approach questionnaire and have been followed up for lung cancer for over 5 years(34). This cohort included never smokers (47%). The authors found the AUC to be 0.81 for both LLP_{v2} and LLP_{v3} but LLP_{v2} was found to overestimate the absolute risk approximately two fold. The LLP_{v3}, which was calibrated to

contemporary English incidence, achieved substantially more accurate prediction of absolute incidence, and would now be an appropriate update to LLP_{v2} in selecting a high-risk group for screening in the UK.

OTHER WORK IN THE LITERATURE

A study by Li et al (16) compared the performance of four risk prediction models including LLP and PLCO_{m2012} in 20,700 German participants of the European Prospective Investigation into Cancer and Nutrition cohort. This showed better discrimination for the PLCO_{m2012} model (c-index 0.81, 95% CI 0.76 – 0.86) compared with the LLP model (c-index 0.79, 95% 0.73 – 0.83). However, the cohort had an overall rate of lung cancer of less than 0.5% with fewer than 100 lung cancer events.

Weber et al (35) externally validated PLCO_{m2012} in a cohort of 95,882 Australian ever-smokers aged 45 years and older. They used questionnaire data completed as part of the 45 and Up Study (36), linked to a number of population datasets. They demonstrated an AUC of 0.80 (95% CI, 0.78–0.81) with good calibration in their population (mean and 90th percentile absolute difference between observed and predicted probabilities of 0.006 and 0.016, respectively). The authors assessed the model performance at a risk threshold of 1.51% and showed a sensitivity of 70% (95% CI 67.1–72.7) and specificity of 75.4% (95% CI 75.2–75.7). In a subset of the population (those aged 55-74 years) they also assessed a variety of additional risk thresholds, namely 1.49%, 1.73% and 2% but did not show that changing the risk threshold made a substantial improvement to the sensitivity and specificity. The 45 and Up Study cohort may be more similar to those who participate in trials, from which PLCO_{m2012} was derived, as people had to personally complete a questionnaire, consent form and mail it to the Study centre to be included. As the authors acknowledge, this means that there may be a selection bias in favour of less deprived people. The good calibration suggests that the population is similar to that from which the model was derived. The data required to compute the risk score, particularly with

regards to detailed smoking data, were largely derived from these questionnaires rather than already available in routinely collected data.

Ten Haaf et al (6) conducted a retrospective validation of 9 risk prediction models using data from NLST and PLCO. Both calibration and discriminative ability was better for all models using PLCO data than NLST. PLCO_{m2012} showed better discrimination than LLP (0.789 (95% CI 0.781-0.797) vs 0.745 (95% CI 0.736-0.755) in the PLCO control arm) but the PLCO_{m2012} was derived from this dataset which places PLCO_{m2012} at an advantage compared with LLP_{v2}. Interestingly, most of the models tested in this study had greater discriminative ability in predicting 6-year lung cancer mortality rather than 6-year lung cancer incidence.

Katki et al (9) evaluated 9 risk prediction models in US data on ever smokers from the National Institutes of Health–AARP Diet and Health Study (NIH-AARP) and the Cancer Prevention Study II Nutrition Survey (ACS CPSII) cohort to compare model performance. Both LLP and PLCO_{m2012} showed some overestimation of risk but PLCO_{m2012} was better calibrated in this cohort. Both showed moderate discrimination; PLCO_{m2012} with AUC of 0.769 (95% CI 0.766-0.772) and 0.754 (95% CI 0.741-0.767) for NIH-AARP and ACS CPSII respectively and LLP with AUC values of 0.726 (95% CI 0.722-0.731) and 0.726 (95% CI 0.711-0.740)) respectively. When the authors set screening eligibility at 2% lung cancer risk over 5 years the well calibrated models, including PLCO_{m2012} selected fewer participants for inclusion (ranges between 7.6-10.9 million, compared with 14.5-26 million for the less well calibrated models).

