Supplementary Online Content

Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Consortium for Early Detection of Lung Cancer. Assessment of lung cancer risk on the basis of a biomarker panel of circulating proteins. *JAMA Oncol*. Published online July 12, 2018. doi:10.1001/jamaoncol.2018.2078

eMethods. Supplementary Methods

eResults. Supplementary Results

eTable 1. Discriminative Performance of the Biomarker Score in the CARET Training Study

eTable 2. Discriminative Performance of the Individual Biomarkers in the CARET Study (Training Sample) and Model Specification of the Biomarker Score

eTable 3. Discriminative Performance of the Smoking-Based and Integrated Risk Prediction Models in the Validation Study (Ever Smokers)

eTable 4. Specification of the PLCO Risk Score Based on the $PLCO_{M2012}$ Model

eTable 5. Subject Baseline Characteristics in the Training (CARET) and Validation Studies (EPIC and NSHDS)

eTable 6. Subject Baseline Characteristics in the EPIC and NSHDS Studies (Cases Diagnosed 2-10 years after Blood Collection) Used to Train the Smoking Model

eTable 7. Specification of the Smoking Models Developed on the Basis of EPIC and NSHDS Cases Diagnosed 2-10 Years after Blood Collection

eTable 8. Discriminative Performance of the Smoking-Based and Integrated Risk Prediction Models in EPIC and NSHDS Validation Samples (All Subjects Diagnosed Within 1 Year of Blood Collection, Including Never Smokers)

eTable 9. Discriminative Performance of the Smoking-Based and Integrated Risk Prediction Models in Ever Smokers from the Validation Study, by Histologic Subtypes and Stage of the Cancer

eTable 10. Apparent Discriminative Performance of 3 Risk Prediction Models in Ever Smokers from EPIC and NSHDS Validation Samples (Cases Diagnosed Within 1 Year of Blood Collection)

eFigure 1: Flow Diagram Depicting the Selection of Lung Cancer Cases Included in the Validation Study from the EPIC and NSHDS Cohorts

eFigure 2: Receiver Operating Characteristic (ROC) Curves for Each of the 5 Biomarkers in the CARET Training Study and for the 4-Marker Panel

eFigure 3. Receiver Operating Characteristic (ROC) Curves for All 4 Biomarkers and the Biomarker Score Stratified for Cases Diagnosed Within 6 Months (Panel A) and 6 to 12 Months (Panel B) of Blood Draw in the CARET Training Study

eFigure 4. Extension of the Receiver Operating Characteristic (ROC) Curve Analysis to EPIC and NSHDS Ever-Smoking Subjects With a Diagnosis Within 2 Years of Blood Collection for 2 Risk Prediction Models, Smoking Variables Only and an Integrated Model with the Smoking Variables and the Biomarker Score Combined

eFigure 5. Discriminative Performance of the Original and Reduced $PLCO_{M2012}$ Models in the PLCO Cancer Screening Trial

eFigure 6. Probability of Lung Cancer Within 1 Year Predicted from a Smoking-Based Risk Prediction Model in Ever Smokers from the EPIC Cohort and Presented for a Man According to His Possible Smoking History and Age

eFigure 7. Calibration of the Prediction Models in Ever Smokers from the EPIC and NSHDS Samples

eFigure 8. Receiver Operating Characteristic (ROC) Curve Analysis in the Validation Study (EPIC and NSHDS Subjects With Diagnosis Within 1 Year of Blood Collection) for 2 Risk Prediction Models, Smoking Variables Only and an Integrated Model with the Smoking Variables and the Biomarker Score Combined

eFigure 9. Apparent Receiver Operating Characteristic (ROC) Curve Analysis Among Ever Smokers from the Validation Study (EPIC and NSHDS Subjects With a Diagnosis Within 1 Year of Blood Collection) for 3 Risk Prediction Models With Smoking Variables Only, Biomarker Score Only, and an Integrated Model With the Smoking Variables and the Biomarker Score Combined

eFigure 10. Receiver Operating Characteristic (ROC) Curve Analysis in Ever Smokers in the Validation Study (EPIC and NSHDS Subjects With diagnosis Within 1 Year of Blood Collection) for the PLCO Risk Score Compared with the Risk Prediction Models

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Supplementary Methods

Development of a biomarker score in the CARET training study

CARET was a randomized, double-blind, placebo-controlled trial evaluating the cancer prevention efficacy and the safety of daily supplementation with beta-carotene and retinol palmitate in 18,314 persons at high risk for lung cancer.[1,2] Participants were enrolled at 6 US centers from 1985 to 1994 and were followed for cancer and mortality outcomes until 2005. Aliquots of pre-diagnostic serum samples from CARET participants previously utilized in a blinded validation study of Pro-SFTPB [23] were used to test the individual performance of CA125, CEA, HE4 and CYFRA 21-1 and develop a risk prediction biomarker score. In total, samples were assayed from 108 subjects who subsequently developed non–small-cell lung cancer (NSCLC) within 12 months after providing a blood sample, and 216 controls comprising two controls matched to each case based on age at baseline (5-yr groups), sex, baseline smoking status (current vs former), and study enrollment period.

Performance of the biomarker score in the EPIC and NSHDS validation study

The EPIC study is an ongoing multi-center prospective cohort that recruited participants between 1992 and 1998. The current study was defined amongst 267,377 participants from Greece, Netherlands, UK, France, Germany, Spain, and Italy who donated a blood sample at study recruitment. NSHDS is an ongoing prospective cohort of the general population of the Västerbotten County in Sweden. As of 2014, the cohort had recruited 99,404 study participants who donated a blood sample at recruitment.[3,4]
Incident cancer cases within the studies were identified using combination of passive and active follow-up.[3,4] Lung cancer was defined based on the International Classification of Diseases for Oncology (ICD-O-2), and included all invasive cancers that were coded as C34. The validation study included incident cases diagnosed within 1 year of blood draw, (N=67) and cases diagnosed within 2 years of blood draw (N=85) (eFigure 1).
For each index case, two controls were chosen from risk sets consisting of all cohort members alive and free of cancer (except non-melanoma skin cancer) at the time of diagnosis of the index case. Matching criteria were study center, sex, date of blood collection (±12 months), and age at blood collection (± 3 months, relaxed up to ± 5 years). In order to improve the statistical power in smoking stratified analyses, one of the controls was additionally matched based on smoking status of the index case from 5 categories; never smokers, short and long term quitters among former smokers (<10 years and ≥10 years since quitting, respectively), and light and heavy smokers among current smokers (<15 cigarettes and ≥15 cigarettes per day, respectively).
All study participants gave written informed consent to participate in the study and the research was approved by the local ethics committees in the participating countries, as well as the IARC and MD Anderson Ethical Review Committees.

Laboratory methods

Samples from all study participants for both training and testing, were sent on dry ice blinded to case-control status to the laboratory at MD Anderson Cancer Center, where they were kept below −80°C until analysis. Concentrations for Pro-SFTPB, CA125, CEA, CYFRA 21-1 and HE4 were determined using bead-based immunoassays on the MAGPIX® instrument (Luminex Corporation, Austin TX). Samples were analyzed in batches of 36 samples in duplicates with matched cases and controls in the same batch in random order. Quality control procedures included 7 calibration standards, 2 Quality Control samples, and 1 blank sample run in duplicate in each batch. The coefficients of variation (CVs) within and between batches were, 6.86% and 15.54% for CA125, 1.45% and 9.32% for CEA, 6.55% and 17.26% for Pro-SFTPB, 5.56% and 28.71% for CYFRA 21-1, and 10.334 % and 12.997% for HE4, respectively.

Statistical methods

Data for each evaluated biomarker was initially log-transformed. Data from the CARET training study was used to develop a biomarker score. Because data for CYFRA 21-1 was missing from some CARET samples due to prior sample depletion, model building employed a two-stage approach wherein the first stage involved selecting a biomarker panel using data for CA125, CEA, HE4, and Pro-SFTPB by logistic regression based on Akaike Information Criterion (AIC). The second stage involved combining the risk model attained from the first stage with data on CYFRA 21-1 using logistic regression. Using Receiver Operating Curve (ROC) analyses, we evaluated the AUC for a biomarker score with and without CYFRA 21-1 in order to establish a final biomarker score. Because of lack of additive performance, HE4 was dropped out from the biomarker panel.
The biomarker score was subsequently validated in ever smokers from EPIC and NSHDS (63 cases and 90 matched controls). To determine the extent to which the biomarker score could improve on a risk prediction model based on smoking exposure history, we fitted a smoking-model using data from EPIC and NSHDS that were not used in the validation study, defined by cases diagnosed 2 to 10 years after study recruitment with controls individually matched with the same matching criteria as in the validation study (886 ever-smoking cases and 1,349 ever-smoking controls). With use of conditional logistic regression, parameters for the smoking-model were fitted for smoking status (former vs. never, current vs. never), number of cigarettes per day for current smokers (continuous [not available in former smokers]), smoking duration (continuous in former and current smokers), and time since quitting for former smokers (continuous).
The extent to which the biomarker score and smoking-based score could discriminate between incident lung cancer cases and controls was subsequently evaluated externally and non-parametrically by assigning the respective risk scores to each participant in the validation study (cases diagnosed 0 to 1 year after blood draw, and subsequently expanded to 0 to 2 years after blood draw). In addition, in order to evaluate the potential of combining the two risk scores, an integrated risk prediction model was developed by

fitting a conditional logistic regression model using the smoking-based score and biomarker score as two separate covariates in the validation study.

In order to provide absolute risk and discrimination estimates reflecting the background population of the validation studies, we used the pseudo-likelihood approach of Samuelsen *et al.* [5]. We subsequently modelled the cumulative hazards of lung cancer using flexible parametric survival models[6] for the smoking and integrated risk prediction models to estimate the baseline hazards and the absolute risks of lung cancer over 1 and 2 years.

