Supplementary Materials

Supplementary Methods

Imputation of missing covariate data: We first imputed missing information for demographic, smoking, and health-related covariates needed to implement a standard smoking-based lung cancer risk prediction model, PLCOm2012. We assigned the mean value within each cohort for the participants who missed BMI (0.6% of study participants missing BMI). For the partially missing variables (education, smoking duration, smoking intensity, and years since quitting for former smokers), we stratified the imputation procedure by cohort and smoking status (current or former) and applied multiple imputation by chained equations (MICE). In NSHDS (Sweden), smoking intensity information was not available for people who used to smoke, nor for some people who currently smoke. We therefore additionally combined HUNT (Norway) with NSHDS to impute the smoking intensity for participants in NSHDS. We assigned history of chronic obstructive pulmonary disease (COPD) or emphysema, first-degree family history of lung cancer, and personal history of cancer as null for all participants from cohorts that did not collect information about these variables. The detailed missingness distribution are presented in Supplementary Table 1.

R packages used for analysis: The R package "glmnet" was applied to perform the Lasso logistic regression analysis for multiprotein analyses,[1] “pROC” was applied for the AUC calculation, comparation and adjustment of cut-point for the protein-based risk model that yields the same specificity as EarlyCDT®-Lung,[2] and “DTComPair” was used for the McNemar test.[3]

References:

1. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33(1):1-22.
2. Robin X, Turck N, Hainard A*, et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.
3. Christian Stock, Hielscher T. *DTComPair: comparison of binary diagnostic tests in a paired study design*. http://CRAN.R-project.org/package=DTComPair.

Supplementary Table 1. Characteristics of the study participants before imputation.

	CPS (N=228)	**HUNT** (N=324)	**MCCS** (N=208)	**SCHS** (N=180)	**EPIC** (N=180)	**NSHDS** (N=128)	**Total** (N=1248)
Age, Median (Q1-Q3)	71 (67 - 74)	68 (60 - 73)	67 (62 - 73)	69 (65 - 74)	57 (53 - 63)	60 (57 - 60)	66 (60 - 72)
Female participants (%)	84 (36.8%)	122 (37.7%)	64 (30.8%)	20 (11.1%)	60 (33.3%)	64 (50.0%)	414 (33.2%)
Number of cigarettes smoked per day							
Median (Q1-Q3)	20 (9.5 - 30)	10 (8.0 - 15)	20 (10 - 30)	17 (10 - 20)	16 (11 - 21)	15 (15 - 20)	16 (10 - 23)
Missing, N	3	43	2	0	9	119	176
Years smoked							
Median (Q1-Q3)	34 (20 - 45)	44 (36 - 50)	40 (29 - 48)	44 (37 - 53)	36 (29 - 42)	38 (28 - 43)	40 (30 - 47)
Missing, N	0	24	0	1	6	17	48
Quit years*							
Median (Q1-Q3)	23 (9.5 - 33)	13 (5.6 - 20)	15 (6.3 - 27)	8.0 (3.0 - 18)	12 (5.4 - 22)	9.6 (4.0 - 19)	15 (6.0 - 27)
Missing, N	0	17	0	0	2	5	24
Education							
Less than grade 12	19 (8.41%)	153 (53.5%)	133 (63.9%)	146 (81.1%)	94 (55.6%)	86 (68.3%)	631 (52.8%)
high-school graduate	54 (23.9%)	99 (34.6%)	47 (22.6%)	28 (15.6%)	50 (29.6%)	21 (16.7%)	299 (25.0%)
Above high school	153 (67.7%)	34 (11.9%)	28 (13.5%)	6 (3.33%)	25 (14.8%)	19 (15.1%)	265 (22.2%)
Missing, N	2	38	0	0	11	2	53
BMI							
Median (Q1-Q3)	26 (23 - 28)	26 (23 - 28)	27 (24 - 30)	23 (20 - 25)	26 (23 - 29)	26 (23 - 28)	25 (23 - 28)
Missing, N	2	4	0	0	0	2	8
History of cancer	3 (1.32%)	15 (4.81%)	30 (14.4%)	0 (0%)	0 (0%)	0 (NA)	48 (4.33%)
Missing, N	0	12	0	0	0	128	140
Family history of cancer	18 (7.89%)	0 (NA%)	22 (20.8%)	0 (NA)	0 (NA)	0 (NA)	40 (12.0%)
Missing, N	0	324	102	180	180	128	914
COPD	20 (8.77%)	23 (14.0%)	0 (NA)	0 (NA)	0 (NA)	0 (NA)	43 (11.0%)
Missing, N	0	160	208	180	180	128	856

*Only former smokers

Supplementary Table 2. 500 bootstrap internal validation and external validation of PLCOm2012 model and Protein marker model in 6 datasets. Using the protein panels selected by themselves.