CLINICAL RELEVANCE AND CONCLUSIONS

Primary care records in the UK are likely to be used to identify those who may be eligible for CT screening because they provide an efficient way to identify ever-smokers and thus minimise contact with people who would not be eligible and reduce cost and potential distress from being contacted about cancer screening when there is no benefit. Other countries, where similar data exist will likely do the same. Ideally, a model with good sensitivity and specificity should be applied directly to primary

care data and only then would potential participants be contacted. We have shown that two existing models, even at very low risk thresholds, would miss a significant number of people if applied in this way. We do not know how much better the models perform when applied in the lower risk categories to more detailed, directly-derived data from participant questionnaires but it is likely that a significant proportion of people who develop cancer would be below the threshold. If models are to be used to derive a first-step “enriched” population, then the second step would likely involve increasing the risk threshold to comply with cost-effectiveness standards. The principle of the two-step approach is the use of an initial model at a low risk threshold in order to maximise sensitivity, with a second model that uses the integrity of detailed and directly-acquired data to improve specificity and reduce cost. This study has tested models at different thresholds and we conclude that specificity in the first step of the two-step approach would only be improved by obtaining more accurate data to use in the risk prediction, or by the development of new models. Obtaining better data in a first step approach could place a considerable burden on services with limited gain and extra cost. However, once national screening programmes are in place, this could be the subject of data quality improvement in primary care, with additional data fields completed that are important in risk prediction e.g. detail of family history. It will be important to compare the performance of two-step approaches with newer, single step models developed in primary care data. The main value of improved models is in identifying those who are at lower or intermediate risk on the basis of current risk models, but who arguably may have more life-years to gain from screening due to younger age, lower smoking intensity and consequently fewer co-morbidities. It is key that future risk prediction models are able to predict not only eligibility for entry into screening but also whether, and by how much, participants can expect to benefit. Some studies have suggested that risk models may identify those who are less likely to benefit from screening due to competing causes of mortality and morbidity.(37) Optimal risk thresholds need to be identified based on local population data and further work is needed to determine what the best strategy is for identifying and inviting those who have most to gain from screening for lung cancer.

In conclusion, two validated multivariable models perform less well than previously reported when applied to routinely collected primary care data restricted to ever smokers aged 50 to 80 years, which may be explained by the quality and completeness of the data. However, they may be used as a way to reduce the total number of people in this higher risk group who are contacted as part of a screening programme by a third to a half but with 10-15% of people who develop lung cancer excluded from more detailed evaluation. Whilst many of these excluded people may not be at high enough risk to be eligible, further work is needed to establish how many are incorrectly excluded and to what extent newer models can improve on this, as even the best models will miss some lung cancers.

REFERENCES

1. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine*. 2011;365(5):395-409.
2. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med*. 2020;382(6):503-13.
3. Pastorino U, Silva M, Sestini S, Sabia F, Boeri M, Cantarutti A, et al. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Annals of Oncology*. 2019;30(10):1672.
4. Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, Korch M, et al. Targeting of low-dose CT screening according to the risk of lung-cancer death. *N Engl J Med*. 2013;369(3):245-54.
5. Tammemagi MC, Schmidt H, Martel S, McWilliams A, Goffin JR, Johnston MR, et al. Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study. *Lancet Oncol*. 2017;18(11):1523-31.
6. Ten Haaf K, Jeon J, Tammemagi MC, Han SS, Kong CY, Plevritis SK, et al. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. *PLoS Med*. 2017;14(4):e1002277.
7. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. *New England Journal of Medicine*. 2013;368(8):728-36.
8. Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, et al. Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. *Ann Intern Med*. 2012;157(4):242-50.
9. Katki HA, Kovalchik SA, Petito LC, Cheung LC, Jacobs E, Jemal A, et al. Implications of Nine Risk Prediction Models for Selecting Ever-Smokers for Computed Tomography Lung Cancer Screening. *Ann Intern Med*. 2018;169(1):10-9.
10. Ten Haaf K, Bastani M, Cao P, Jeon J, Toumazis I, Han SS, et al. A Comparative Modeling Analysis of Risk-Based Lung Cancer Screening Strategies. *J Natl Cancer Inst*. 2020;112(5):466-79.
11. Baldwin DR, Duffy SW, Wald NJ, Page R, Hansell DM, Field JK. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. *Thorax*. 2011;66(4):308-13.
12. Field JK, Duffy SW, Baldwin DR, Whynes DK, Devaraj A, Brain KE, et al. UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. *Thorax*. 2016;71(2):161-70.
13. Quaife SL, Ruparel M, Dickson JL, Beeken RJ, McEwen A, Baldwin DR, et al. Lung Screen Uptake Trial (LSUT): Randomized Controlled Clinical Trial Testing Targeted Invitation Materials. *Am J Respir Crit Care Med*. 2020;201(8):965-75.
14. National Health Service England -N, Cancer, Programme,. Targeted Screening for Lung Cancer with Low Radiation Dose Computed Tomography. Standard Protocol prepared for the Targeted Lung Health Checks Programme 2019 [cited UK Government. Available from: <https://www.england.nhs.uk/wp-content/uploads/2019/02/targeted-lung-health-checks-standard-protocol-v1.pdf>.
15. D'Amelio AM, Jr., Cassidy A, Asomaning K, Raji OY, Duffy SW, Field JK, et al. Comparison of discriminatory power and accuracy of three lung cancer risk models. *British Journal of Cancer*. 2010;103(3):423-9.
16. Li K, Hüsing A, Sookthai D, Bergmann M, Boeing H, Becker N, et al. Selecting High-Risk Individuals for Lung Cancer Screening: A Prospective Evaluation of Existing Risk Models and Eligibility Criteria in the German EPIC Cohort. *Cancer prevention research (Philadelphia, Pa)*. 2015;8(9):777-85.

