DOI: 10.1093/jnci/djr173
Advance Access publication on May 23, 2011.

ARTICLE

Lung Cancer Risk Prediction: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Models and Validation

C. Martin Tammemagi, Paul F. Pinsky, Neil E. Caporaso, Paul A. Kvale, William G. Hocking, Timothy R. Church, Thomas L. Riley, John Commins, Martin M. Oken, Christine D. Berg, Philip C. Prorok

Manuscript received August 30, 2010; revised March 29, 2011; accepted April 19, 2011.

Correspondence to: C. Martin Tammemagi, PhD, Department of Community Health Sciences, Brock University, 500 Glenridge Ave, St Catharines, ON, Canada L2S 3A1 (e-mail: martin.tammemagi@brocku.ca).

Introduction Identification of individuals at high risk for lung cancer should be of value to individuals, patients, clinicians, and researchers. Existing prediction models have only modest capabilities to classify persons at risk accurately.

Methods Prospective data from 70962 control subjects in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) were used in models for the general population (model 1) and for a subcohort of ever-smokers (N = 38254) (model 2). Both models included age, socioeconomic status (education), body mass index, family history of lung cancer, chronic obstructive pulmonary disease, recent chest x-ray, smoking status (never, former, or current), pack-years smoked, and smoking duration. Model 2 also included smoking quit-time (time in years since ever-smokers permanently quit smoking). External validation was performed with 44223 PLCO intervention arm participants who completed a supplemental questionnaire and were subsequently followed. Known available risk factors were included in logistic regression models. Bootstrap optimism-corrected estimates of predictive performance were calculated (internal validation). Nonlinear relationships for age, pack-years smoked, smoking duration, and quit-time were modeled using restricted cubic splines. All reported *P* values are two-sided.

Results During follow-up (median 9.2 years) of the control arm subjects, 1040 lung cancers occurred. During follow-up of the external validation sample (median 3.0 years), 213 lung cancers occurred. For models 1 and 2, bootstrap optimism-corrected receiver operator characteristic area under the curves were 0.857 and 0.805, and calibration slopes (model-predicted probabilities vs observed probabilities) were 0.987 and 0.979, respectively. In the external validation sample, models 1 and 2 had area under the curves of 0.841 and 0.784, respectively. These models had high discrimination in women, men, whites, and nonwhites.

Conclusion The PLCO lung cancer risk models demonstrate high discrimination and calibration.

J Natl Cancer Inst 2011;103:1058–1068

Lung cancer is the leading cause of cancer death in North America and worldwide (1–3). Accurate lung cancer prediction might help reduce lung cancer mortality by motivating current smokers to quit and by identifying current smokers at high risk who might benefit from intensive smoking cessation programs. Clinicians might consider increased monitoring of patient behaviors and health based on lung cancer risk. Lung cancer chemoprevention trials can be made more efficient by increasing enrollment of high-risk individuals, and if effective chemoprevention is discovered, it should be more cost-effective if applied to high-risk individuals. Recently, the National Cancer Institute announced that the National Lung Screening Trial (4) found that low-dose computed tomography screening statistically significantly reduced lung cancer mortality by 20% in high-risk individuals (5). Cost-effective adoption of computed tomography lung cancer screening programs will require identification and application of such programs to high-risk individuals. Thus, accurate lung cancer risk prediction models would benefit patients, clinicians, researchers, and public health administrators.

Several lung cancer risk prediction models have been proposed (6–13). However, some limitations of these models are a restricted number of potential predictors, generally low overall predictive performance, and methodological limitations. Predictor variables in existing models include age; cigarette smoking history; second-hand smoke exposure in never-smokers (10); history of bronchitis, emphysema, or pneumonia (9); asbestos exposure (8); and family history of lung cancer (11). Several factors associated with lung cancer in epidemiological studies might also be useful for prediction, including socioeconomic status (14), body mass index (BMI; weight in kilograms per height in meters squared), and recent history of chest radiograph (15). Lower socioeconomic status has been associated with increased lung cancer risk in many countries, including the United States, Canada, the Netherlands, and China

(14). High BMI has been inversely associated with lung cancer in several studies (16–25). Chronic obstructive pulmonary disease (COPD), respiratory illnesses, and chronic inflammation play a role in carcinogenesis (26–28), including lung carcinogenesis (29–32), and may explain the observed association between recent chest x-ray and lung cancer (15).