The apparent discriminatory accuracy of the smoking-based score, the biomarker score, and the integrated risk prediction model, were evaluated using ROC analyses in the validation study. The 95% confidence intervals were estimated using 2,000 stratified bootstrap replicates, and differences in AUC estimates were determined using nonparametric methods.[7] To determine the fraction of future lung cancer cases that would have been identified using the different models, we estimated the sensitivity of each model at the specificity level obtained by applying the USPSTF screening eligibility criteria to each subject in the validation study. However, because the controls were individually matched to the cases in the validation study, the apparent specificity estimates will be biased. The final ROC analysis was thus conducted using predicted 1-year and 2-year lung cancer risks as scoring rule, with population-based sensitivity and specificity estimates weighted according to their sampling probability.

Statistical significance was assumed at a two-sided P-value below 0.05. All statistical analyses were conducted using R 3.3.0 (R Core Team (2016)) and STATA v.14.2 (StataCorp LP, College Station, TX).

Comparison between the smoking-based risk prediction model used in the current paper and a validated lung cancer risk prediction model.

We compared the discriminative performance of our smoking-based risk prediction model with a model based on the validated $PLCO_{M2012}$ lung cancer risk prediction model developed in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial [8].

The $PLCO_{M2012}$ model includes 11 variables and a constant, and was developed for ever smoking subjects only.

In the validation study, we did not have access to data on history of chronic obstructive pulmonary disease (COPD), family history of lung cancer, nor intensity of smoking among former smokers and could therefore not include these variables in the comparison model. Since all of our subjects are of white ethnicity, and were selected based on no previous history of cancer, these coefficients were not included in the comparison model. The coefficients we used to build the risk score based on the $PLCO_{M2012}$ model are shown in eTable 4.

First, We compared the discriminative performance of the reduced PLCO model (excluding history of COPD and family history of lung cancer and intensity among former smokers) to the original $PLCO_{M2012}$ model using the data where the $PLCO_{M2012}$ model was validated in (80,375 persons in the PLCO control and intervention groups who had ever smoked). The AUC were very similar: 0.80 and 0.78 for the original and modified PLCO models respectively (eFigure 5).

The comparison between our smoking-based risk prediction model and the reduced $PLCO_{M2012}$ model was performed among ever smokers in EPIC and NSHDS who were diagnosed with lung cancer within 1 year of blood collection. Due to some missing values on education, body-mass index, and intensity of smoking, we performed these analyses on 136 subjects (57 cases and 79 controls).

We used the PLCO risk score as a covariate in our weighted flexible parametric survival models as described in the methods section of the manuscript for the smoking-based score. An integrated risk prediction model was also built by fitting a model using the PLCO risk score and the biomarker risk score as separate covariates. The discriminatory accuracy of our smoking-based risk prediction model, the PLCO-based risk prediction model, and the two integrated risk prediction models, were evaluated using ROC analyses.

eResults. Supplementary Results

The baseline characteristics of subjects in the training and validation studies are presented in eTable 5.

Training of the smoking risk prediction model

The characteristics of the EPIC and NSHDS sample set used to train the smoking model are presented in eTables 6 and 7. The 1-year probability of lung cancer predicted from the smoking model was similar in the validation study and in the full EPIC cohort (1,161 cases and 114,204 controls; eFigure 6).

Development of a biomarker score based on the CARET training study

The discrimination performances of each candidate biomarker in the CARET training study are presented in eTable 1 and eTable 2. Their AUC estimates ranged from 0.60 (95% CI: 0.53-0.67, CA125) to 0.70 (95% CI: 0.64-0.77, ProSFTPB) at P-value < 0.05. Based on AIC, HE4 was excluded from the model, and the final biomarker score based on four markers (CA125, CEA, CYFRA 21-1 and Pro-SFTPB) yielded an AUC of 0.80 (95% CI 0.72-0.87) in the CARET training study.

Performance of the risk prediction models in discriminating between future lung cancer cases and controls in the validation study

All the weighted (according to the sampling probability) models were well calibrated (eFigure 7).

Integrating never smokers in the ROC analyses yielded comparable findings to the overall analysis with a 10% increase in AUC when biomarkers were combined to smoking variables (eTable 8, eFigure 8).

Among ever smokers, the integrated risk prediction model discriminated similarly for early and late lung cancer stages and the two most prevalent histologic types, with a consistently higher AUC than the smoking-based risk prediction model (eTable 9).

Apparent (unweighted) discrimination estimates among subjects diagnosed within 1 year of blood collection are provided in eFigure 9 and eTable 10. This analysis showed a 13% improvement in AUC from the smoking model (AUC=0.77 (95% CI: 0.70-0.85)) to the integrated model (AUC=0.90 (95% CI: 0.86-0.95)).

Comparison between the smoking-based risk prediction model used in the current paper and a validated lung cancer risk prediction model.

In the validation study, the reduced PLCO model yielded similar AUC estimates to our smoking-based risk prediction model (P for difference in AUC > 0.5, eFigure 10).

REFERENCES

1. Goodman GE, Thornquist MD, Balmes J, et al. The Beta-Carotene and Retinol Efficacy Trial: incidence of lung cancer and cardiovascular disease mortality during 6-year follow-up after stopping beta-carotene and retinol supplements. *J Natl Cancer Inst.* 2004;96(23):1743-1750.
2. Omenn GS, Goodman GE, Thornquist MD, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *N Engl J Med.* 1996;334(18):1150-1155.
3. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* 2002;5(6B):1113-1124.
4. Bingham S, Riboli E. Diet and cancer--the European Prospective Investigation into Cancer and Nutrition. *Nat Rev Cancer.* 2004;4(3):206-215.
5. Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika.* 1997;84(2):379-394.
6. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med.* 2002;21(15):2175-2197.
7. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837-845.
8. Tammemagi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. *N Engl J Med.* 2013;368(8):728-736.

eTable 1. Discriminative Performance of the Biomarker Score in the CARET Training Study

	Biomarker score		CA125		CEA		Pro-SFTPB		CYFRA21-1	
	AUC	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**
All	0.80	[0.72-0.87]	0.60	[0.53-0.67]	0.69	[0.62-0.75]	0.70	[0.64-0.77]	0.66	[0.56-0.75]
Men	0.78	[0.68-0.88]	0.63	[0.55-0.71]	0.66	[0.58-0.74]	0.71	[0.64-0.78]	0.68	[0.58-0.79]
Women	0.83	[0.70-0.96]	0.50	[0.37-0.63]	0.76	[0.65-0.88]	0.70	[0.58-0.82]	0.60	[0.42-0.77]
Stage										
Stage I-II	0.68	[0.45-0.90]	0.62	[0.47-0.76]	0.58	[0.42-0.74]	0.66	[0.53-0.79]	0.55	[0.32-0.78]
Stage III-IV	0.83	[0.75-0.91]	0.61	[0.52-0.70]	0.73	[0.65-0.82]	0.71	[0.63-0.79]	0.68	[0.58-0.79]
Time from blood collection to subsequent diagnosis										
0-6 months	0.86	[0.76-0.96]	0.67	[0.56-0.78]	0.73	[0.63-0.84]	0.76	[0.67-0.85]	0.74	[0.62-0.87]
>6-12 months	0.77	[0.66-0.88]	0.56	[0.47-0.65]	0.66	[0.57-0.75]	0.67	[0.60-0.75]	0.59	[0.46-0.71]
Pack-years (PY)										
PY >30	0.81	[0.73-0.89]	0.59	[0.51-0.67]	0.69	[0.62-0.76]	0.70	[0.63-0.76]	0.65	[0.55-0.75]
PY 20 to 30	0.80	[0.54-1.05]	0.67	[0.49-0.84]	0.70	[0.50-0.90]	0.75	[0.59-0.91]	0.74	[0.48-1.00]
Histological subtype										
ADC	0.79	[0.67-0.92]	0.55	[0.44-0.67]	0.70	[0.59-0.80]	0.64	[0.54-0.74]	0.59	[0.42-0.76]
Other NSCLC	0.80	[0.68-0.93]	0.57	[0.43-0.70]	0.70	[0.57-0.83]	0.76	[0.65-0.87]	0.68	[0.53-0.83]
SCC	0.79	[0.62-0.96]	0.67	[0.56-0.78]	0.67	[0.56-0.79]	0.74	[0.63-0.84]	0.69	[0.53-0.86]

Abbreviations: ADC, Adenocarcinoma; NSCLC, Non small-cell lung cancer; SCC, Squamous cell carcinoma.

eTable 2. Discriminative Performance of the Individual Biomarkers in the CARET Study (Training Sample) and Model Specification of the Biomarker Score

	Discriminative performance		Model specification of the biomarker score		Odds Ratio Per unit increase	
Biomarker	**AUC**	**95% CI**	**Beta-estimate**	**95% CI**	**Odds atio**	**95% CI**
CA125	0.60	[0.53-0.67]	0.4730	[0.0886 - 0.8583]	1.55	[1.16-2.08]
CEA	0.69	[0.62-0.75]	0.6531	[0.1364 - 1.1698]	2.35	[1.68-3.29]
CYFRA 21-1	0.66	[0.56-0.75]	0.2612	[-0.1601 - 0.6825]	1.85	[1.25-2.76]
ProSFTPB	0.70	[0.64-0.77]	0.9238	[0.3627 - 1.4849]	2.55	[1.82-3.59]
HE4	0.65	[0.58-0.71]	N/A		1.61	[1.13-2.28]

eTable 3. Discriminative Performance of the Smoking-Based and Integrated Risk Prediction Models in the Validation Study (Ever Smokers)

