Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model

Supplementary Materials

Study Populations

Lung cancer OncoArray project of the International Lung Cancer Consortium (ILCCO). The ILCCO Lung cancer OncoArray project has been previously published[1]. In brief, it included 26 lung cancer studies that were genotyped by the Illumina OncoArray [2], which comprised a GWAS backbone and a custom cancer panel to facilitate in-depth interrogation of the cancer susceptibility genes[3]. For this study, we included only unrelated individuals with European descents, and a total of 18,316 lung cancer cases and 14,025 controls were used for PRS constriction. All lung cancer patients were histologically confirmed. The imputation was conducted based on 1000 Genome v.3 as described previously [1]. A total of 13,119 cases and 10,008 controls had epidemiological data required for the risk prediction modeling (such as demographics, smoking history, COPD and family history of lung cancer) and was used for the downstream analysis combining genetic and epidemiological data (**Supplementary Figure 1**). The protocol of the pooled analysis was approved by the Research Ethics Review Board at the Sinai Health System. The recruitment and data collection of all participating research institutes was approved by the local ethics review committee.

UK Biobank. UK Biobank is a population-based cohort study of over 500,000 participants, aged 40-69 at entry, recruited throughout the United Kingdom between 2006 to 2010 [4]. The details of the study design and data elements have been previously described [4]. In brief, epidemiological information such as lifestyle risk factors, medical history and family history of lung cancer were collected via study questionnaires. In addition, extensive physical measurement and biospecimens were collected at baseline. Lung cancer diagnosis was obtained through record linkage with death and cancer registries with the follow-up time up to date of death, lung cancer diagnosis, or March 31, 2016 (in England and Wales) and Oct 31, 2015 (in Scotland) per censor date defined by UKB. To minimize the possibility of including lung cancer metastasis, we excluded lung cancer that occurred

within 5 years of different primary cancer. In addition, prevalent lung cancer cases diagnosed prior to baseline enrollment were excluded. A total of 1,768 primary lung cancer cases and 334,163 unrelated controls with European ancestry were available for analysis (**Supplementary Figure 1**). Genotyping was completed using the UK BiLEVE Axiom array and the UK Biobank Axiom array [5]. Imputation was performed based on the Haplotype Reference Consortium (HRC) reference panel as the first choice and supplemented with those with a combination of UK10K and 1000 Genomes panels. This research was conducted with approved access to UK Biobank data under applications number 23261.

Statistical Analysis

Construction of Polygenic Risk Score

In general, the polygenic risk score (PRS) is constructed as the sum of the number of minor alleles weighted by their effect coefficients.

$$PRS = \sum_{k} \beta_k \, g_k$$

Where, $\beta_k$is the estimated per allele log-odds ratio for the association between lung cancer and the minor allele of the k^{th} variant and g_k is the number of genotyped minor alleles 0,1,2 of the k^{th} variant or genotype dosage.

There are two components included in PRS: one is comprised of the known lung cancer susceptibility loci previously identified, and one included additional loci that previously did not reach genome-wide significance, but were identified in this analysis through application of a machine learning algorithm. The list of known lung cancer loci were compiled based on literature and NHGRI-EBI GWAS Catalog [6], including variants that were associated with either overall or histology-specific lung cancer. We also included several variants that did not reach the stringent GWAS level of significance, but could potentially improve risk stratification: variants identified on the basis of their functional significance[7], uncovered through their association with first-degree family history of lung cancer [8], and those identified by a fine-mapping investigation of lung cancer

susceptibility loci 5p15.33 [9]. In addition, we included genetic variants identified for related disease traits, such as lung function impairment at the genome-wide significance level ($p<5x10^{-8}$)[1, 7-11]. Correlated variants with r^2 more than 0.2 based on the 1000 Genome v3 panel and the variants representing independent loci with the strongest statistical significance were retained. The final component of known lung cancer loci included 35 variants (PRS-35), as shown in the **Supplementary Table 1**, along with their log-odds ratio estimated based on the OncoArray meta-analysis [1], the largest lung cancer study to date, thus providing the most reliable effect estimates.

To maximize the prediction performance of the PRS, we went beyond the previously known loci and performed a penalized regression using *lasso* on a pre-selected set of SNPS that passed the suggestive significance-level ($p<5x10^{-6}$) in either overall or histology-specific lung cancer based on the combined analysis of OncoArray and previous ILCCO genome-wide studies[1]. All pre-selected SNPs had minor allele frequencies of at least 0.05 and were filtered for IMPUTE2 imputation quality score (INFO>0.3). The model selection was performed based on the lung cancer OncoArray data with 32,341 subjects of European ancestry with genetic data (18,316 lung cancer patients and 14,025 controls) as the training set. The most optimal penalty parameter (lambda) was selected based on a 10-fold cross-validation [12]. Each fold of the cross-validation analysis was adjusted by age, sex and top five principal components (PCs). Each variant selected was weighted by the lasso-shrunken parameter estimate in the PRS.

The best performing lasso model selected 221 variants, and among those, 93 variants remained after applying an r^2 threshold of 0.2. The final PRS (PRS-128) was constructed by combining PRS-35 which represents the known loci, and the additional 93 SNPs selected from the lasso analysis (**Supplementary Table 1**).

We compared effect sizes of PRS for lung cancer risk by groups defined by PRS deciles (<10%, 10–20%, 20–40%,

60–80%, 80–90%, >90%); by histologic-subtypes (adenocarcinoma, squamous cell, small cell); smoking status and family history of lung cancer in first degree relatives.

Assessment of multiplicative interactions between PRS and epidemiologic factors

We performed likelihood ratio tests to evaluate multiplicative interactions assumption between PRS and the epidemiologic risk factors age, family history of lung cancer and smoking variables in the OncoArray datasets. We did not observe consistent evidence of interactions between PRS and risk factors, except with age (interaction p=0.02) and smoking status (interaction p=0.01). The AUC however did not change when we incorporate the interaction terms into the model. We therefore report the parsimonious model, which reached the same predictive accuracy without interaction terms. None of the other risk factors showed consistent evidence of interactions with the polygenic risk scores.

PRS Validation and Model Evaluation based on the UK Biobank dataset

Standard quality control criteria were applied to the UK Biobank data to remove duplicates, relatedness, and sex discrepancies as previously described [5]. The PRS in the UK Biobank was computed based on the same weights derived and applied in the OncoArray dataset to avoid model overfitting. Fourteen (2 from PRS-35) variants were not genotyped or imputed based on Haplotype Reference Consortium (HRC) panel, and thus were not included in the PRS used in UK Biobank, which resulted in a total of 114 variants in the PRS for the analysis in UK Biobank. All of the variants in the PRS passed imputation quality threshold (INFO>0.3). To validate the PRS constructed in OncoArray, we used the same effect coefficients for the parameters included in the model (**Supplementary Table 2**).

To eliminate the potential over- or under-estimation when importing coefficients of a risk model previously built in a different population and to integrate PRS into the model, we recalibrated the $PLCO_{all2014}$ model based on random sample of 50% of UKB data, while holding the remaining 50% of data for strict prospective validation.