The aim of this study was to produce improved lung cancer risk prediction models by incorporating a wider range of lung cancer risk factors than in past models, using a prospective study design, and evaluating nonlinear effects using improved statistical methodologies. The models are based on data collected in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO), a randomized clinical trial, which is studying whether selected screening modalities reduce respective cancer mortality rates. Models were prepared for the entire PLCO control arm participants and for control subjects from the PLCO who were ever-smokers (former or current smokers). The incremental value of adding detailed smoking, nonsmoking variables, and nonlinear terms to simpler prediction models was evaluated by comparing receiver operator characteristic areas under the curve (ROC AUC). External validation was carried out using a subset of individuals in the PLCO intervention arm.

CONTEXTS AND CAVEATS

Prior knowledge

Current lung cancer risk prediction models are hampered by a restricted number of potential predictors, generally low overall predictive performance, and methodological limitations.

Study design

Data on socioeconomic variables, medical history, and smoking history from 70 962 control subjects in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) were used to test predictive models for the general population and a subcohort of ever-smokers. The models were externally validated in follow-up of 44 223 PLCO intervention arm participants.

Contribution

Both models demonstrated high predictive accuracy when applied to the external validation sample. Predictive accuracy remained high for subsamples of women, men, whites, and nonwhites separately. Sex and race/ethnicity did not appear to be important predictors of lung cancer risk.

Implications

Accurate lung cancer prediction might help reduce lung cancer mortality by motivating current smokers to quit and by identifying high-risk individuals who might benefit from intensive smoking cessation programs and chemoprevention.

Limitations

The models may not be generalizable to other populations because the PLCO study participants were on average of higher socioeconomic status than the general population, and the external validation sample came from the same referent population as the model development sample. Data on exposure to radon, asbestos, secondhand smoke, occupational carcinogens, and history of adult pneumonia were not available for analysis.

From the Editors

Methods

Study Design

Data for this study came from the PLCO trial, which has been reported previously (33–35). The PLCO is a 10-site randomized screening trial, which began enrolling participants in November 1993 and completed recruitment of 154 938 subjects in June 2001. Subjects at study entry were men and women aged 55–74 years who were thought to be free of the cancers under study. The screening intervention for lung cancer consisted of four annual posterior–anterior chest radiographs. Control subjects received regular care as recommended by their physicians. A baseline epidemiological questionnaire was administered at enrollment and a supplemental questionnaire (SQ) was administered during 2004–2005. The PLCO received institutional review board approval from all sites and, informed consent was obtained from each participant.

To avoid possible effects of overdiagnosis bias (36,37), models were developed using PLCO control subject data, which included data from 77 461 individuals. Individuals who did not complete the baseline questionnaire (n = 3091), who had cancer before study entry (3405), or who dropped out of the study just after being assigned a study identification number (n = 3) were excluded from the current study, leaving 70 962 individuals eligible for study.

Statistical Analysis

Logistic regression models were prepared for prediction of lung cancer in the entire PLCO control arm (N = 70 962) and in the subcohort of ever-smokers (N = 38 254). Potential predictor variables, including sociodemographic, medical history, and exposure risk factors, were selected by review of the scientific literature. Prespecified available model predictor variables included age, socioeconomic status (estimated by education), race/ethnicity (self-reported), sex, family history of lung cancer, BMI (in kilogram per meter squared), history of COPD, history of chest x-ray in the 3 years before baseline, smoking history (smoking status [never, former, or current]; smoking intensity [cigarettes per day], duration [years], quit-time [time in years since former smokers permanently quit smoking] and pack-years smoked [average number of packages smoked multiplied by the number of years smoked]) (14,15,18). Regarding sex and race/ethnicity, some data indicate that black men are at higher risk of lung cancer compared with other sex–race groups (38–40), and we evaluated this association in our prediction models. Data were not available for some lung cancer risk factors, such as asbestos, radon, or secondhand smoke exposures. Backward model reduction removed only variables having very small or implausible effects ($P > .20$). To minimize overfitting of models, no exploratory search beyond the a priori predictors was carried out, and no additional subset exploratory analyses were conducted.

Nonlinear effects of continuous variables were evaluated using restricted cubic splines (41). The number and location of knots used to fix splines in modeling followed the recommendations of Harrell (41) and Steyerberg (42). Interactions of variables in final models were evaluated by the likelihood ratio test. To enhance the simplicity and utility of the models, it was decided a priori to include interaction terms only if the global significance level for interactions was statistically significant if P was less than or equal to .05.