	Cases	Controls	Specificity of USPSTF criteria	Sensitivity of USPSTF criteria	Risk model	AUC	95% CI	Sensitivity at USPSTF Specificity	95% CI	Specificity at USPSTF Sensitivity	95% CI
All	63	90	0.83	0.42	Smoking[c]	0.73	[0.64-0.82]	0.43	[0.23-0.65]	0.86	[0.72-0.94]
					Smoking + Biomarkers[d]	0.83	[0.76-0.90]	0.63	[0.49-0.76]	0.95	[0.85-0.99]
Men	42	63	0.76	0.55	Smoking	0.74	[0.63-0.85]	0.57	[0.28-0.83]	0.76	[0.58-0.89]
					Smoking + Biomarkers	0.84	[0.76-0.93]	0.79	[0.62-0.90]	0.92	[0.79-0.98]
Women	21	27	0.95	0.15	Smoking	0.69	[0.52-0.86]	0.29	[0.01-0.88]	0.95	[0.74-1.00]
					Smoking + Biomarkers	0.82	[0.69-0.95]	0.38	[0.15-0.66]	1.00	[N/A-N/A]
Former smokers	24	43	0.93	0.35	Smoking	0.79	[0.66-0.92]	0.50	[0.23-0.80]	0.97	[0.80-1.00]
					Smoking + Biomarkers	0.85	[0.75-0.96]	0.58	[0.37-0.81]	0.99	[0.83-1.00]
Current smokers	39	47	0.76	0.46	Smoking	0.68	[0.55-0.81]	0.51	[0.25-0.75]	0.76	[0.54-0.92]
					Smoking + Biomarkers	0.88	[0.80-0.95]	0.85	[0.66-0.93]	0.97	[0.82-1.00]
NSCLC Cases	51	90	0.83	0.40	Smoking	0.71	[0.61-0.80]	0.39	[0.20-0.62]	0.87	[0.65-0.90]
					Smoking + Biomarkers	0.83	[0.76-0.91]	0.67	[0.52-0.79]	0.98	[0.88-0.99]
Heavy smokers[a]	20	17	N/A	N/A	Smoking	0.56	[0.35-0.77]	N/A		N/A	
					Smoking + Biomarkers	0.77	[0.60-0.94]	N/A		N/A	
Light smokers[b]	19	31	N/A	N/A	Smoking	0.63	[0.45-0.81]	N/A		N/A	
					Smoking + Biomarkers	0.89	[0.79-0.99]	N/A		N/A	
0-6 months from BC	28	90	0.83	0.44	Smoking	0.70	[0.58-0.82]	0.43	[0.20-0.68]	0.86	[0.63-0.92]
					Smoking + Biomarkers	0.81	[0.71-0.92]	0.64	[0.45-0.80]	0.90	[0.73-0.98]
6-12 months from BC	35	90	0.83	0.40	Smoking	0.75	[0.65-0.85]	0.43	[0.20-0.68]	0.87	[0.71-0.94]
					Smoking + Biomarkers	0.85	[0.77-0.93]	0.63	[0.45-0.78]	0.96	[0.87-0.99]

Abbreviation: BC, Blood collection; NSCLC, Non small-cell lung cancer.

[a]Heavy smoker: current smokers that smoke ≥30 pack-years; [b]Light smokers: current smokers that smoke < 30 pack-years; [c]Smoking model: logistic model including smoking status, smoking duration, mean quantity of cigarettes smoked/day (for current smokers), time since quitting smoking (for former smokers) fitted in EPIC and NSHDS samples including cases diagnosed between 2 to 10 years from blood draw; [d]Smoking + Biomarkers model: logistic model including smoking score from the smoking model and the biomarker score fitted in the CARET data.

eTable 4. Specification of the PLCO Risk Score Based on the $PLCO_{M2012}$ Model

Variables	Beta Coefficient
Age, per 1–yr increase	0.0778868
Education, per increase of 1 level	−0.0812744
Body-mass index, per 1-unit increase	−0.0274194
Smoking status (current vs. former)	0.2597431
Smoking intensity	−1.822606
Duration of smoking, per 1-yr increase	0.0317321
Smoking quit time, per 1-yr increase	−0.0308572
Model constant	−4.532506

eTable 5. Subject Baseline Characteristics in the Training (CARET) and Validation Studies (EPIC and NSHDS)

		Training study (CARET)		Validation study (EPIC and NSHDS)			
				Diagnostic 0 to 1 year from BC		Diagnostic 0 to 2 year from BC	
	N (%)	Cases	Controls	Cases	Controls	Cases	Controls
Overall		108	216	67	126	152	288
Sex	Male	75 (69.4)	150 (69.4)	43 (64.2)	79 (62.7)	93 (61.2)	172 (59.7)
	Female	33 (30.6)	66 (30.6)	24 (35.8)	47 (37.3)	59 (38.8)	116 (40.3)
Age, years	≤40	-	-	3 (4.5)	6 (4.8)	3 (2)	6 (2.1)
	40-50	2 (1.9)	4 (1.9)	7 (10.4)	14 (11.1)	14 (9.2)	27 (9.4)
	50-60	35 (32.4)	72 (33.3)	30 (44.8)	55 (43.7)	64 (42.1)	124 (43.1)
	60-70	69 (63.9)	136 (63.0)	22 (32.8)	42 (33.3)	57 (37.5)	106 (36.8)
	>70	2 (1.9)	4 (1.9)	5 (7.5)	9 (7.1)	14 (9.2)	25 (8.7)
Years from BC to diagnosis	0-0.5	40 (37.0)	-	31 (46.3)	-	31 (20.4)	-
	0.5-1	68 (63.0)	-	36 (53.7)	-	36 (23.7)	-
	1-2	-	-	-	-	85 (55.9)	-
Smoking status	Never	-	-	4 (6)	36 (28.6)	16 (10.5)	80 (27.8)
	Former	36 (33.3)	72 (33.3)	24 (35.8)	43 (34.1)	47 (30.9)	99 (34.4)
	Current	72 (66.7)	144 (66.7)	39 (58.2)	47 (37.3)	89 (58.6)	109 (37.8)
Histological subtype	ADC	40 (37.0)	-	23 (34.3)	-	56 (36.8)	-
	SCC	38 (35.2)	-	17 (25.4)	-	32 (21.1)	-
	Other	30 (27.8)	-	27 (40.3)	-	64 (42.1)	-
Stage	I and II	26 (24.1)	-	11 (16.4)	-	19 (12.5)	-
	III and IV	64 (59.3)	-	36 (53.7)	-	77 (50.7)	-
	Unknown	18 (16.7)	-	20 (29.9)	-	56 (36.8)	-
Eligible for lung cancer screening (USPSTF)	Not Eligible	29 (26.9)	57 (26.4)	40 (59.7)	104 (82.5)	90 (59.2)	235 (81.6)
	Eligible	79 (73.1)	159 (73.6)	26 (38.8)	20 (15.9)	60 (39.5)	50 (17.4)
	N/A	-	-	1 (1.5)	2 (1.6)	2 (1.3)	3 (1)

eTable 6. Subject Baseline Characteristics in the EPIC and NSHDS Studies (Cases Diagnosed 2-10 years after Blood Collection) Used to Train the Smoking Model

			Training sample for the smoking-model (EPIC and NSHDS)	
	N (%)		**Cases**	**Controls**
Overall			1008	1873
Sex	Male		605 (60)	1088 (58.1)
	Female		403 (40)	785 (41.9)
Age, years	≤40		20 (2.0)	45 (2.4)
	40-50		167 (16.6)	315 (16.8)
	50-60		430 (42.7)	787 (42.0)
	60-70		315 (31.2)	599 (32.0)
	>70		76 (7.5)	127 (6.8)
Years from blood collection to diagnosis	2-5		351 (34.8)	-
	5-10		657 (65.2)	-
Smoking status	Never		122 (12.1)	524 (28.0)
	Former		296 (29.4)	606 (32.3)
	Current		590 (58.5)	743 (39.7)
Histological subtype	ADC		366 (36.3)	-
	SCC		200 (19.8)	-
	Other		442 (43.6)	
Eligible for lung cancer screening (USPSTF)	Not Eligible		688 (68.7)	1599 (86.6)
	Eligible		313 (31.3)	248 (13.4)
	N/A		7	26

ADC: Adenocarcinoma; SCC: Squamous cell carcinoma

eTable 7. Specification of the Smoking Models Developed on the Basis of EPIC and NSHDS Cases Diagnosed 2-10 Years after Blood Collection

Variables included in the smoking score	Beta estimates for the smoking-score	OR	95% CI
Ever smokers			
Current vs Former	0.761658	2.14	[1.45-3.17]
Duration of smoking (years)	0.032454	1.03	[1.01-1.06]
Time since smoking cessation (years) for former smokers	-0.032156	0.97	[0.94-0.99]
Number of cigarette smoked per day for current smokers	0.067843	1.07	[1.05-1.09]
All subjects			
Former vs never	1.635706	5.13	[3.44-7.66]
Current vs never	2.276509	9.74	[6.63-14.33]
Duration of smoking (years) among ever smokers	0.038906	1.04	[1.02-1.06]
Time since smoking cessation (years) for former smokers	-0.027166	0.97	[0.95-0.99]
Number of cigarette smoked per day for current smokers	0.066884	1.07	[1.05-1.09]

eTable 8. Discriminative Performance of the Smoking-Based and Integrated Risk Prediction Models in EPIC and NSHDS Validation Samples (All Subjects Diagnosed Within 1 Year of Blood Collection, Including Never Smokers)