We computed the log-odds of lung cancer (Z) in UKB based on the original $PLCO_{all2014}$ coefficients with the addition of two PRS coefficients. Then we fit a logistic regression model in the 50% training sample with lung cancer status as the outcome and Z as the sole predictor. The beta coefficient for Z , $\hat{\beta}_Z$, is the re-calibrated slope (i.e. the adjustment factor). For absolute risk trajectories, $\hat{\beta}_Z$ was applied to regularize the $PLCO_{all2014}$ coefficients.

In addition, to acknowledge the markedly different baseline risk and potential risk factors for never smoker population, we built a *de novo* model for never smokers based potential predictors defined *a priori*, including age, sex, education, BMI, personal history of cancer, family history of lung cancer in first degree relatives, lung function (FEV_1/FVC), ambient air pollution and second hand smoke. We adapted the split design and used 80% of the UKB data for training and 20% was set aside for hold-out testing set. Within the 80% training data, we applied 10-fold cross-validation to select the parsimonious model. The model with ambient air pollution and second-hand smoke did not improve the AUC (0.670, 95%CI=0.611-0.728), therefore the final parsimonious model includes age, sex, education, BMI, personal history of cancer, family history of lung cancer and lung function.

Evaluation of all model performance , including model calibration and discrimination were evaluated based on the hold-out set only. Model calibration was assessed by evaluating how much the slope of the calibration line (plotting the predicted vs the observed probabilities) deviates from the ideal of 1. The 95% confidence intervals of the predicted risk were computed with the percentile-based bootstrap method using 100 replicates. Calibration was formally tested using Spiegelhalter's *z* statistic and the corresponding p-values [13, 14]. The risk model's ability to discriminate was assessed by the area under the receiver operator characteristic curves (AUC). Risk discrimination improvement of the developed PRSs was evaluated by comparing a base model with epidemiologic risk factors and a model that includes epidemiologic risk factors and PRS.

Absolute risk estimation

The absolute risk of developing lung cancer was estimated based on Cox proportional hazards model accounting for the presence of competing risk of all causes of death other than lung cancer, as originally described by Benichou and Gail[15]. The risk in a given time interval $(a, a+\tau)$ is estimated by integrating a model of relative risks, age-specific lung cancer incidence rates and a representative distribution of risk factors of the population of interest, where, *X* represents the risk factors, $h_0(t)$ is the baseline hazard function, $m(t)$ is age-specific competing hazards of mortality, u as the time interval for the estimation of the integral, and β is a vector of log-odds ratios. The underlying assumption of the integrated risk prediction model is that risk factors act in a multiplicative fashion on the baseline hazard function.

$$AR(a, a+\tau) = \int_a^{a+\tau} h_0(t)\exp(X\beta)\exp\left(-\int_a^t [h_0(u)\exp(X\beta) + m(u)]\,du\right)dt$$

To estimate the absolute risk in the UK Biobank, the frequency distribution of epidemiologic risk factors was estimated based on the full UKB cohort. Age-specific lung cancer rates and competing rates for mortality rates obtained from Cancer Research UK, 2012[16]. The age-specific lung cancer rates specifically for never smokers were derived from the UK Million Women Cohort[17], and the average male to female incidence ratio of lung cancer in never smokers previously reported in population cohorts[18]. The underlying assumption of the integrated risk prediction model is that risk factors act in a multiplicative fashion on the baseline hazard function.

NLST PRS Simulation and Projection

PRS distributions in NLST were simulated conditional on lung cancer status and family-history of lung cancer based on the effect estimation and allele frequency from the validation set (UK Biobank) using iCARE package as previously described [19, 20]. Age-specific overall lung cancer incidence rates were obtained from the Center for Disease Control, 2013[21]. The majority (>90%) of all NLST participants are of European ancestry, thus we used

incidence rates of non-Hispanic white population in the US for the absolute risk projection. Since NLST represents a selected high risk population, not the general population, the overall incidence rates of lung cancers in the US population were multiplied by an adjustment ratio of 4.3, derived by the ratio of the percentage of all lung cancer that are eligible for NLST (26.7%) and the US population that meet the NLST-eligibility criteria (6.2%)[22]. We simulated five independent PRS distributions for the NLST cohort. The weights of the PRS were based on the coefficient estimated from the validation set (UK Biobank).

Reference list

1. McKay JD, Hung RJ, Han Y*, et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet 2017;49(7):1126-1132.

2. Amos CI, Dennis J, Wang Z*, et al.* The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2016.

3. Amos CI, Dennis J, Wang Z*, et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev 2017;26(1):126-135.

4. Sudlow C, Gallacher J, Allen N*, et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Medicine 2015;12(3):1-10.

5. Bycroft C, Freeman C, Petkova D*, et al.* The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562(7726):203-209.

6. MacArthur J, Bowler E, Cerezo M*, et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 2017;45(D1):D896-D901.

7. Brenner DR, Amos CI, Brhane Y*, et al.* Identification of lung cancer histology-specific variants applying Bayesian framework variant prioritization approaches within the TRICL and ILCCO consortia. Carcinogenesis 2015;36(11):1314-26.

8. Poirier JG, Brennan P, McKay JD*, et al.* Informed genome-wide association analysis with family history as a secondary phenotype identifies novel loci of lung cancer. Genetic Epidemiology 2015;39(3):197-206.

9. Kachuri L, Amos CI, McKay JD*, et al.* Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. Carcinogenesis 2016;37(1):96-105.

10. Bosse Y, Amos CI. A Decade of GWAS Results in Lung Cancer. Cancer Epidemiol Biomarkers Prev 2018;27(4):363-379.

11. Weissfeld JL, Lin Y, Lin HM*, et al.* Lung Cancer Risk Prediction Using Common SNPs Located in GWAS-Identified Susceptibility Regions. J Thorac Oncol 2015;10(11):1538-45.

12. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33(1):1-22.

13. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020;27(4):621-633.

14. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5(5):421-33.

15. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. Biometrics 1990;46(3):813-26.

16. CRUK. Lung cancer, age-specific incidence rates, 2012-2014. In: Cancer Research UK; 2017.

17. Pirie K, Peto R, Green J*, et al.* Lung cancer in never smokers in the UK Million Women Study. Int J Cancer 2016;139(2):347-54.

18. Wakelee HA, Chang ET, Gomez SL*, et al.* Lung cancer incidence in never smokers. J Clin Oncol 2007;25(5):472-8.

19. Maas P, Wheeler W, N. BM*, et al.* iCARE (individualized Coherent Absolute Risk Estimators). In: the National Cancer Institute; 2016.

20. Choudhury PP, P. M, Wilcox A*, et al.* iCARE: R package to build, validate and apply absolute risk models. BioRxiv 2018.

21. CDC. United States Cancer Statistics: 1999 - 2013 Incidence Archive. In: United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; 2016.