17. Lebrecht MB, Balata H, Evison M, Colligan D, Duerden R, Elton P, et al. Analysis of lung cancer risk model (PLCOM2012 and LLPv2) performance in a community-based lung cancer screening programme. *Thorax*. 2020;75(8):661-8.
18. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology*. 2015;44(3):827-36.
19. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer*. 2008;98(2):270-6.
20. Gorelick MH. Bias arising from missing data in predictive models. *J Clin Epidemiol*. 2006;59(10):1115-23.
21. The Children and Young Persons (Sale of Tobacco etc.) Order 2007., 767 (2007).
22. West R. Tobacco smoking: Health impact, prevalence, correlates and interventions. *Psychol Health*. 2017;32(8):1018-36.
23. Dierker L, He J, Kalaydjian A, Swendsen J, Degenhardt L, Glantz M, et al. The importance of timing of transitions for risk of regular smoking and nicotine dependence. *Ann Behav Med*. 2008;36(1):87-92.
24. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*. 2015;162(1):W1-W73.
25. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Stat Med*. 1991;10(4):585-98.
26. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*. 1996;49(12):1373-9.
27. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer; 2009.
28. Coleman T, Lewis S, Hubbard R, Smith C. Impact of contractual financial incentives on the ascertainment and management of smoking in primary care. *Addiction*. 2007;102(5):803-8.
29. Digital N. *Statistics on Smoking 2020: Data tables 2020* [Available from: <https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-smoking/statistics-on-smoking-england-2020/statistics-on-smoking-2020-data-tables>].
30. Langley TE, Szatkowski LC, Wythe S, Lewis SA. Can primary care data be used to monitor regional smoking prevalence? An analysis of The Health Improvement Network primary care data. *BMC Public Health*. 2011;11(1):773.
31. Szatkowski L, Lewis S, McNeill A, Huang Y, Coleman T. Can data from primary care medical records be used to monitor national smoking prevalence? *Journal of Epidemiology and Community Health*. 2012;66(9):791-5.
32. Booth HP, Prevost AT, Gulliford MC. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011. *Pharmacoepidemiol Drug Saf*. 2013;22(12):1357-61.
33. Shiffman S. How many cigarettes did you smoke? Assessing cigarette consumption by global report, Time-Line Follow-Back, and ecological momentary assessment. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*. 2009;28(5):519-26.
34. Field JK, Vulkan D, Davies MPA, Duffy SW, Gabe R. Liverpool Lung Project lung cancer risk stratification model: calibration and prospective validation. *Thorax* 2021;76:161-168.
35. Weber M, Yap S, Goldsbury D, Manners D, Tammemagi M, Marshall H, et al. Identifying high risk individuals for targeted lung cancer screening: Independent validation of the PLCOM2012 risk prediction tool. *International Journal of Cancer*. 2017;141(2):242-53.
36. Banks E, Redman S, Jorm L, Armstrong B, Bauman A, Beard J, et al. Cohort profile: the 45 and up study. *Int J Epidemiol*. 2008;37(5):941-7.

37. Han SS, Ten Haaf K, Hazelton WD, Munshi VN, Jeon J, Erdogan SA, et al. The impact of overdiagnosis on the selection of efficient lung cancer screening strategies. *International journal of cancer Journal international du cancer*. 2017;140(11):2436-43.