	Cases	Controls	Specificity of USPSTF criteria	Sensitivity of USPSTF criteria	Risk model	AUC	95% CI	Sensitivity at USPSTF Specificity	95% CI	Specificity at USPSTF Sensitivity	95% CI
All	67	126	0.89	0.39	Smoking[c]	0.78	[0.71-0.85]	0.40	[0.22-0.60]	0.90	[0.79-0.96]
					Smoking + Biomarkers[d]	0.88	[0.83-0.93]	0.60	[0.47-0.72]	0.96	[0.89-0.99]
Men	43	79	0.81	0.53	Smoking	0.77	[0.68-0.87]	0.58	[0.23-0.87]	0.83	[0.69-0.92]
					Smoking + Biomarkers	0.87	[0.79-0.94]	0.77	[0.61-0.88]	0.94	[0.84-0.98]
Women	24	47	0.97	0.13	Smoking	0.78	[0.66-0.90]	0.25	[0.07-0.56]	0.97	[0.86-1.00]
					Smoking + Biomarkers	0.89	[0.82-0.97]	0.33	[0.16-0.54]	1.00	[N/A-N/A]
Never smokers	4	36	N/A	N/A	N/A	N/A		N/A		N/A	
Former smokers	24	43	0.93	0.35	Smoking	0.79	[0.67-0.92]	0.50	[0.20-0.73]	0.97	[0.82-1.00]
					Smoking + Biomarkers	0.85	[0.75-0.96]	0.50	[0.26-0.74]	0.99	[0.83-1.00]
Current smokers	39	47	0.76	0.46	Smoking	0.69	[0.56-0.82]	0.49	[0.18-0.67]	0.76	[0.52-0.91]
					Smoking + Biomarkers	0.87	[0.80-0.95]	0.82	[0.65-0.92]	0.97	[0.76-1.00]
NSCLC Cases	55	126	0.89	0.37	Smoking	0.76	[0.68-0.84]	0.40	[0.22-0.60]	0.91	[0.79-0.96]
					Smoking + Biomarkers	0.88	[0.82-0.93]	0.62	[0.48-0.74]	0.99	[0.91-0.99]
Heavy smokers[a]	20	17	N/A	N/A	Smoking	0.57	[0.36-0.78]	N/A		N/A	
					Smoking + Biomarkers	0.75	[0.58-0.93]	N/A		N/A	
Light smokers[b]	19	31	N/A	N/A	Smoking	0.65	[0.47-0.83]	N/A		N/A	
					Smoking + Biomarkers	0.89	[0.79-0.99]	N/A		N/A	
0-6 months from BC	31	126	0.89	0.40	Smoking	0.74	[0.64-0.85]	0.42	[0.21-0.65]	0.90	[0.75-0.97]
					Smoking + Biomarkers	0.86	[0.79-0.94]	0.58	[0.40-0.74]	0.93	[0.85-0.97]
6-12 months from BC	36	126	0.89	0.39	Smoking	0.82	[0.74-0.89]	0.39	[0.19-0.63]	0.91	[0.76-0.94]
					Smoking + Biomarkers	0.89	[0.84-0.95]	0.61	[0.44-0.76]	0.98	[0.90-0.99]

Abbreviation: BC, Blood collection; NSCLC, Non small-cell lung cancer.
[a]Heavy smoker: current smokers that smoke ≥30 pack-years;
[b]Light smokers: current smokers that smoke < 30 pack-years;
[c]Smoking model: logistic model including smoking status, smoking duration, mean quantity of cigarettes smoked/day (for current smokers), time since quitting smoking (for former smokers) fitted in EPIC and NSHDS samples including cases diagnosed between 2 to 10 years from blood draw.
[d]Smoking + Biomarkers model: logistic model including smoking score from the smoking model and the biomarker score fitted in the CARET data

eTable 9. Discriminative Performance of the Smoking-Based and Integrated Risk Prediction Models in Ever Smokers from the Validation Study, by Histologic Subtypes and Stage of the Cancer

	Cases	Controls	Specificity of USPSTF criteria	Sensitivity of USPSTF criteria	Risk model	AUC	95% CI	Sensitivity at USPSTF Specificity	95% CI	Specificity at USPSTF Sensitivity	95% CI
Adenocarcinomas	20	90	0.83	0.20	Smoking[a]	0.63	[0.50-0.75]	0.15	[0.03-0.42]	0.79	[0.58-0.92]
					Smoking + Biomarkers[b]	0.86	[0.77-0.95]	0.70	[0.47-0.87]	1.00	[N/A-N/A]
Squamous cell carcinoma	16	90	0.83	0.56	Smoking	0.81	[0.68-0.93]	0.63	[0.31-0.87]	0.87	[0.70-0.95]
					Smoking + Biomarkers	0.90	[0.82-0.99]	0.81	[0.57-0.95]	0.98	[0.71-1.00]
Stage I/II	11	90	0.83	0.36	Smoking	0.63	[0.46-0.80]	0.18	[0.02-0.58]	0.79	[0.55-0.93]
					Smoking + Biomarkers	0.78	[0.44-0.93]	0.55	[0.26-0.81]	0.95	[0.50-1.00]
Stage IIII/IV	33	90	0.83	0.44	Smoking	0.72	[0.61-0.83]	0.52	[0.29-0.74]	0.87	[0.72-0.95]
					Smoking + Biomarkers	0.83	[0.74-0.92]	0.70	[0.51-0.84]	0.96	[0.85-0.99]

[a]Smoking model: logistic model including smoking status, smoking duration, mean quantity of cigarettes smoked/day (for current smokers), time since quitting smoking (for former smokers) fitted in EPIC and NSHDS samples including cases diagnosed between 2 to 10 years from blood draw.
[b]Smoking + Biomarkers model: logistic model including smoking score from the smoking model and the biomarker score fitted in the CARET data.

eTable 10. Apparent Discriminative Performance of 3 Risk Prediction Models in Ever Smokers from EPIC and NSHDS Validation Samples (Cases Diagnosed Within 1 Year of Blood Collection)

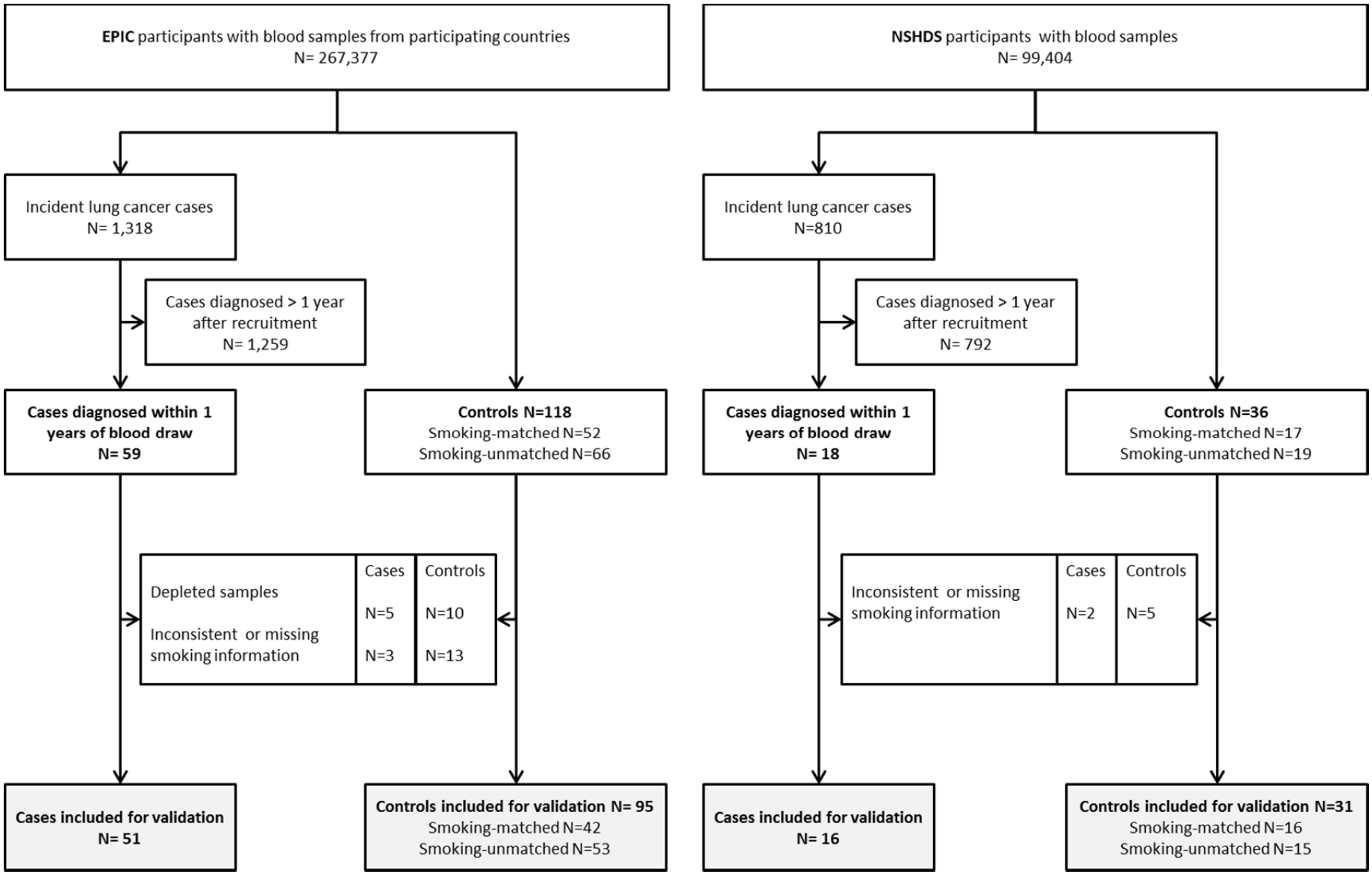
	Cases	Controls	Specificity of USPSTF criteria	95% CI	Sensitivity of USPSTF criteria	95% CI	Risk model	AUC	95% CI	Sensitivity at USPSTF Specificity	95% CI	Specificity at USPSTF Sensitivity	95% CI
All	61	88	0.77	[0.61-0.86]	0.42	[0.26-0.55]	Smoking[c]	0.77	[0.70-0.85]	0.64	[0.46-0.77]	0.90	[0.80-0.97]
							Biomarkers[d]	0.90	[0.84-0.95]	0.84	[0.72-0.95]	0.98	[0.93-1.00]
							Smoking + Biomarkers[e]	0.90	[0.86-0.95]	0.92	[0.74-0.98]	0.97	[0.91-1.00]
Men	40	63	0.73	[0.51-0.84]	0.55	[0.34-0.7]	Smoking	0.81	[0.72-0.89]	0.73	[0.58-0.90]	0.87	[0.76-0.95]
							Biomarkers	0.90	[0.84-0.96]	0.85	[0.73-0.98]	0.95	[0.89-1.00]
							Smoking + Biomarkers	0.91	[0.85-0.96]	0.93	[0.73-1.00]	0.95	[0.90-1.00]
Women	21	25	0.88	[0.71-0.98]	0.15	[0.02-0.34]	Smoking	0.67	[0.51-0.83]	0.33	[0.10-0.57]	0.96	[0.80-1.00]
							Biomarkers	0.89	[0.79-0.99]	0.71	[0.48-0.95]	1.00	[1.00-1.00]
							Smoking + Biomarkers	0.93	[0.85-1.00]	0.95	[0.38-1.00]	1.00	[1.00-1.00]
Former smokers	23	34	0.88	[0.65-0.97]	0.36	[0.13-0.57]	Smoking	0.80	[0.68-0.92]	0.61	[0.30-0.83]	0.94	[0.85-1.00]
							Biomarkers	0.82	[0.71-0.94]	0.65	[0.35-0.83]	0.97	[0.85-1.00]
							Smoking + Biomarkers	0.87	[0.77-0.96]	0.70	[0.39-0.91]	0.97	[0.88-1.00]
Current smokers	35	43	0.63	[0.42-0.79]	0.46	[0.26-0.64]	Smoking	0.72	[0.61-0.83]	0.66	[0.49-0.89]	0.81	[0.65-0.95]
							Biomarkers	0.94	[0.90-0.99]	0.97	[0.89-1.00]	1.00	[0.93-1.00]
							Smoking + Biomarkers	0.95	[0.90-0.99]	0.97	[0.91-1.00]	1.00	[1.00-1.00]
NSCLC Cases	49	71	0.77	[0.6-0.87]	0.40	[0.23-0.54]	Smoking	0.76	[0.67-0.85]	0.63	[0.43-0.78]	0.90	[0.79-0.97]
							Biomarkers	0.90	[0.84-0.96]	0.86	[0.71-0.98]	0.99	[0.94-1.00]
							Smoking + Biomarkers	0.90	[0.84-0.96]	0.90	[0.69-0.98]	0.97	[0.92-1.00]
Heavy smokers[a]	12	15	N/A		N/A		Smoking	0.56	[0.31-0.81]	N/A		N/A	
							Biomarkers	0.91	[0.80-1.00]	N/A		N/A	
							Smoking + Biomarkers	0.89	[0.77-1.00]	N/A		N/A	
Light smokers[b]	16	18	N/A		N/A		Smoking	0.76	[0.59-0.93]	N/A		N/A	
							Biomarkers	0.99	[0.97-1.00]	N/A		N/A	
							Smoking + Biomarkers	0.99	[0.97-1.00]	N/A		N/A	
0-6 months from BC	26	38	0.76	[0.53-0.89]	0.44	[0.21-0.64]	Smoking	0.72	[0.60-0.85]	0.46	[0.27-0.77]	0.82	[0.60-1.00]
							Biomarkers	0.88	[0.79-0.97]	0.77	[0.62-0.96]	1.00	[0.89-1.00]
							Smoking + Biomarkers	0.88	[0.80-0.97]	0.85	[0.65-0.96]	0.97	[0.89-1.00]
6-12 months from BC	35	50	0.78	[0.59-0.89]	0.40	[0.2-0.57]	Smoking	0.81	[0.72-0.91]	0.69	[0.51-0.86]	0.92	[0.84-0.98]
							Biomarkers	0.91	[0.84-0.97]	0.89	[0.69-1.00]	0.98	[0.92-1.00]
							Smoking + Biomarkers	0.92	[0.86-0.98]	0.97	[0.69-1.00]	0.96	[0.90-1.00]