	Internal validation: Mean AUC (SD)			External validation: AUC (95% CI)	
Training dataset	**PLCOm2012 model**	**Protein-based risk model**	**Testing dataset**	**PLCOm2012 model**	**Protein-based risk model**
EPIC, HUNT, MCCS, NSHDS, SCHS	0.59 (0.03)	0.71 (0.03)	CPS	0.63 (0.55 to 0.70)	0.77 (0.71 to 0.83)
CPS, HUNT, MCCS, NSHDS, SCHS	0.59 (0.03)	0.73 (0.03)	EPIC	0.70 (0.62 to 0.77)	0.71 (0.63 to 0.78)
CPS, EPIC, MCCS, NSHDS, SCHS	0.62 (0.03)	0.74 (0.03)	HUNT	0.55 (0.49 to 0.62)	0.70 (0.64 to 0.76)
CPS, EPIC, HUNT, NSHDS, SCHS	0.59 (0.03)	0.74 (0.02)	MCCS	0.68 (0.60 to 0.75)	0.68 (0.61 to 0.75)
CPS, EPIC, HUNT, MCCS, NSHDS	0.62 (0.03)	0.73 (0.03)	SCHS	0.53 (0.45 to 0.62)	0.71 (0.63 to 0.78)
CPS, EPIC, HUNT, MCCS, SCHS	0.61 (0.03)	0.71 (0.03)	NSHDS	0.56 (0.46 to 0.66)	0.81 (0.74 to 0.89)

Supplementary Table 3. 500 bootstrap internal validation and external validation of PLCOm2012 model and Protein marker model in 6 datasets. Using the 4 proteins selected in the main manuscript.

	Internal validation: Mean AUC (SD)			External validation: AUC (95% CI)	
Training dataset	**PLCOm2012 model**	**Protein-based risk model**	**Testing dataset**	**PLCOm2012 model**	**Protein-based risk model**
EPIC, HUNT, MCCS, NSHDS, SCHS	0.59 (0.03)	0.72 (0.03)	CPS	0.63 (0.55 to 0.70)	0.77 (0.71 to 0.83)
CPS, HUNT, MCCS, NSHDS, SCHS	0.59 (0.03)	0.73 (0.03)	EPIC	0.70 (0.62 to 0.77)	0.72 (0.64 to 0.79)
CPS, EPIC, MCCS, NSHDS, SCHS	0.62 (0.03)	0.73 (0.03)	HUNT	0.55 (0.49 to 0.62)	0.70 (0.64 to 0.75)
CPS, EPIC, HUNT, NSHDS, SCHS	0.59 (0.03)	0.72 (0.03)	MCCS	0.68 (0.60 to 0.75)	0.72(0.65 to 0.79)
CPS, EPIC, HUNT, MCCS, NSHDS	0.62 (0.03)	0.72 (0.03)	SCHS	0.53 (0.45 to 0.62)	0.72 (0.65 to 0.80)
CPS, EPIC, HUNT, MCCS, SCHS	0.61 (0.03)	0.71 (0.03)	NSHDS	0.56 (0.46 to 0.66)	0.81 (0.74 to 0.89)

Supplementary Table 4. AUC of PLCOm2012 model and protein-based risk model in the validation set by stratifying the histology.

Models	Number Of cases	Number Of controls	PLCOm2012 model AUC (95% CI)	Protein-based risk model AUC (95% CI)
Histology subtype and all controls				
Adenocarcinoma	50	154	0.60 (0.51 to 0.69)	0.77 (0.69 to 0.85)
Small Cell Carcinoma	24	154	0.63 (0.52 to 0.74)	0.78 (0.69 to 0.88)
Squamous Cell Carcinoma	26	154	0.64 (0.52 to 0.76)	0.77 (0.66 to 0.88)
Other/NOS	54	154	0.67 (0.58 to 0.75)	0.72 (0.64 to 0.80)

Supplementary Table 5. Comparison of Diagnostic Performance of Protein Algorithm and EarlyCDT®-Lung in the validation set by stratifying the histology.