Table 1 (a): CPRD cohort and original LLP cohort demographic comparison

	CPRD		LLP _{v2}	
	Cohort (n=842,109)		Cohort (n=1736)	
	Percent or interquartile range in parentheses		Percent or interquartile range in parentheses	
	Non Lung cancer cases	Lung cancer cases	Controls	Cases
Number of patients	834986 (99.1)	7123 (0.9)	1157 (67)	579 (33)
Sex				
Females	373255 (45)	3060 (43)	444 (38)	222 (38)
Males	461731 (55)	4063 (57)	713 (62)	357 (62)
Age	62 (56 – 70)	69 (63 – 74)	66 (57 – 75)	66 (57 – 75)
Pneumonia				
No	811244 (97.2)	6789 (95.3)	989 (86)	361 (62)
Yes	23742 (2.8)	334 (4.7)	168 (14)	104 (18)
Personal History				
No	834770 (99.9)	7116 (99.9)	1091 (94)	509 (88)
Yes	216 (0.03)	7 (0.1)	66 (6)	72 (12)
Family History				
No	834151 (99.9)	7115 (99.9)	947 (82)	456 (79)
Yes	835 (0.1)	8 (0.1)	62 (5)	46 (8)
Late onset (>=60 years)†	-	-	148 (13)	77 (13)
Asbestosis*				
No	831734 (99.6)	7065 (99.2)	664 (76)	287 (65)
Yes	3252 (0.4)	58 (0.8)	206 (24)	155 (35)
Smoking Duration				
Never†			335 (29)	27 (5)
<=20 years	133770 (14)	266 (4)	236 (20)	43 (7)
>20- ≤40 years	392089 (47)	1985 (28)	337 (29)	1577
>40- ≤60 years	320646 (38)	4653 (65)	234 (20)	321 (55)
>60 years	8481 (1)	219 (3)	15 (1)	31 (5)

†Variables only in LLP_{v2} model – excluded from CPRD analysis. *Original LLP_{v2} model uses Exposure to Asbestos

Table 1 (b): CPRD cohort and PLCO_{m2012} cohort demographic comparison

	CPRD		NLST				PLCO _{m2012}			
	Cohort (n=842,109)		CT Arm (n=26,722)		Chest x-ray arm (n=26,730)		Chest x-ray arm (n=40,600)		Control arm (n=40,072)	
	Percent or interquartile range in parentheses									
	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases
Number of patients	834233 (99.1)	7876 (0.9)	25692 (96)	1030 (4)	25835 (97)	895 (3)	39846 (98)	754 (2)	39363 (98)	709 (2)
Age	62 (56 – 70)	68 (62 – 74)	60 (57 – 65)	63 (59 – 68)	60 (57 – 65)	64 (60 – 68)	62 (58 – 66)	65 (60 – 69)	62 (58 – 66)	65 (60–69)
BMI	27 (24 – 30.5)	26 (23 – 30)	27 (24 – 31)	26 (24 – 29)	27 (24.5 – 31)	26 (24 – 29)	27 (24 – 30)	26 (23 – 29)	27 (24 – 30)	26 (23 – 29)
Missing BMI n (%)	83260 (10)	721 (9)	146 (1)	13 (1)	206 (1)	7 (1)	494 (1)	10 (1)	742 (2)	15 (2)
Personal History										
No	834017 (99.9)	7869 (99.9)	24588 (96)	956 (93)	24554 (95)	833 (93)	38033 (95)	709 (94)	37532 (95)	653 (92)
Yes	216 (0.03)	7 (0.1)	1028 (4)	68 (7)	1154 (4)	58 (6)	1813 (5)	45 (6)	1831 (5)	56 (8)
Missing	0 (0)	0 (0)	76 (0)	6 (1)	127 (0.5)	4 (1)	0 (0)	0 (0)	0 (0)	0 (0)
Family History										
No	833399 (99.9)	7867 (99.9)	19741 (77)	746 (72)	19812 (77)	640 (72)	33718 (85)	565 (75)	33485 (85)	541 (76)
Yes	834 (0.1)	9 (0.1)	5554 (22)	261 (25)	5570 (22)	236 (26)	4514 (11)	139 (18)	4414 (11)	130 (18)
Missing	0 (0)	0 (0)	397 (2)	23 (2)	453 (2)	19 (2)	1614 (4)	50 (7)	1464 (4)	38 (5)
COPD										
No	775255 (93)	6122 (78)	21283 (83)	765 (74)	21435 (83)	643 (72)	36381 (91)	602 (80)	35899 (91)	567 (80)