22. Pinsky PF, Berg CD. Applying the National Lung Screening Trial eligibility criteria to the US population: what percent of the population and of incident lung cancers would be covered? J Med Screen 2012;19(3):154-6.

Supplementary Table 1: Annotation of the genetic variants included in the polygenic risk scores

									OncoArray & ILCCO meta analysis		OncoArray lasso analysis
Polygenic risk score	SNP	Chr	Position	Locus	Gene	reference	effect	EAF	Odds ratio (95% CI)	Pvalue	Odds ratio
PRS-35	rs71658797	1	77967507	p31.1	*AK5*	T	A	0.1	1.14(1.09,1.18)	3.25E-11	NA
	rs13080835	3	189357199	q28	*TP63*	G	T	0.49	0.94(0.92,0.97)	1.25E-06	NA
	rs7705526	5	1285974	p15.33	*TERT*	C	A	0.34	1.12(1.10,1.15)	1.01E-18	NA
	rs112290073	5	1286032	p15.33	*TERT*	G	A	0.01	**1.39(1.20,1.61)**[b]	1.16E-05	NA
	rs2736098	5	1294086	p15.33	*TERT*	C	T	0.28	**1.14(1.10,1.18)**[b]	4.36E-13	NA
	rs2853668	5	1300025	p15.33	*TERT*	G	T	0.24	1.08(1.05,1.11)	9.36E-09	NA
	rs401681	5	1322087	p15.33	*CLPTM1L*	C	T	0.43	0.87(0.85,0.89)	3.25E-30	NA
	rs466502	5	1325767	p15.33	*CLPTM1L*	A	G	0.44	0.91(0.89,0.93)	1.99E-15	NA
	rs6903823	6	28354519	p22.1	*ZKSCAN3*	A	G	0.2	1.07(1.04,1.10)	1.09E-05	NA
	rs116822326	6	31434111	p21.33		A	G	0.16	1.15(1.12,1.19)	5.29E-19	NA
	rs2855812	6	31504943	p21.33	*MICB*	G	T	0.22	1.05(1.03,1.08)	1.26E-04	NA
	rs805262	6	31628733	p21.33	*C6orf47*	C	T	0.47	1.07(1.04,1.09)	1.50E-08	NA
	rs6916278	6	31678774	p21.33	*LY6G6F-LY6G6D*	G	A	0.05	0.90(0.86,0.95)	1.15E-04	NA
	rs3129763	6	32590925	p21.32		G	A	0.22	1.12(1.09,1.15)	8.48E-16	NA
	rs114544105[a]	6	32667852	p21.32	*HLA-DQB1*	G	A	0.2	1.06(1.02,1.09)	9.14E-04	NA
	rs6920364[a]	6	167376466	q27		G	C	0.46	1.07(1.05,1.10)	1.29E-08	NA
	rs11780471	8	27344719	p21.2	*EPHX2*	G	A	0.06	0.87(0.83,0.91)	1.69E-08	NA
	rs4236709	8	32410110	p12	*NRG1*	A	G	0.22	1.07(1.04,1.10)	5.88E-06	NA
	rs885518	9	21830157	p21.3	*MTAP*	A	G	0.1	1.09(1.05,1.13)	2.13E-06	NA
	rs2007153	9	136503819	q34.2	*DBH*	C	T	0.37	0.96(0.93,0.98)	2.49E-04	NA
	rs11591710	10	105687632	q24.33		A	C	0.14	1.07(1.04,1.11)	3.53E-05	NA
	rs1056562	11	118125625	q23.3	*MPZL2*	C	T	0.48	1.07(1.04,1.09)	1.92E-08	NA
	rs7953330	12	998819	p13.33	*WNK1*	G	C	0.31	0.92(0.89,0.94)	6.10E-12	NA
	rs11571833	13	32972626	q13.1	*BRCA2*	A	T	0.01	1.60(1.43,1.80)	6.12E-16	NA
	rs689647	15	43762196	q15.3	*TP53BP1*	C	T	0.11	0.93(0.90,0.97)	2.11E-04	NA
	rs66759488	15	47577451	q21.1	*SEMA6D*	G	A	0.36	1.07(1.04,1.10)	2.83E-08	NA
	rs77468143	15	49376624	q21.1		T	G	0.25	0.92(0.90,0.95)	1.00E-09	NA
	rs3885951	15	78825917	q25.1	*HYKK*	A	G	0.11	**1.17(1.12,1.23)**[b]	4.09E-10	NA
	rs55781567	15	78857986	q25.1	*CHRNA5*	C	G	0.37	1.30(1.27,1.33)	3.08E-103	NA
	rs7177699	15	79089734	q25.1	*ADAMTS7*	T	C	0.44	1.13(1.11,1.16)	5.98E-26	NA
	rs62070270	17	29936962	q11.2	*EFCAB5*	A	G	0.45	1.03(1.01,1.06)	7.24E-03	NA
	rs1542752	17	72938100	q25.1	*OTOP3*	C	T	0.16	1.04(1.01,1.07)	1.24E-02	NA
	rs56113850	19	41353107	q13.2	*CYP2A6*	C	T	0.44	0.88(0.86,0.91)	5.02E-19	NA
	rs41309931	20	62326579	q13.33	*RTEL1*	G	T	0.12	1.08(1.04,1.12)	2.23E-05	NA
	rs17879961	22	29121087	q12.1	*CHEK2*	A	G	0.01	0.60(0.52,0.70)	1.54E-10	NA
	rs71641333	1	78743005	p31.1	*MGC27382*	T	A	0.06	1.14(1.09,1.21)	4.49E-07	1.03
	rs78062588	1	154566225	q21.3	*ADAR*	T	C	0.06	0.88(0.84,0.93)	4.60E-07	0.95
	rs114737056	1	168511081	q24.2	*XCL2*	G	A	0.12	0.91(0.88,0.94)	5.80E-07	0.95
	rs145733018	2	38567201	p22.2	*ATL2*	T	C	0.02	2.20(1.61,3.00)	7.24E-07	1.04
	rs79368540	2	45189737	p21		C	T	0.15	1.09(1.06,1.13)	6.13E-07	1.05
	rs11692700[a]	2	67510377	p14	*LINC01828*	T	C	0.03	1.20(1.12,1.29)	6.44E-07	1.05
	rs114928225	2	119449740	q14.2		T	A	0.01	1.65(1.35,2.01)	7.24E-07	1.17
	rs7592999	2	140398327	q22.1		T	C	0.04	0.81(0.74,0.88)	7.35E-07	0.92
	rs722864	2	173983204	q31.1	*MAP3K20*	G	A	0.19	0.93(0.90,0.96)	5.53E-07	0.99
	rs1866631[a]	2	174075761	q31.1	*MAP3K20*	A	G	0.4	0.94(0.92,0.96)	6.97E-07	0.99
	rs185666783	4	67833774	q13.2	*LOC105377262*	C	G	0.29	0.92(0.90,0.95)	9.20E-08	1.06
	rs7676823	4	164007992	q32.2		A	G	0.34	0.92(0.89,0.95)	6.71E-07	0.98
	rs78154696	5	1000156	p15.33		G	A	0.03	1.22(1.13,1.32)	7.55E-07	1.09
	rs112333466	5	1249816	p15.33	*TERT*	C	T	0.01	1.55(1.32,1.81)	8.11E-08	1.07
	rs56345976	5	1276873	p15.33	*TERT*	A	G	0.42	1.10(1.07,1.13)	2.60E-12	1.02
	rs2853677	5	1287194	p15.33	*TERT*	A	G	0.42	1.12(1.09,1.15)	2.66E-18	1.06
	rs112401627	5	1300269	p15.33		G	A	0.03	1.30(1.20,1.42)	3.03E-10	1.05