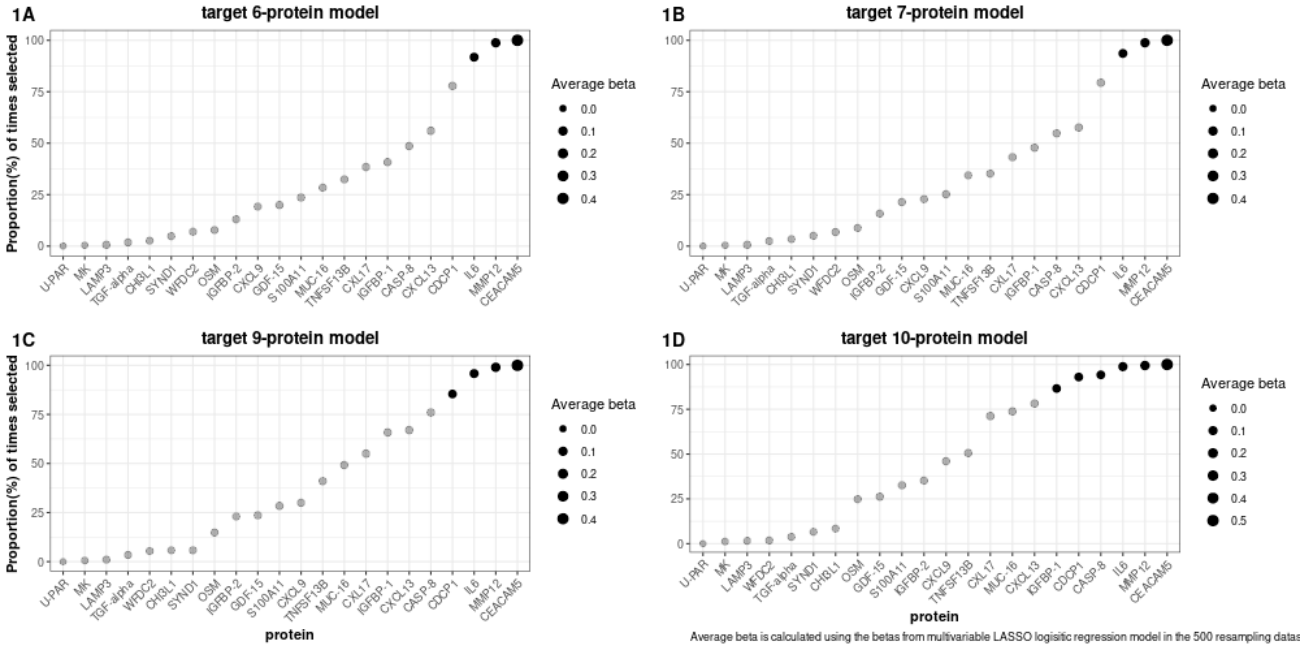
Models	Number Of cases	Number Of controls	EarlyCDT-Lung test		PLCOm2012 model		Protein-based risk model
			Sensitivity (95% CI)	P value[#]	Sensitivity* (95% CI)	P value[†]	Sensitivity* (95% CI)
Histology subtype and all controls							
Adenocarcinoma	50	154	6% (0 to 13)	0.000012	28% (16 to 40)	0.03	48% (34 to 62)
Small Cell Carcinoma	24	154	21% (4.6 to 37)	0.03	12% (0 to 26)	0.02	50% (30 to 70)
Squamous Cell Carcinoma	26	154	15% (1.5 to 29)	0.002	38% (20 to 57)	0.12	58% (39 to 77)
Other/NOS	54	154	17% (6.7 to 27)		35% (22 to 48)	0.24	46% (33 to 60)

*Sensitivities for the PLCOm2012model and Protein-based risk model were estimated by adjusting the cut-off of each respective risk model that yielded the same specificity as the EarlyCDT®-Lung test which was estimated at 86% in the overall smoking-matched control population and varied between 84% and 90% depending on the strata.

[#]p value for the sensitivity difference between Protein-based risk model and EarlyCDT®-Lung at the same specificity level.

[†]p value for the sensitivity difference between Protein-based risk model and PLCOm2012 model at the same specificity level.

Supplementary Figure 1. Proportion of proteins selected by the LASSO logistic regression model during 500 resampling for a set number of targeted proteins. Proteins selected more than 400 times are marked as black. 1A: target 6-protein model; 1B: target 7-protein model; 1C: target 9-protein model; 1D: target 10-protein model.



Supplementary Figure 2. Proportion of proteins selected by the LASSO logistic regression model during 500 resampling in 6 development datasets. Proteins selected more than 400 times are marked as black. 2A: EPIC, HUNT, MCCS, NSHDS, SCHS; 2B: CPS, HUNT, MCCS, NSHDS, SCHS; 2C: CPS, EPIC, MCCS, NSHDS, SCHS; 2D: CPS, EPIC, HUNT, NSHDS, SCHS; 2E: CPS, EPIC, HUNT, MCCS, NSHDS; 2F: CPS, EPIC, HUNT, MCCS, SCHS.