Abbreviation: BC, Blood collection; NSCLC, Non small-cell lung cancer.

[a]Heavy smokers: current smokers that smoke ≥30 pack-years; [b]Light smokers: current smokers that smoke < 30 pack-years

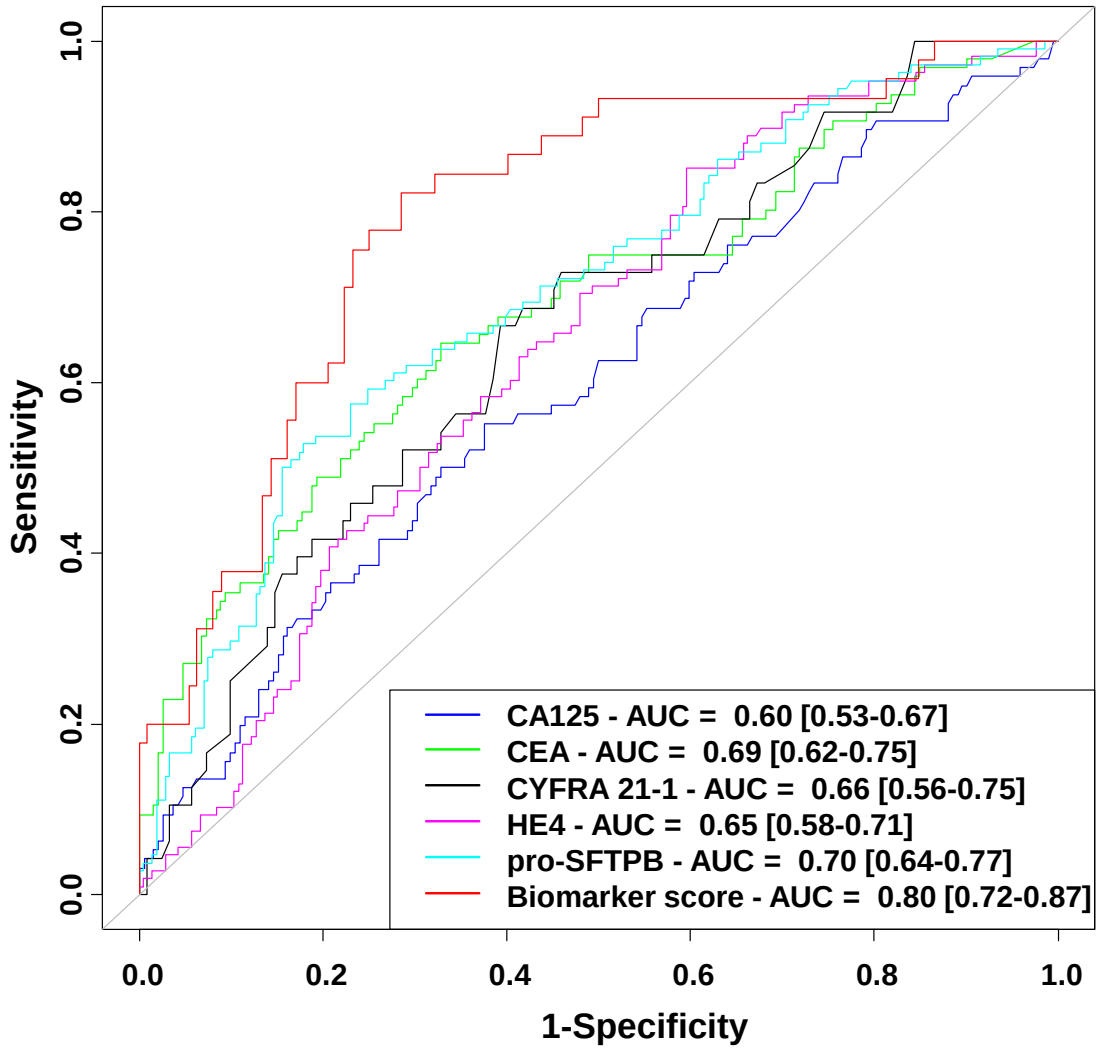
[c]Smoking model: logistic model including smoking status, smoking duration mean quantity of cigarettes smoked/day (for current smokers), time since quitting smoking (for former smokers) fitted in EPIC and NSHDS samples including cases diagnosed between 2 to 10 years from blood draw.
[d]Biomarkers model: logistic model including the biomarker score fitted in the CARET data.
[e]Smoking + Biomarkers model: logistic model including smoking score from the smoking model and the biomarker score fitted in the CARET data.

eFigure 1: Flow Diagram Depicting the Selection of Lung Cancer Cases Included in the Validation Study from the EPIC and NSHDS Cohorts

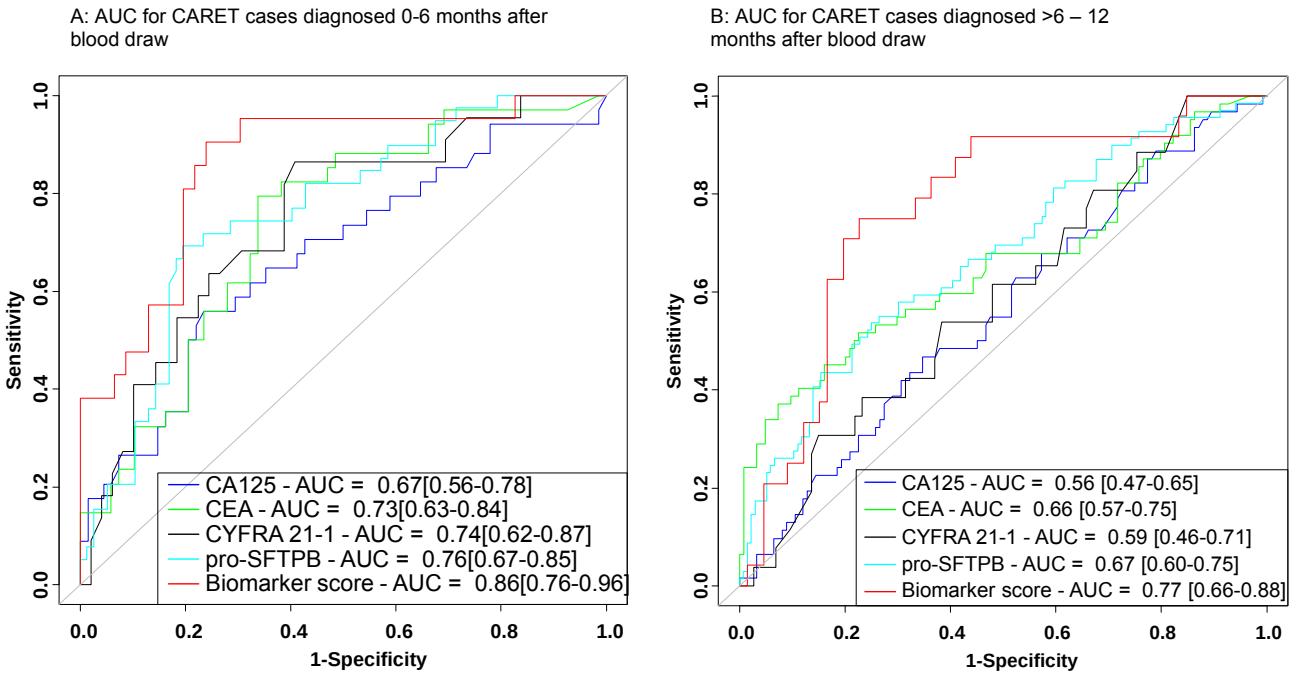


EPIC participating countries: Greece, Netherlands, UK, France, Germany, Spain, and Italy
Matching criteria for all controls: study center, sex, date of blood collection (± 1 month, relaxed to ± 12 months for sets without available controls), time at blood collection (± 1 hour, relaxed to ± 12 hours), and age at blood collection (± 3 months, relaxed to ± 5 years). If possible, one of the controls was additionally matched based on smoking status of the index case from 5 categories; never smokers, short and long term quitters among former smokers (<10 years, ≥10 years since quitting), and light and heavy smokers among current smokers (< 15 years, ≥15 cigarettes per day).