Yes	58978 (7)	1754 (22)	4409 (17)	265 (26)	4400 (17)	252 (28)	3465 (9)	152 (20)	3464 (9)	142 (20)
Smoking Status										
Ex	473551 (57)	3633 (46)	1350 (53)	429 (42)	13561 (52)	337 (38)	32102 (81)	422 (56)	31708 (81)	385 (54)
Current	360682 (43)	4243 (54)	12183 (47)	601 (58)	12274 (48)	558 (62)	7744 (19)	332 (44)	7655 (19)	324 (46)
Smoking Intensity (cig/d)	15 (7 – 24)	17 (9 – 27)	25 (20 – 35)	30 (20 – 40)	25 (20 – 30.5)	25 (20 – 40)	20 (10 – 30)	30 (20 – 40)	20 (10 – 30)	30 (20 – 40)
Missing Smoking Intensity n (%)	231508 (28)	1212 (15)	0 (0)	0 (0)	0 (0)	0 (0)	78 (0.2)	4 (0.5)	112 (0.3)	2 (0.3)
Smoking Duration	37 (30 – 45)	45 (38 – 52)	40 (35 – 44)	44 (40 – 49)	40 (35 – 44)	44 (40 – 49)	28 (16 – 39)	42 (35 – 48)	28 (16 – 39)	42 (35 – 47)
Quit Years*	9 (6 – 24)	10 (5 – 15)	7 (3 – 11)	5 (2 – 10)	7 (3 – 11)	6 (2 – 11)	20 (10 – 30)	10 (4 – 19)	20 (10 – 30)	10 (4 – 18)

Table 2: AUC

	CPRD	NLST		PLCO	
	Cohort	CT arm	Chest x-ray arm	Chest x-ray arm	Control arm
LLP_{V2}*	0.700 (0.694 – 0.710)	0.66 (0.64 – 0.67)	0.68 (0.67 – 0.69)	0.75 (0.74 – 0.76)	0.75 (0.74 – 0.76)
PLCO_{m2012}†	0.679 (0.673 – 0.685)	0.69 (0.68 – 0.70)	0.71 (0.70 – 0.72)	0.80 (0.79 – 0.81)	0.79 (0.78 – 0.80)

* Five year incidence model performance

† Six year Model performance

Table 3: LLP_{v2} (recalibrated) performance in CPRD showing the proportion of people with lung cancer that might have been screened at selected thresholds