									OncoArray & ILCCO meta analysis		OncoArray lasso analysis
Polygenic risk score	SNP	Chr	Position	Locus	Gene	reference	effect	EAF	Odds ratio (95% CI)	Pvalue	Odds ratio
PRS-93	rs6875416	5	90250631	q14.3	*ADGRV1*	A	T	0.21	0.83(0.78,0.90)	6.99E-07	0.96
	rs114136906	5	150121458	q33.1	*DCTN4*	G	C	0.02	1.46(1.27,1.68)	1.17E-07	1.14
	rs2316515	6	410848	p25.3	*IRF4*	G	A	0.41	0.92(0.89,0.95)	1.42E-07	0.96
	rs629444	6	25885814	p22.2	*HIST1H2APS2*	C	T	0.1	1.11(1.07,1.15)	1.40E-08	1.00
	rs2179517	6	26198845	p22.2	*HIST1H3D*	G	C	0.49	0.94(0.92,0.96)	3.99E-08	1.00
	rs68141011	6	28217797	p22.1	*ZKSCAN4*	G	T	0.13	1.09(1.06,1.13)	1.43E-07	0.99
	rs114722608	6	29223493	p22.1	*LOC101929006*	G	C	0.09	1.15(1.11,1.20)	7.09E-12	0.79
	rs115123779	6	29477821	p22.1	*LOC105375009*	G	T	0.18	1.10(1.07,1.13)	1.91E-09	1.01
	rs138488080	6	29606761	p22.1	*SUMO2P1*	G	A	0.15	1.15(1.11,1.19)	5.96E-18	1.00
	rs114192654	6	29759750	p22.1		G	A	0.34	1.06(1.04,1.09)	5.69E-07	0.98
	rs116675020	6	29922740	p22.1	*HLA-W*	A	G	0.38	0.94(0.91,0.96)	6.72E-07	0.99
	rs115993819	6	30074163	p22.1	*TRIM31*	G	A	0.24	1.09(1.06,1.12)	9.64E-10	1.00
	rs116534499	6	30138162	p22.1	*TRIM15*	C	G	0.43	1.07(1.04,1.09)	2.94E-08	1.01
	rs116629156	6	30864829	p21.33	*DDR1*	T	C	0.41	1.07(1.05,1.10)	1.52E-08	1.00
	rs114103504	6	31002452	p21.33	*MUC22*	A	G	0.49	0.93(0.91,0.95)	3.29E-09	0.98
	rs114052224	6	31067852	p21.33		A	G	0.48	1.06(1.04,1.09)	1.26E-07	1.02
	rs2233959	6	31081065	p21.33	*C6orf15*	T	C	0.41	1.07(1.05,1.10)	5.14E-09	1.01
	rs114689412	6	31117577	p21.33	*CCHCR1*	C	G	0.13	0.91(0.88,0.94)	3.64E-07	1.00
	rs2596499	6	31321429	p21.33	*HLA-B*	T	A	0.3	1.07(1.04,1.10)	3.42E-07	0.99
	rs2596496	6	31322782	p21.33	*HLA-B*	G	C	0.36	0.90(0.87,0.94)	3.64E-07	0.99
	rs2596490 [a]	6	31324996	p21.33		C	G	0.23	0.87(0.83,0.92)	2.82E-08	1.01
	rs115176861	6	31412961	p21.33	*HCP5*	T	C	0.48	1.06(1.04,1.09)	1.11E-07	1.03
	rs553108	6	31840455	p21.33	*SLC44A4*	G	A	0.38	1.07(1.04,1.09)	6.63E-08	1.00
	rs115200960 [a]	6	32335204	p21.32	*C6orf10*	G	A	0.17	1.11(1.07,1.14)	4.77E-11	1.00
	rs12722051	6	32609147	p21.32	*HLA-DQA1*	A	T	0.18	0.87(0.83,0.92)	4.96E-07	0.98
	rs116767258	6	32757737	p21.32		A	G	0.39	0.94(0.92,0.96)	9.10E-07	0.96
	rs7383287	6	32783086	p21.32	*HLA-DOB*	A	G	0.2	1.10(1.07,1.13)	1.12E-10	1.03
	rs117534741	6	72384541	q13		G	A	0.02	1.24(1.14,1.34)	4.68E-07	1.17
	rs1321817	6	117734267	q22.1	*ROS1*	A	G	0.37	0.92(0.89,0.95)	5.67E-07	0.99
	rs6957511	7	130668618	q32.3	*LINC-PINT*	T	C	0.4	1.10(1.06,1.14)	9.78E-07	1.02
	rs2565064	8	27327841	p21.2	*CHRNA2*	G	C	0.29	1.07(1.04,1.10)	4.58E-07	1.03
	rs67749759	8	27397087	p21.2	*EPHX2*	C	T	0.07	1.13(1.08,1.19)	2.66E-07	1.03
	rs111960002	8	144722420	q24.3	*ZNF623*	T	C	0.05	1.36(1.21,1.54)	4.62E-07	1.04
	rs10118776	9	6227418	p24.1	*IL33*	A	G	0.06	1.34(1.20,1.50)	3.91E-07	1.08
	rs17185553 [a]	9	17934120	p22.2		G	C	0.08	1.29(1.17,1.43)	8.94E-07	1.02
	rs2518717	9	21959751	p21.3	*RP11-145E5.5*	T	C	0.36	1.09(1.06,1.13)	3.35E-07	1.00
	rs28557075	9	22066572	p21.3	*CDKN2B-AS1*	G	A	0.09	1.11(1.06,1.16)	8.43E-07	1.11
	rs1333040	9	22083404	p21.3	*CDKN2B-AS1*	T	C	0.46	1.10(1.06,1.14)	7.02E-07	1.00
	rs4879704	9	33427322	p13.3		A	C	0.33	0.92(0.89,0.95)	9.16E-07	0.98
	rs191205566	9	102587233	q22.33	*NR4A3*	C	T	0.02	1.40(1.24,1.59)	1.17E-07	1.21
	rs75685923	9	136275229	q34.2	*REXO4*	C	T	0.03	1.38(1.22,1.57)	6.16E-07	1.09
	rs7897454	10	102011702	q24.31	*CWF19L1*	G	A	0.04	1.25(1.14,1.36)	6.64E-07	1.06
	rs62621207	10	102672248	q24.31	*SLF2*	A	T	0.05	1.16(1.09,1.23)	5.85E-07	1.07
	rs78853063	11	57250026	q12.1	*SLC43A1*	C	T	0.08	0.89(0.85,0.93)	4.65E-07	0.97
	rs78334599	11	115998756	q23.3		G	A	0.04	0.86(0.80,0.91)	7.93E-07	0.89
	rs7487683	12	1036042	p13.33	*RAD52*	C	T	0.04	0.83(0.77,0.89)	6.58E-08	0.94
	rs73351723	12	58831070	q14.1		G	A	0.13	1.09(1.06,1.13)	3.81E-07	1.00
	rs9668978	12	64913237	q14.2	*RP11-439H13.2*	G	T	0.29	1.10(1.06,1.14)	6.24E-07	1.05
	rs9602270	13	84281063	q31.1		A	T	0.05	1.27(1.16,1.39)	3.28E-07	1.04
	rs8003466	14	34013721	q13.1	*NPAS3*	G	A	0.18	0.89(0.84,0.93)	5.27E-07	0.96
	rs8031813	15	49253961	q21.1	*SHC4*	A	C	0.31	0.91(0.87,0.94)	4.29E-08	0.99
	rs6493361	15	49615952	q21.2	*GALK2*	C	G	0.34	1.09(1.05,1.12)	7.09E-07	1.02
	rs11855650	15	70431773	q23		G	T	0.38	1.09(1.05,1.12)	5.60E-07	1.06
	rs79149102	15	75055819	q24.1		C	T	0.03	1.18(1.11,1.25)	1.54E-07	1.09