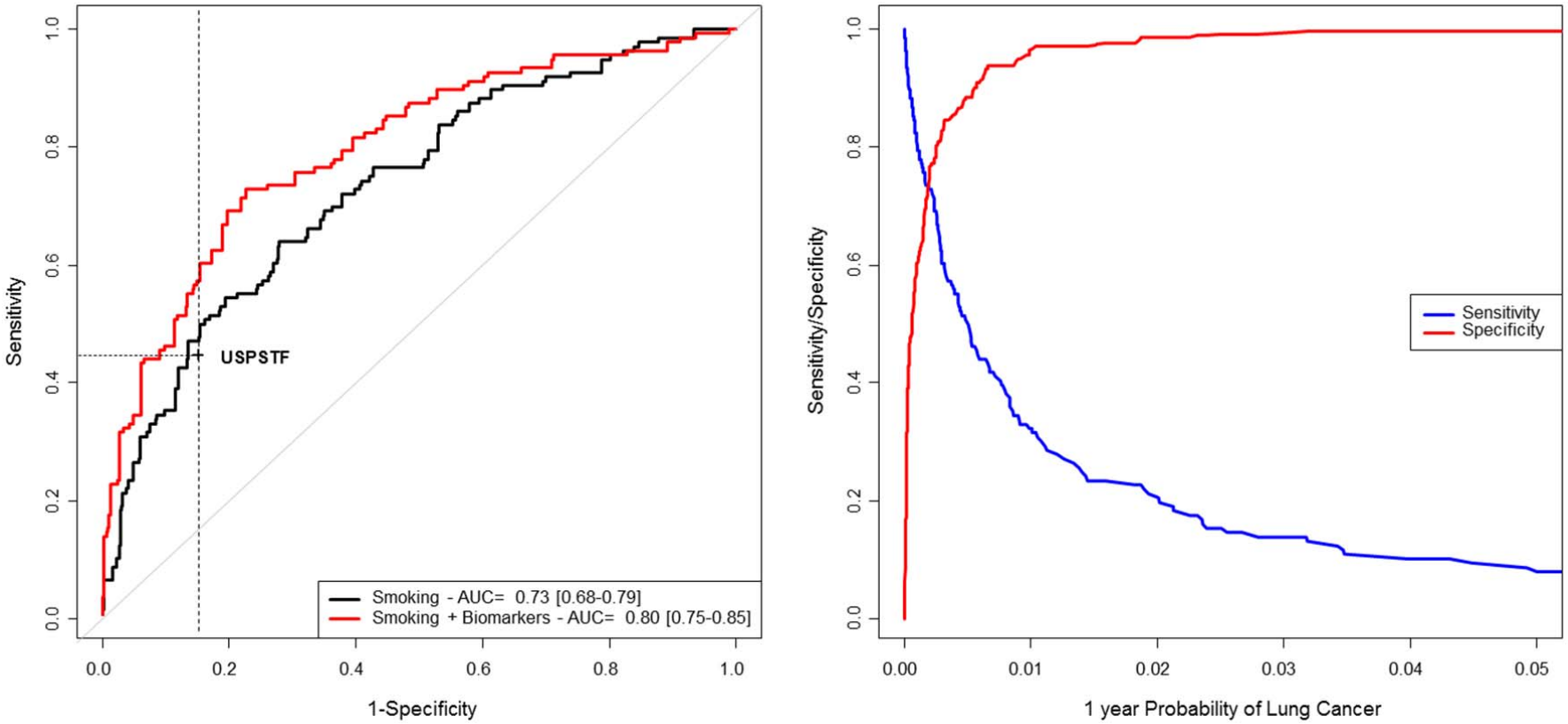
eFigure 2: Receiver Operating Characteristic (ROC) Curves for Each of the 5 Biomarkers in the CARET Training Study and for the 4-Marker Panel



eFigure 3. Receiver Operating Characteristic (ROC) Curves for All 4 Biomarkers and the Biomarker Score Stratified for Cases Diagnosed Within 6 Months (Panel A) and 6 to 12 Months (Panel B) of Blood Draw in the CARET Training Study

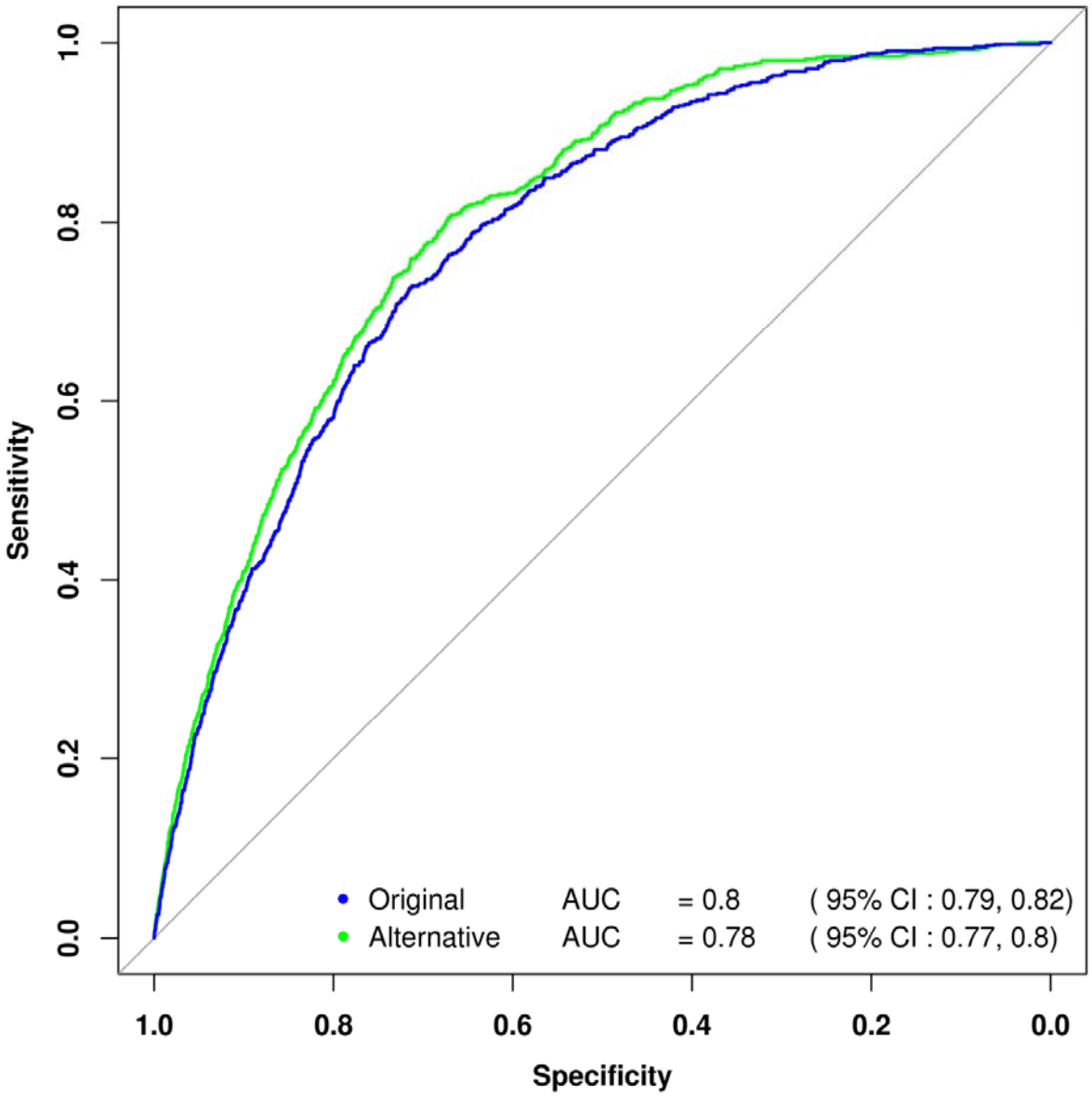


eFigure 4. Extension of the Receiver Operating Characteristic (ROC) Curve Analysis to EPIC and NSHDS Ever-Smoking Subjects With a Diagnosis Within 2 Years of Blood Collection for 2 Risk Prediction Models, Smoking Variables Only and an Integrated Model with the Smoking Variables and the Biomarker Score Combined