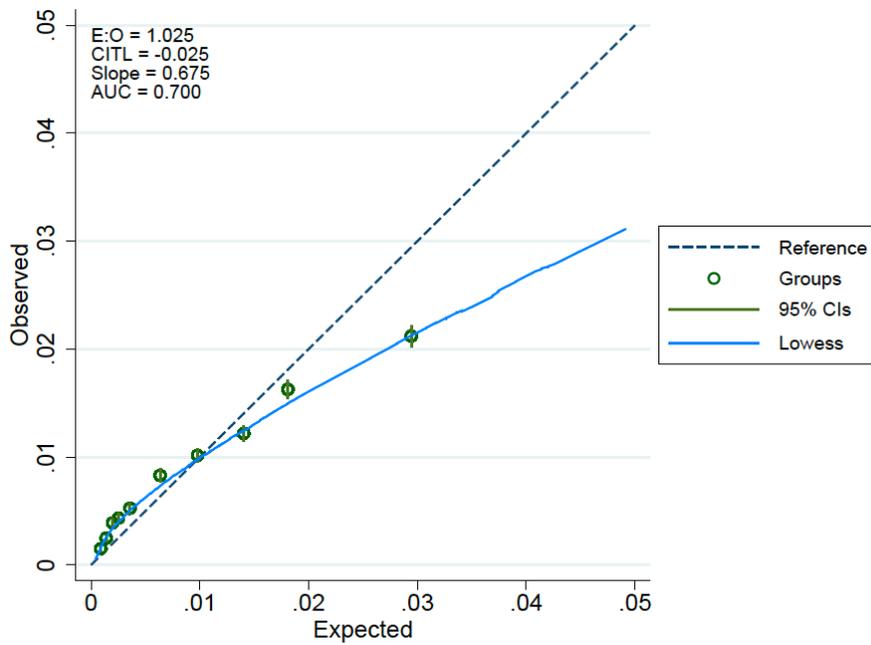
Corresponding risk threshold	LLP _{v2} (recalibrated) performance in CPRD							
	<1% (n=357,595)		1% - <2.5% (n=140,646)		2.5% - <5% (n=174,985)		≥5% (n=168,883)	
	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases
Number of patients	356485 (99.7)	1110 (0.3)	139672 (99.3)	974 (0.7)	173057 (98.9)	1928 (1.1)	165772 (98.2)	3111 (1.8)
Mean predicted risk	0.56%	0.64%	1.62%	1.69%	3.64%	3.71%	7.10%	7.50%
Proportion of total Lung Cancer	15.6%		13.7%		27.1%		43.7%	
Need to Screen to detect 1 lung cancer case	322		144		91		54	
Sex								
Females	178059 (50)	598 (54)	55285 (40)	454 (47)	78656 (45)	923 (48)	61255 (37)	1085 (35)
Males	178426 (50)	512 (46)	84387 (60)	520 (53)	94401 (55)	1005 (52)	104517 (63)	2026 (65)
Age	55 (52 – 58)	56 (53 – 58)	65 (61 – 71)	66 (61 – 70)	65 (62 – 68)	65 (63 – 68)	73 (71 – 77)	74 (71 – 77)
Pneumonia								
No	352184 (98.8)	1097 (98.8)	133567 (95.6)	935 (96)	169776 (98.1)	1887 (97.9)	155717 (93.9)	2870 (92.2)
Yes	4301 (1.2)	13 (1.2)	6105 (4.4)	39 (4)	3281 (1.9)	41 (2.1)	10055 (6.1)	241 (7.8)
Personal History								
No	356463 (99.9)	1110 (100)	139636 (99.9)	974 (100)	173026 (99.9)	1925 (99.8)	165645 (99.9)	3107 (99.9)
Yes	22 (0.01)	0 (0)	36 (0.03)	0 (0)	31 (0.02)	3 (0.2)	127 (0.1)	4 (0.1)
Family History								
No	356094 (99.9)	1109 (99.9)	139513 (99.9)	973 (99.9)	172917 (99.9)	1924 (99.8)	165627 (99.9)	3109 (99.9)
Yes	391 (0.1)	1 (0.1)	159 (0.1)	1 (0.1)	140 (0.1)	4 (0.2)	145 (0.1)	2 (0.1)
Asbestosis								

No	356270 (99.9)	1110 (100)	139036 (99.5)	973 (99.9)	172489 (99.7)	1916 (99.4)	163939 (98.9)	3066 (98.6)
Yes	215 (0.1)	0 (0)	636 (0.5)	1 (0.1)	568 (0.3)	12 (0.6)	1833 (1.1)	45 (1.4)
Smoking Duration								
<=20 years	90384 (25)	164 (15)	22885 (16)	94 (10)	494 (0.3)	7 (0.4)	7 (0)	1 (0)
>20- ≤40 years	266101 (75)	946 (85)	100929 (72)	749 (77)	23906 (14)	269 (14)	1153 (1)	21 1)
>40- ≤60 years	0 (0)	0 (0)	15858 (12)	131 (13)	148657 (86)	1652 (86)	156131 (94)	2870 (92)
>60 years	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	8481 (5)	219 (7)

Table 4: PLCO_{m2012} (recalibrated) performance in CPRD showing the proportion of people with lung cancer that might have been screened at selected thresholds