									OncoArray & ILCCO meta analysis		OncoArray lasso analysis
Polygenic risk score	SNP	Chr	Position	Locus	Gene	reference	effect	EAF	Odds ratio (95% CI)	Pvalue	Odds ratio
	rs2229961	15	78880752	q25.1	*CHRNA5*	G	A	0.02	1.43(1.30,1.57)	5.01E-14	1.05
	rs8192479	15	78909398	q25.1	*CHRNA3*	C	T	0.02	1.28(1.17,1.40)	2.29E-08	1.08
	rs2869551	15	78981423	q25.1	*CHRNB4*	A	G	0.01	0.75(0.67,0.84)	5.09E-07	0.91
	rs12593207[a]	15	78987225	q25.1	*CHRNB4*	G	A	0.1	0.85(0.82,0.89)	2.88E-14	1.01
	rs189146505[a]	15	79058730	q25.1	*ADAMTS7*	A	G	0.04	0.83(0.78,0.89)	8.34E-08	1.00
	rs28450923[a]	15	79065557	q25.1	*ADAMTS7*	A	G	0.13	0.88(0.84,0.92)	7.76E-08	0.99
	rs28624856	15	79075233	q25.1	*ADAMTS7*	T	C	0.24	0.91(0.88,0.94)	1.06E-09	1.00
	rs77719127	15	79110783	q25.1	*MORF4L1*	C	T	0.17	1.14(1.10,1.18)	1.79E-13	1.00
	rs76164573	15	79198760	q25.1		T	G	0.04	0.86(0.81,0.91)	8.83E-07	0.97
	rs78442819	16	10740982	p13.13	*TEKT5*	G	C	0.19	0.89(0.85,0.93)	6.48E-07	0.97
	rs9926896	16	26980646	p12.1		T	C	0.01	2.70(1.90,3.83)	3.06E-08	1.22
	rs17181550	17	70299958	q24.3		T	G	0.43	0.94(0.92,0.96)	1.98E-07	0.97
	rs79421398	18	20741135	q11.2	*CABLES1*	T	C	0.05	1.23(1.14,1.34)	2.75E-07	1.04
	rs66500423	19	41195170	q13.2	*NUMBL*	T	C	0.3	1.07(1.04,1.09)	2.23E-07	1.01
	rs4803356	19	41207206	q13.2	*ADCK4*	C	G	0.07	0.89(0.85,0.93)	8.31E-07	1.00
	rs11881918[a]	19	41334199	q13.2	*CTC-490E21.12*	G	A	0.09	0.86(0.82,0.91)	4.10E-09	0.93
	rs2258380[a]	19	41338988	q13.2	*CTC-490E21.12*	C	G	0.23	1.08(1.05,1.11)	9.15E-07	0.99
	rs67210567	19	41357457	q13.2	*CYP2A6*	G	T	0.03	0.79(0.73,0.86)	2.96E-08	0.86
	rs184589612	19	41412192	q13.2	*CTC-490E21.13*	T	C	0.02	0.77(0.70,0.85)	3.23E-07	0.90
	rs12981718[a]	19	54567858	q13.42	*VSTM1*	G	A	0.07	1.36(1.21,1.53)	2.13E-07	0.99
	rs13036436	20	61988382	q13.33	*CHRNA4*	A	G	0.2	1.08(1.05,1.12)	9.03E-07	1.09
	rs61541144	20	62527305	q13.33	*DNAJC5*	G	A	0.07	0.89(0.85,0.93)	5.61E-07	0.91

[a] (bolded) SNPs not genotyped or imputed in the UK biobank based on Haplotype Reference Consortium (HRC) panel and therefore not included in PRS analysis in the UK Biobank

[b] Odds ratios for association with lung cancer in OncoArray analysis for SNPs with no OncoArray-ILCCO meta-analysis results

Supplementary Table 2: Risk factors included for absolute risk projection in the UK Biobank and NLST studies

	Recalibrated $PLCO^{c}_{all2014}$ (For overall population)	**Never-Smoker Model**	**$PLCO^{c}_{m2012}$** (For NLST)*
Covariate	**Beta**	**Beta**	**Beta**
Age	0.06962700	0.08374322	0.0778868
Sex	--	0.27877676	--
Education	-0.07691528	0.02989868	-0.0812744
Body Mass Index (kg/m^2)	-0.02532209	-0.01309118	-0.0274194
Chronic obstructive pulmonary disease (0=No; 1=Yes)	0.30376798	--	0.3553063
Personal history of cancer (0=No; 1=Yes)	0.42383167	0.0.48212857	0.4589971
Family history of lung cancer (0=No; 1=Yes)	0.51231807	-0.08546407	0.587185
FEV1/FVC	--	-1.64183399	--
Race/ethnicity			
White	Reference		
Black	0.28093323		0.3944778
Hispanic	-0.71758156		-0.7434744
Asian	-0.45847836		-0.466585
Native Hawaiian/Pacific Islander	-1.19348237		0
American Indian/Alaskan Native	0.83336775		1.027152
Smoking Status			
0 = Former smoker	2.22401218		
1 = Current smoker	2.44904445		
Smoking intensity (average cigarettes/day)	-0.15880855		-1.822606
Duration smoked (per year)	0.02672920		0.0317321
Smoking quit-time	-0.02811095		-0.0308572
Model Development Study	OncoArray	OncoArray	UK Biobank
Polygenic risk score (PRS)			
PRS_128 [a]	0.663		
PRS_114 [b]		0.45901616	0.462234

[a] Beta coefficient estimated in OncoArray adjusted for age, sex and top 5 PCs

[b] Beta coefficient estimated in UK Biobank, adjusted for age, sex and top 5 PCs

[c] Education, BMI, Smoking duration and Smoking quit-time were centered to the mean and Smoking intensity as modelled as a non-linear transformation as previously described.