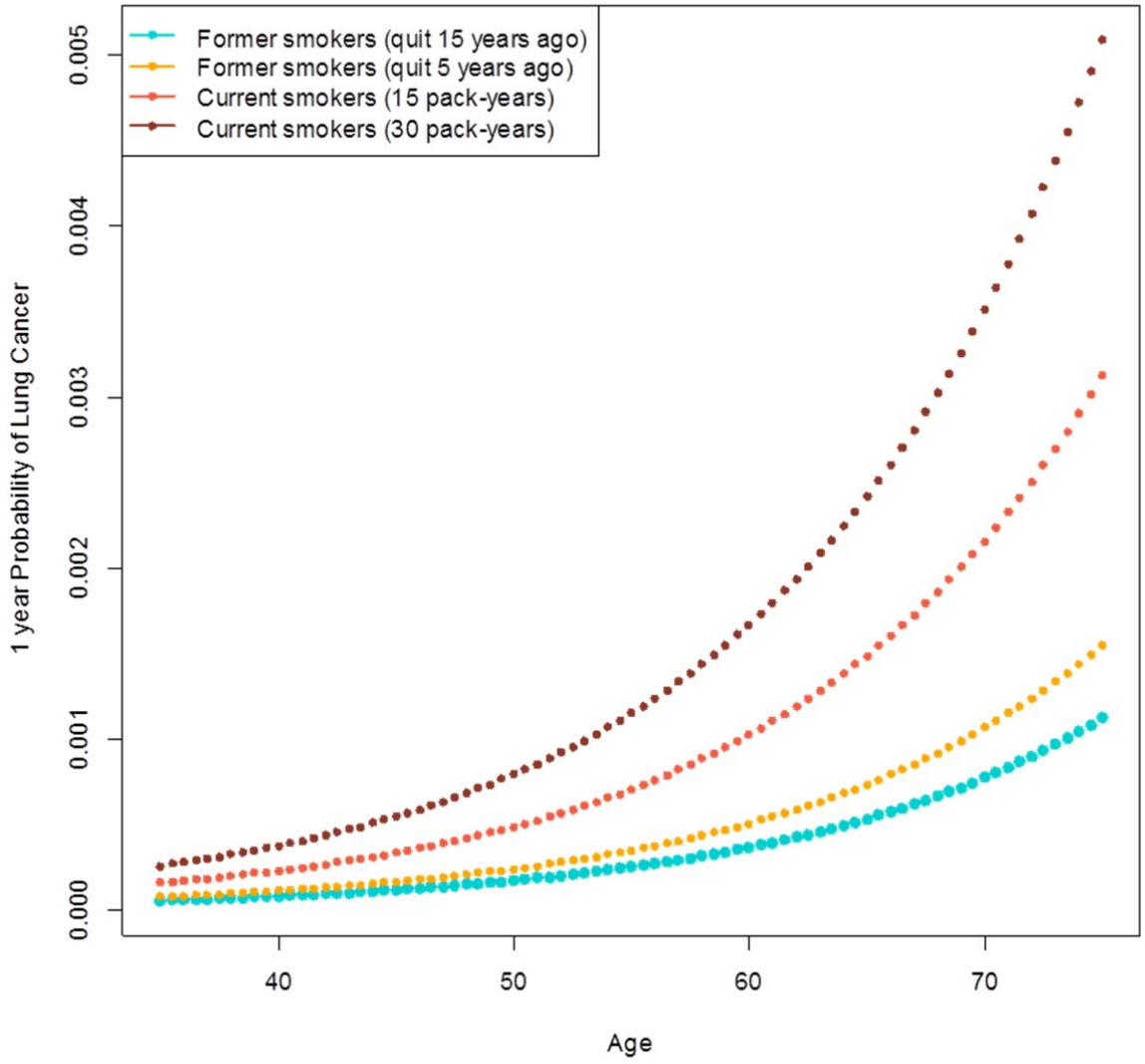
A: ROC curve analysis in EPIC and NSHDS subjects diagnosed within 2 year of blood collection for two risk prediction models, smoking variables only, and an integrated model with the smoking variables and the biomarker score combined. B: Evolution of sensitivity and specificity according to the 2 years predicted lung cancer probability from the integrated model.

eFigure 5. Discriminative Performance of the Original and Reduced PLCO $_{M2012}$ Models in the PLCO Cancer Screening Trial

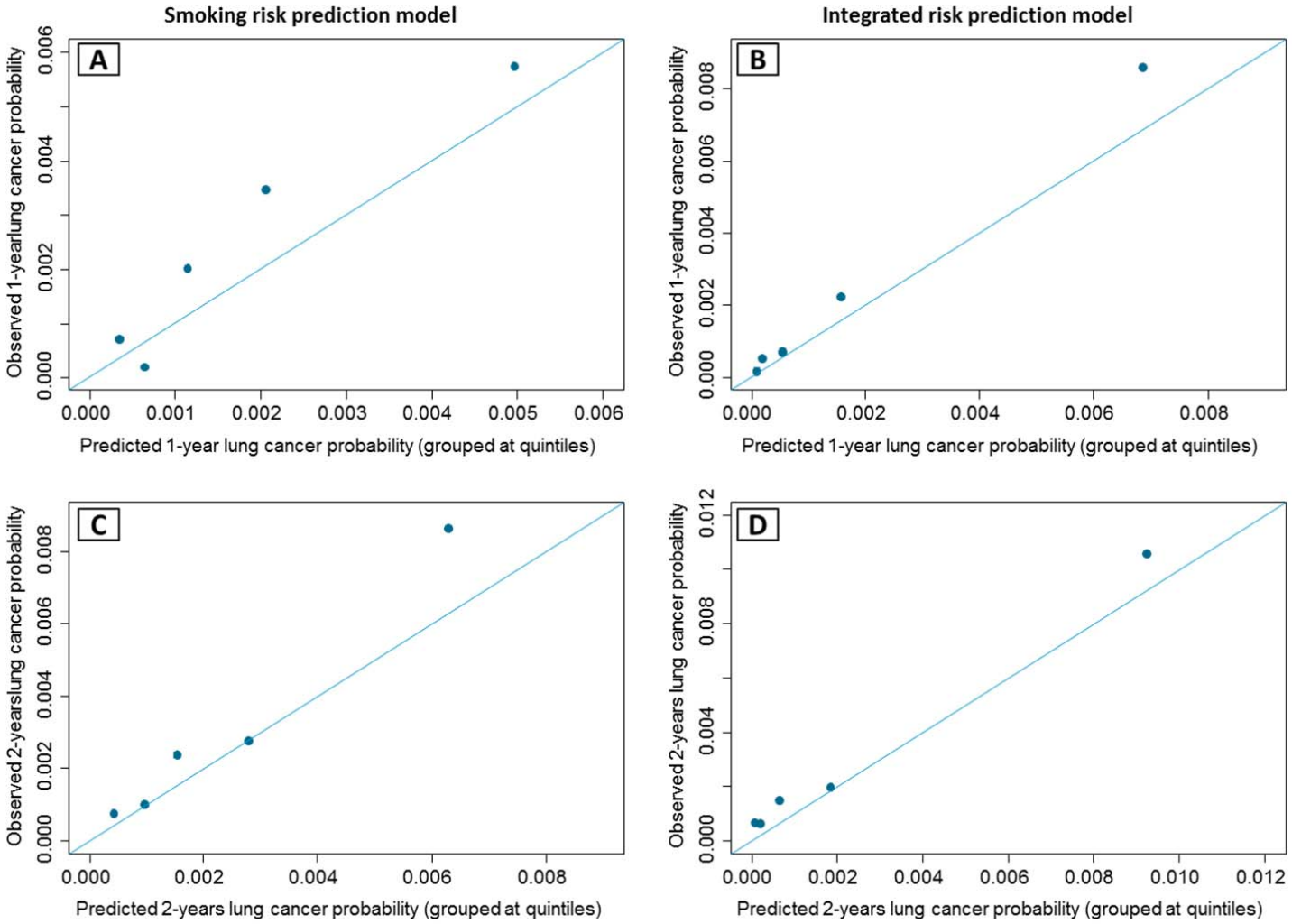


The original PLCO$_{M2012}$ model (blue line) is compared to the same model without history of chronic obstructive pulmonary disease, family history of lung cancer and intensity of smoking for former smokers (green line).

eFigure 6. Probability of Lung Cancer Within 1 Year Predicted from a Smoking-Based Risk Prediction Model in Ever Smokers from the EPIC Cohort and Presented for a Man According to His Possible Smoking History and Age

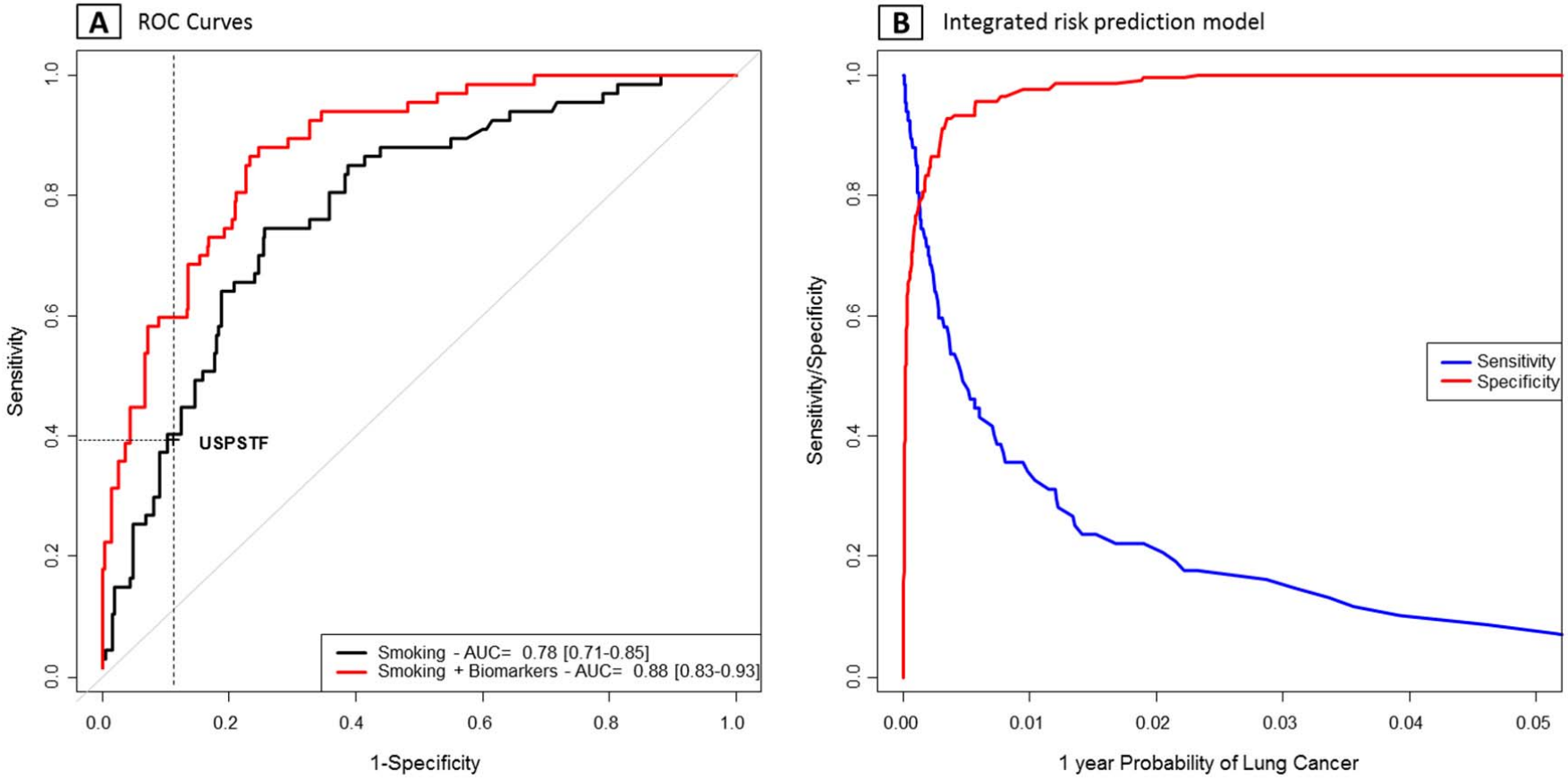


eFigure 7. Calibration of the Prediction Models in Ever Smokers from the EPIC and NSHDS Samples