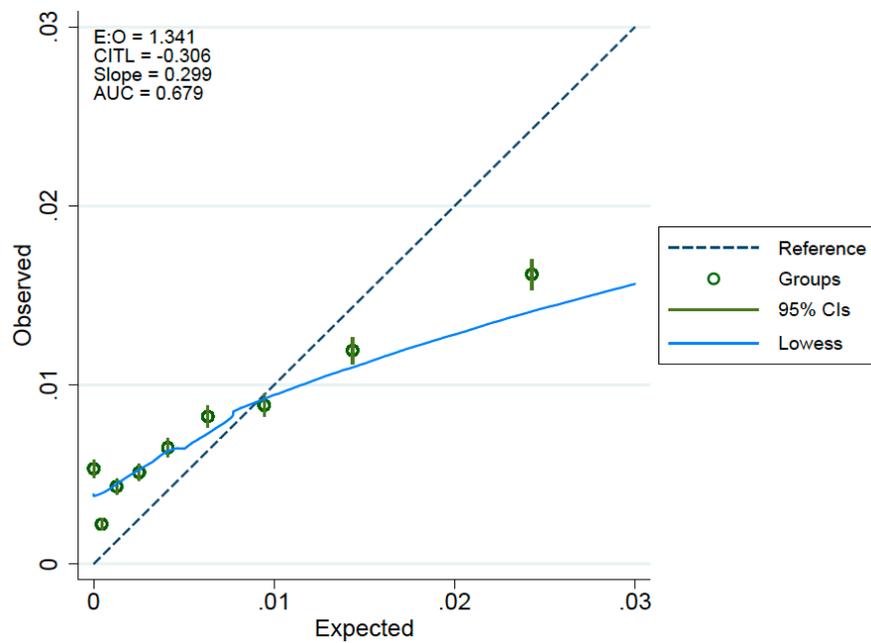
Corresponding risk threshold	PLCO _{m2012} performance in CPRD					
	<0.15% (n=229,443)		0.15% - 1.51% (n=414,032)		>1.51% (n=198,634)	
	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases	Non Lung cancer cases	Lung cancer cases
Number of patients	288544 (99.6)	899 (0.4)	410881 (99.2)	3151 (0.8)	194808 (98.1)	3826 (1.9)
Mean predicted risk	0.047%	0.045%	0.63%	0.74%	3.90%	4.76%
Proportion of total Lung Cancer	11.41%		40.01%		48.58%	
Need to Screen to detect 1 lung cancer case	255		131		52	
Age	60 (54 - 66)	66 (59 - 73)	61 (55 - 67)	65 (59 - 72)	69 (64 - 74)	71 (66 - 75)
BMI	27.3 (24.3 – 30.9)	27 (23.8 – 30.7)	27 (24 - 30.6)	26.3 (23.2 - 29.9)	26.2 (23.2 - 29.5)	25.6 (22.6 - 29)
Personal History						
No	228477 (99.9)	898 (99.9)	410815 (99.9)	3148 (99.9)	194725 (99.9)	3823 (99.9)
Yes	67 (0.03)	1 (0.1)	66 (0.02)	3 (0.1)	83 (0.04)	3 (0.1)
Family History						
No	228317 (99.9)	898 (99.9)	410513 (99.9)	3146 (99.8)	194569 (99.9)	3823 (99.9)
Yes	227 (0.1)	1 (0.1)	368 (0.1)	5 (0.2)	239 (0.1)	3 (0.1)
COPD						
No	221911 (97.1)	772 (85.9)	389474 (94.8)	2617 (83.1)	163870 (84.1)	2733 (71.4)
Yes	6633 (2.9)	127 (14.1)	21407 (5.2)	534 (16.9)	30938 (15.9)	1093 (28.6)
Smoking Status						
Ex	164008 (71.8)	573 (63.7)	231246 (56.3)	1592 (50.5)	78297 (40.2)	1468 (38.4)
Current	64536 (28.2)	326 (36.3)	179635 (43.7)	1559 (49.5)	116511 (59.8)	2358 (61.6)
Smoking Intensity (cig/d)	5 (2 - 8)	2 (2 - 6)	16 (11 - 25)	13 (8 - 19)	24 (17 - 33)	24 (16 - 33)
Smoking Duration (in years)	32 (18 - 39)	38 (31 - 45)	35 (29 - 42)	41 (34 - 48)	47 (42 - 52)	49 (44 - 54)
Quit Yearst	15 (6 - 29)	10 (9 - 23)	9 (6 - 24)	10 (6 - 18)	6 (6 - 11)	9 (5 - 11)

Figure 1: LLP_{v2}



LLP_{v2} external validation in CPRD cohort intercept only (<5% risks)

Figure 2: PLCO_{m2012}



PLCO_{m2012} external validation in CPRD cohort intercept (<3% risks)