*$PLCO_{m2012}$ was applied to NLST because it is an ever-smoking only population.

Supplementary Table 3: Area under the receiver operating characteristic curve (AUC) and 95% confidence intervals in UK Biobank for overall population, by smoking status, and for young -onset lung cancer

	Model			
	Risk factors only		Risk factors + PRS terms	
UKB Data	**AUC**	**95%CI**	**AUC**	**95%CI**
Overall [a]	0.828	0.807-0.850	0.832	0.811-0.853
Ever smokers [a]	0.785	0.762-0.809	0.786	0.762-0.809
Non-smokers [b]	0.670	0.611-0.729	0.687	0.628-0.746
Young onset (<50 years old) [a]	0.798	0.680-0.917	0.811	0.701-0.922

PRS, polygenic risk score

[a] AUCs were based on the 50% hold-out validation set that was not used for re-calibration of the $PLCO_{all2014}$ model

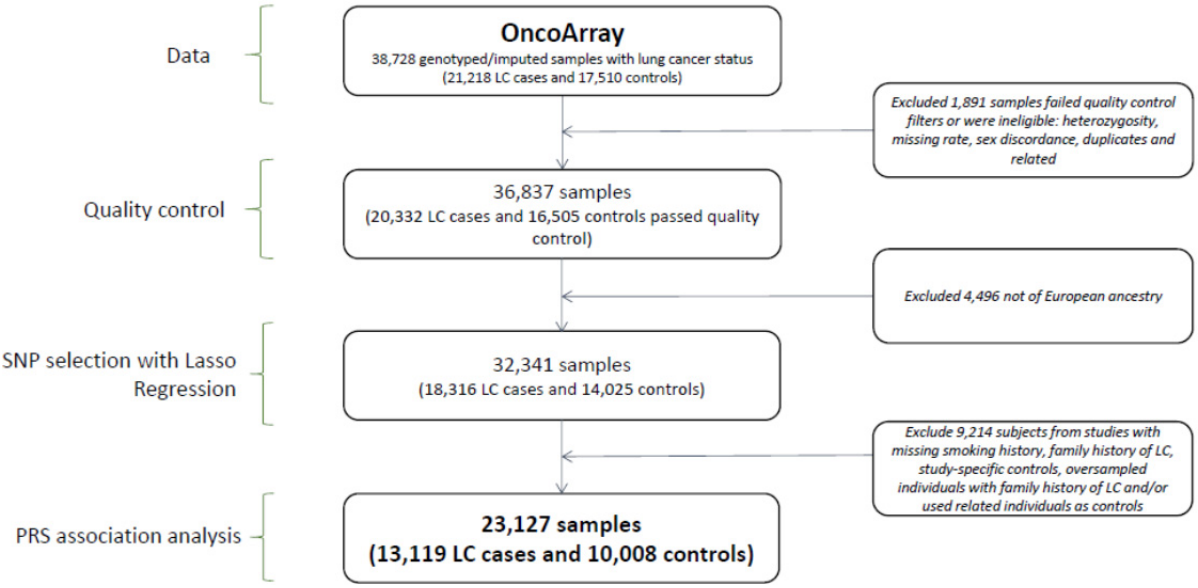
[b] AUCs among never-smokers were based on the 20% testing set that was not used in the model development

Supplementary Table 4: Average age when reaching a 5-year lung cancer absolute risk of at least 1.5% by PRS percentile in UK Biobank, stratified by smoking status and family history of lung cancer (FHLC).

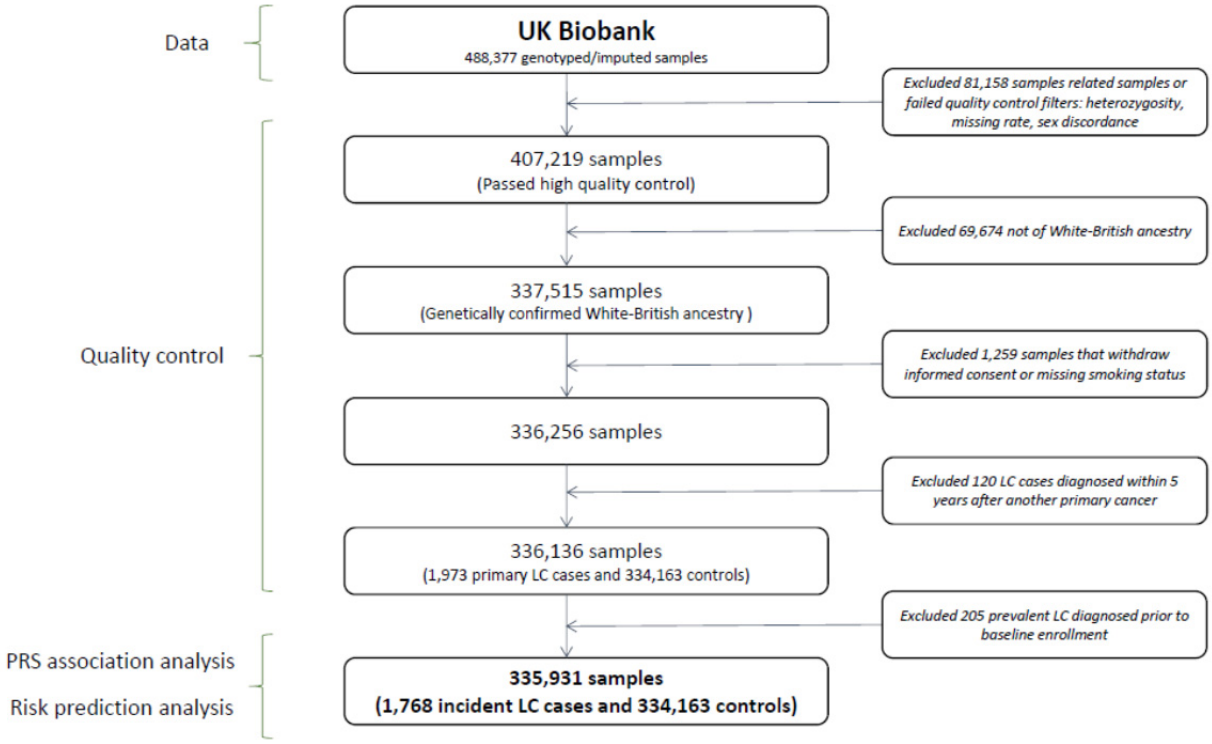
PRS Group	Overall	Ever Smokers		Former Smokers		Current Smokers	
		Without FHLC	With FHLC	Without FHLC	With FHLC	Without FHLC	With FHLC
Top 1%	59	53	51	55	52	51	48
1-5%	61	56	52	57	53	52	48
5-10%	63	57	52	58	54	55	49
Average	69	61	56	63	57	57	52