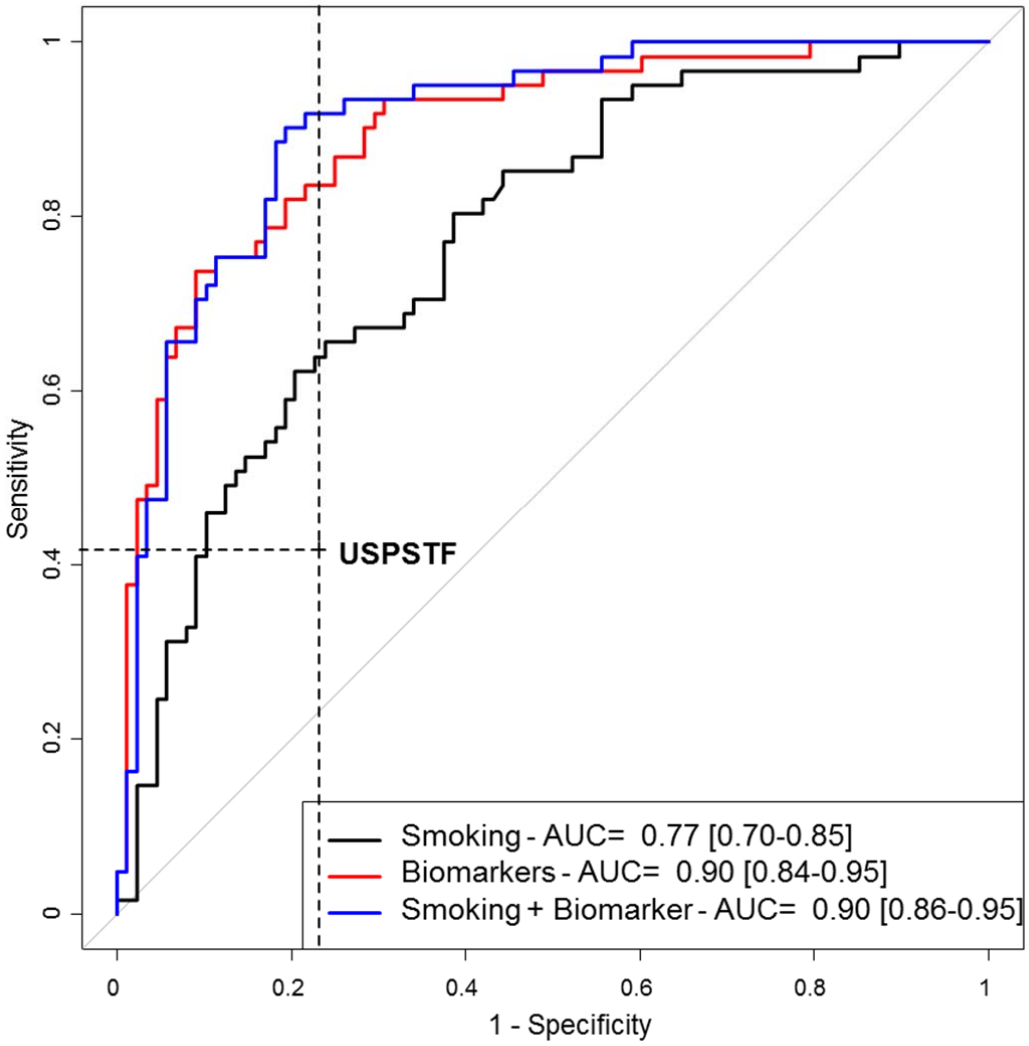
Predicted lung cancer probability are compared to the observed lung cancer probabilities for the same time period (Kaplan-Meier estimates by quintiles of model-predicted probability)
A: Calibration of the smoking risk prediction model for 1-year predicted probability; B: Calibration of the integrated risk prediction model for 1-year predicted probability;
C: Calibration of the smoking risk prediction model for 2-years predicted probability; D: Calibration of the integrated risk prediction model for 2-years predicted probability.

eFigure 8. Receiver Operating Characteristic (ROC) Curve Analysis in the Validation Study (EPIC and NSHDS Subjects With Diagnosis Within 1 Year of Blood Collection) for 2 Risk Prediction Models, Smoking Variables Only and an Integrated Model with the Smoking Variables and the Biomarker Score Combined

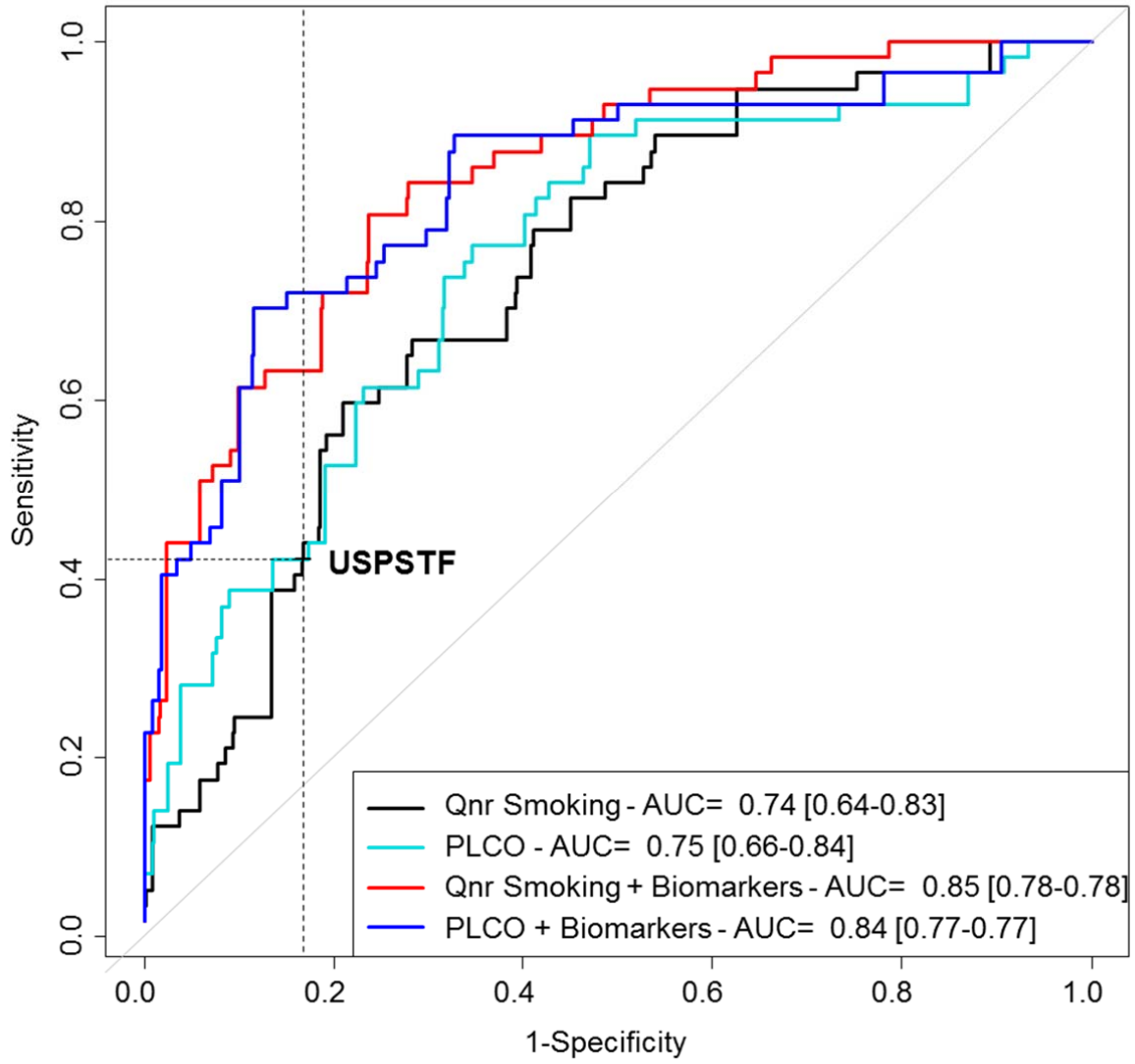


A: ROC curve analysis in ever smokers in the validation study (EPIC and NSHDS subjects diagnosed within 1 year of blood collection) for two risk prediction models, smoking variables only, and an integrated model with the smoking variables and the biomarker score combined. B: Evolution of sensitivity and specificity according to the 1 year predicted lung cancer probability from the integrated model.

eFigure 9. Apparent Receiver Operating Characteristic (ROC) Curve Analysis Among Ever Smokers from the Validation Study (EPIC and NSHDS Subjects With a Diagnosis Within 1 Year of Blood Collection) for 3 Risk Prediction Models With Smoking Variables Only, Biomarker Score Only, and an Integrated Model With the Smoking Variables and the Biomarker Score Combined



eFigure 10. Receiver Operating Characteristic (ROC) Curve Analysis in Ever Smokers in the Validation Study (EPIC and NSHDS Subjects With diagnosis Within 1 Year of Blood Collection) for the PLCO Risk Score Compared with the Risk Prediction Models



Abbreviation: Qnr: Questionnaire.