Supplementary Table 1

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.

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

	CPRD		LLP_{V2}	
	Cohort (n=842,109) Percent or interquartile range in parentheses		Cohort (n=1736) 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)	157 7)
>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
CPRD=Clinical Practice Research Datalink; LLP=Liverpool Lung project

Table 1b: CPRD cohort, NLST and $PLCO_{m2012}$ cohort demographic comparison

	CPRD		NLST				PLCO ever-smokers			
	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-smokers	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)

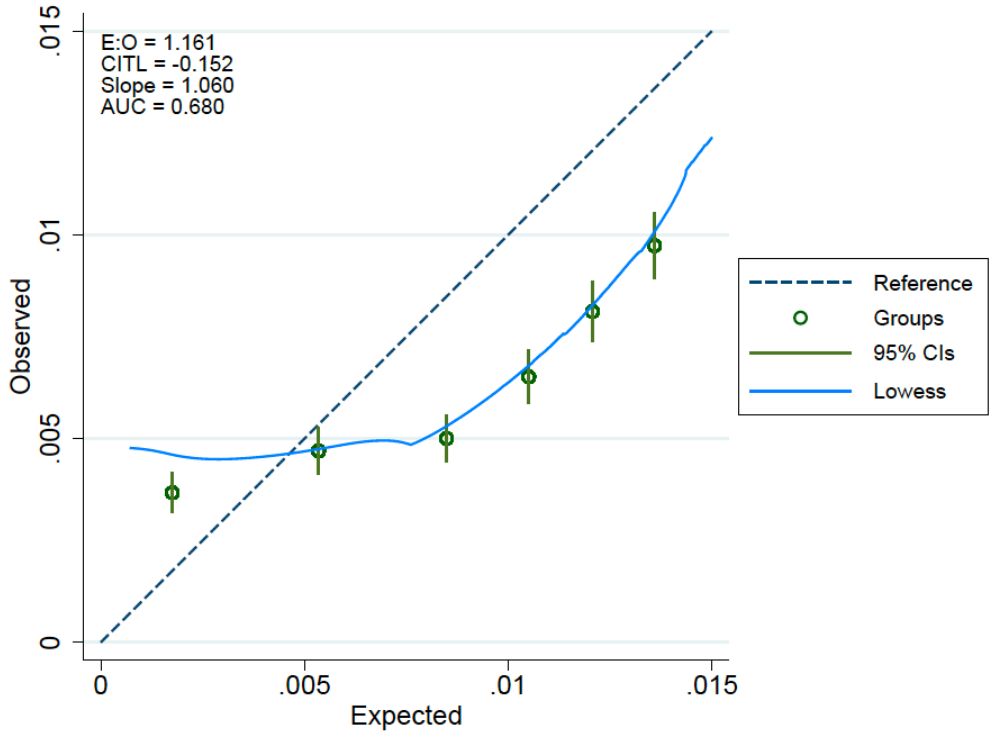
LLP = Liverpool Lung Project; CPRD = Clinical Practice Research Datalink; NLST = National Lung Screen Trial; PLCO= Prostate Lung Colorectal and Ovarian: BMI=Body mass index, COPD=Chronic obstructive pulmonary disease

Supplementary material: sensitivity analysis

Complete case analysis

In the complete case analysis, the c-statistic for $PLCOm_{2012}$ was 0.6800(0.67327-0.68678). There were 9.98 % missing data for BMI and 28% missing on smoking intensity. The total number in the cohort was 555,550. The complete case analysis could have simply deleted participants with missing values leaving a non-random subset of the original study sample, evaluating invalid predictive performance. Therefore, multiple imputation (MI) was used that substituted the missing observations by plausible estimates values derived from the analysis of the available data. These are the results presented in the main paper.

The calibration curve for the complete case analysis is shown below in Supplementary figure 1:



Supplementary Figure 1: Calibration curve for the complete case analysis.
E:O = expected to observed; CITL = Calibration-In-The-Large; AUC = Area Under the Curve; CI=Confidence interval

Discrimination with Family history excluded. The AUCs for original and recalibrated LLP_{v2} and the $PLCO_{m2012}$ did not alter when family history was omitted, as shown below.

Supplementary Table 2:

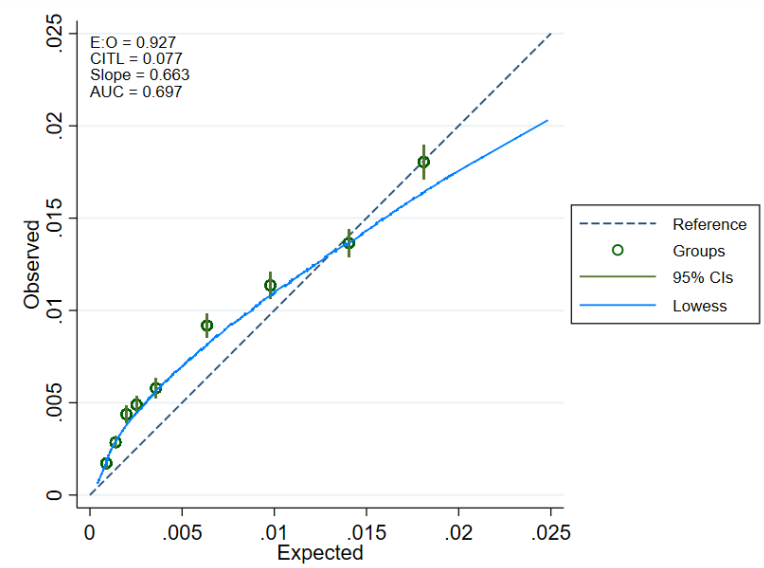
	Observations	**AUC**	**Std. Error**	**(95% CI)**
Original LLP_{v2}	842109	0.6967	0.0028	0.6913 to 0.7021
Recalibrated LLP_{v2}	842109	0.6967	0.0028	0.6913 to 0.7021
$PLCO_{m2012}$	842109	0.6785	0.0031	0.6725 to 0.6845

LLP= Liverpool Lung Project, PLCO= AUC=Area under the curve, CI=confidence interval

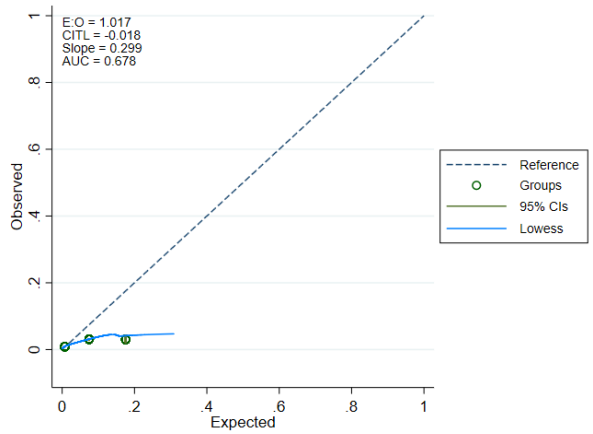
Supplementary table 2 shows AUCs for models with family history excluded

Supplementary figures 2a and 2b show the calibration plot for LLP_{v2} recalibrated and $PLCO_{m2012}$ respectively after family history was omitted.

Supplementary Figure 2a



Supplementary Figure 2b



E:O = expected to observed; CITL = Calibration-In-The-Large; AUC = Area Under the Curve; CI=Confidence interval