Supplementary Figure 1: Flowchart of subject exclusions during the process of quality control procedures in (a) OncoArray for model building, and (b) UK Biobank, model validation

(a) OncoArray

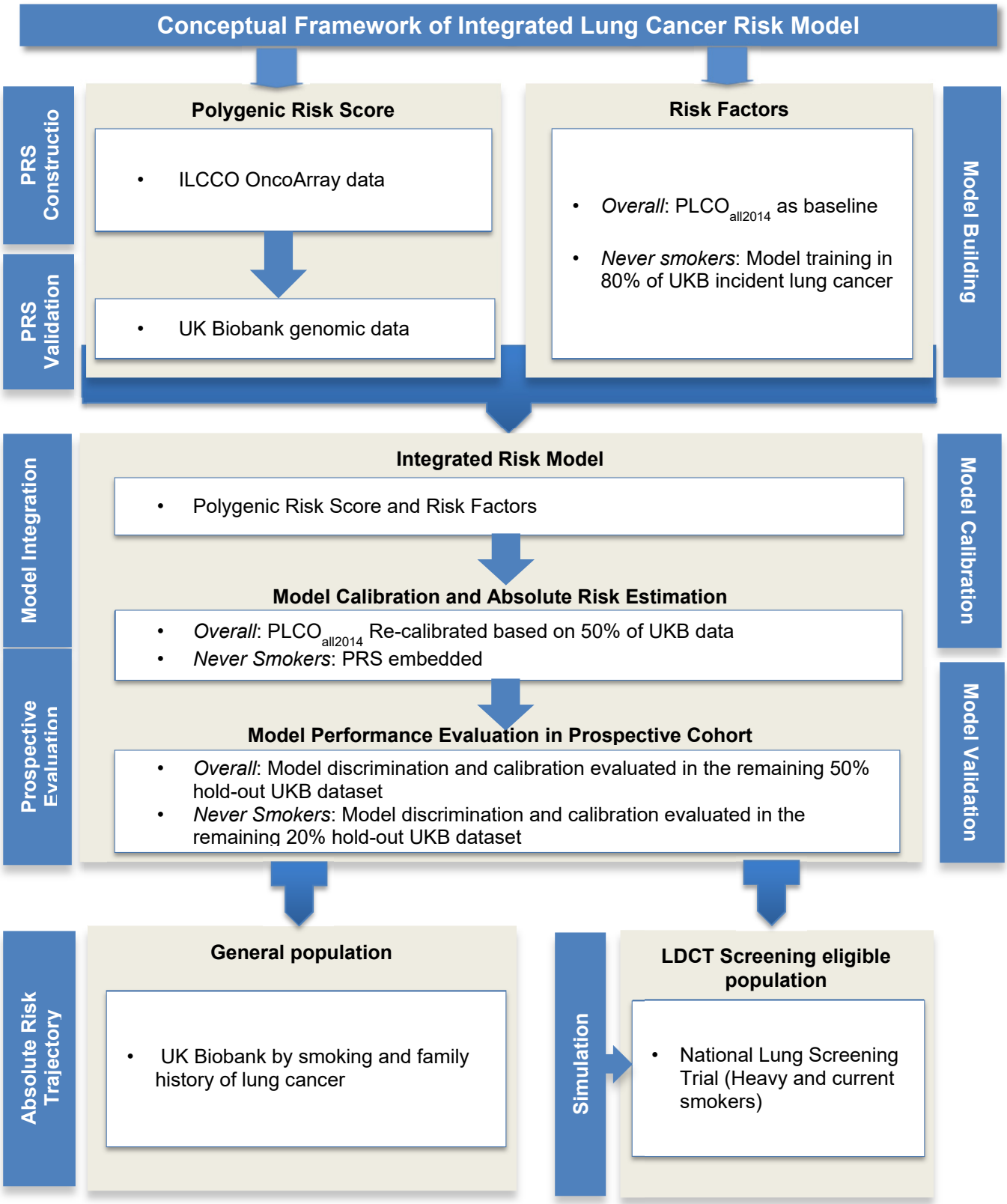


(b) UK Biobank



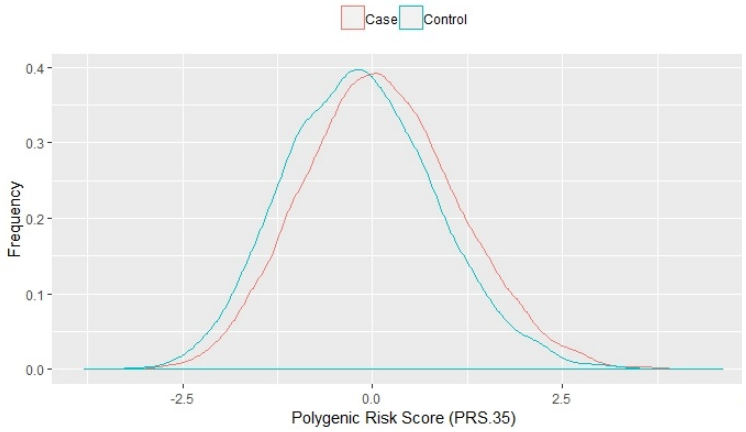
*LC, lung cancer; PRS, polygenetic risk score

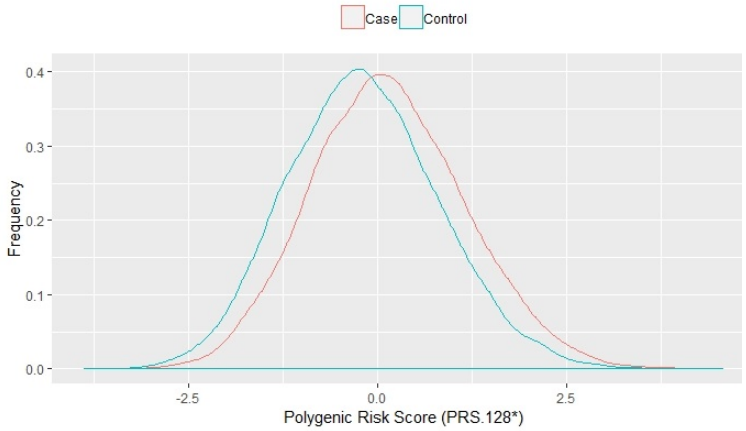
Supplementary Figure 2: Concept Framework and Analysis Flow



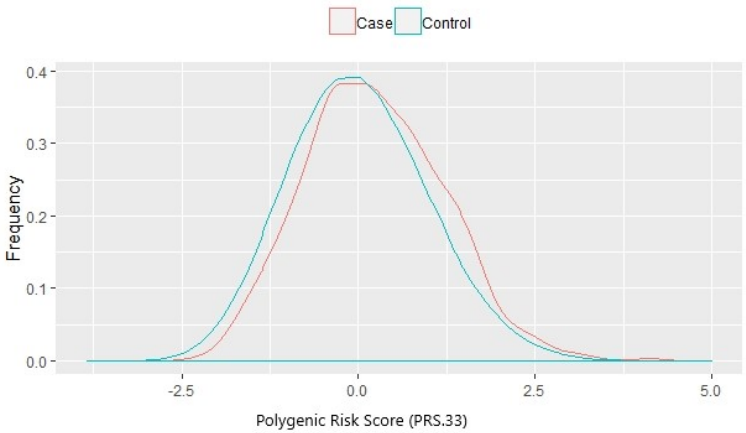
Supplementary Figure 3: PRS distribution in OncoArray and UK Biobank

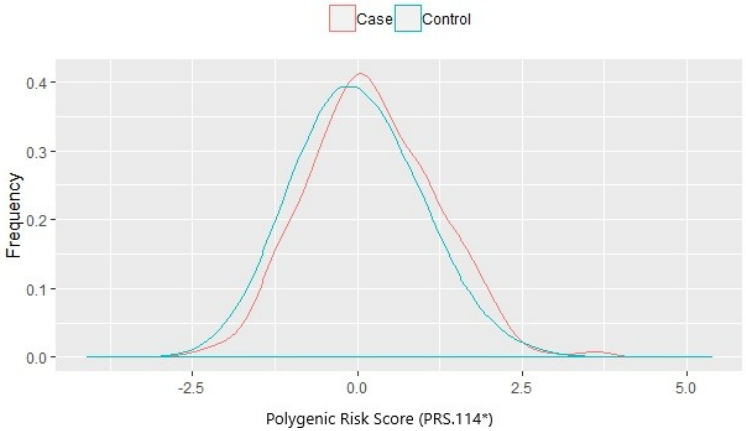
(a) OncoArray (PRS construction)





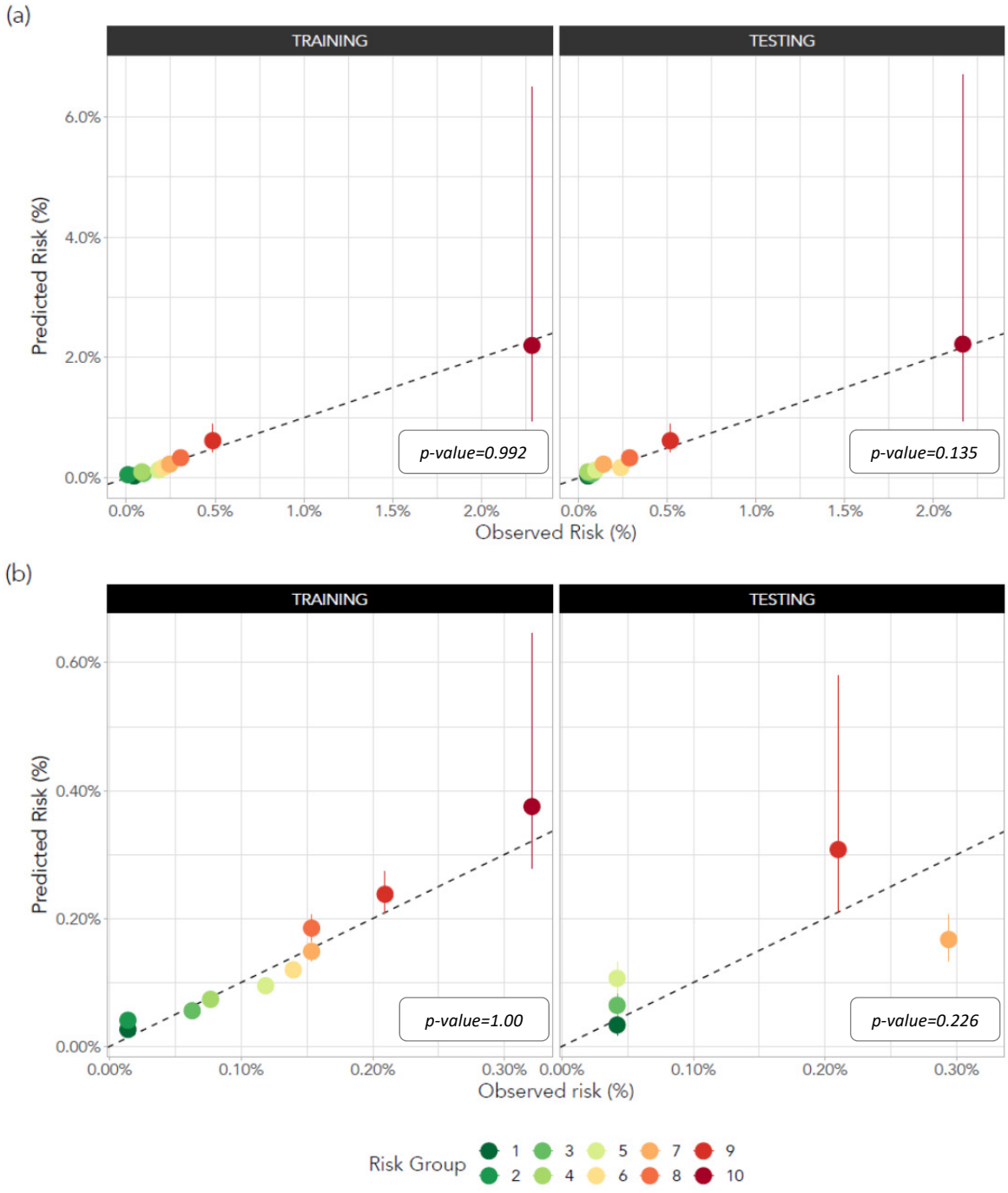
(b) UK Biobank (PRS validation)





PRS	OncoArray			UK Biobank		
	Cases	Controls	Total	Cases	Controls	Total
Mean	0.86	0.68	0.78	0.59	0.48	0.48
SD	0.53	0.52	0.54	0.50	0.50	0.50

Supplementary Figure 4: Model calibration based on UK Biobank cohort including PRS-114 and all described risk factors comparing observed versus predicted risk (a) Overall population: The data were randomly split into 50% for re-calibration and 50% as hold-out testing set for validation, and plotted based on risk deciles; **(b) Never-smokers:** the data were randomly split into 80% training and 20% testing set, where the training set was used to develop the risk model, and the model performance was evaluated in the hold-out testing set. The risks by quintiles are plotted in the hold-out testing set to reduce noise due to fewer lung cancers in each group. P-values are computed based on the Spiegelhalter's *z* statistic.



Supplementary Figure 5: The 5-year absolute risk stratified by smoking status and PRS-114 in UK Biobank. The colored lines define PRS risk groups and the patterned lines define smoking status. Red solid line represents current smokers who are at top 10% of PRS decile, red dotted line represents former smokers who are at top 10% of decile, and red dashed line represents never smokers who are at top 10% of decile. Orange solid, dotted and dashed line represent current, former and never smokers who are at 10-90% of PRS decile, respectively, wheares green solid, dotted and dashed line represent current, former and never smokers at the lowest 10% of PRS decile.