Overall model performance was evaluated with Nagelkerke's pseudo-R^2 statistics (41,42). The models' ability to discriminate (classify correctly) was assessed using the ROC AUC or the equivalent concordance statistic (*c* statistic) (42,43). Model calibration (correspondence between model-predicted probabilities and observed probabilities) was assessed by the Hosmer–Lemeshow goodness-of-fit test and evaluating how much the slope of the calibration line (plotting the predicted probabilities *vs* the observed probabilities) deviates from the ideal of 1.0. The mean absolute error and 90th percentile absolute error in calibration were statistics used to appraise calibration, with error referring to the difference between the observed values and the bias-corrected calibrated values. Gail and Pfeiffer (44) found that in screening applications, discrimination is more important than calibration and, in interpretation of model performance, priority was given to discrimination. Whether one or more predictors substantially improve prediction were tested by evaluating whether the ROC AUC difference equaled zero in nested models.

Internal validation of models was carried out by correcting measures of predictive performance for "optimism" or overfit by using bootstrap methods in the *rms* packages (version 3.1-0) in *R* using 200 resamplings (41). An external validation cohort was established consisting of PLCO intervention arm subjects who were thought to be lung cancer free when they completed the SQ. Follow-up for this cohort began when individuals completed the SQ. Models developed in the control arm subjects were tested to see how well they discriminated in the post-SQ intervention arm cohort. Because the follow-up times differed substantially between development and validation cohorts and thus the cumulative incidence (lung cancer risk) differed between the two groups, calibration was not assessed in the latter. Data for all variables tested in the external validation cohort came from the SQ except for education, COPD, and chest x-ray in the past 3 years, which came from the baseline questionnaire.

Cox proportional hazard models were prepared for comparisons with logistic regression models. Proportionality was verified graphically by plotting $-\ln\{-\ln(\text{survival})\}$ curves *vs* ln(analysis time) and inspecting whether lines were parallel. The hazard ratios and odds ratios from corresponding models were compared. Models, statistics, and figures were prepared using Stata/MP 11.1 (StataCorp, College Station, TX) and *R* (version 2.11) (45) statistical programs. All reported *P* values are two-sided.

Results

The PLCO control arm participants had a mean age of 62.6 years (Table 1). Lung cancer occurred more frequently in older individuals; men; African Americans; and in individuals with lower education, a family history of lung cancer, COPD, or lower BMI; and in those who had smoked more cigarettes per day or for longer durations (Table 1). Lung cancer risk increased from never to former to current smokers (former *vs* never: odds ratio = 7.34, 95% confidence interval [CI] = 5.77 to 9.35; current *vs* never: odds ratio = 26.93, 95% CI = 21.09 to 34.39; both $P <$ .001). The median follow-up was 9.24 years (interquartile range 7.51–10.69), and during follow-up, 1040 lung cancers occurred. The incidence rates of lung cancer per 10 000 person-years in never, former, and current smokers were 2.5, 18.7, and 71.4, respectively.

Of the intervention arm participants, 44 223 completed the SQ. For 16 individuals, information on lung cancer status was absent and they were excluded from external validation, leaving a sample size of 44 207 (Table 2). The median time from last screening chest x-ray to the SQ was 6.5 years (interquartile range 4.97–7.90). The post-SQ median follow-up was 3.04 years (interquartile range 2.53–3.40). Only confirmed incident lung cancers (N = 213) occurring following the SQ were included in the external validation analysis. The incidence rates of lung cancer per 10 000 person-years in never, former, and current smokers were 3.3, 19.3, and 79.9, respectively.

Predictive Model in All PLCO Control Subject Population

In the model 1 sample (PLCO control arm subjects), lung cancer risk increased with age, lower education, lower BMI, family history of lung cancer, presence of COPD, occurrence of chest x-ray in the 3 years before study entry, being a current smoker, pack-years smoked, and smoking duration (Table 3). For the continuous variables, age, pack-years smoked, and smoking duration, the relationships with lung cancer were statistically significantly nonlinear (Wald $P <$ .001 for all three variables). For BMI, the nonlinear component was not statistically significant.

For model 1 performance, the ROC AUC was 0.859 (95% CI = 0.848 to 0.871; bootstrap optimism-corrected $c = 0.857$) and the optimism-corrected calibration slope was 0.987. The mean and 90th percentile absolute errors were 0.0009 and 0.0025, respectively (Figure 1,A and Table 3). These statistics indicate excellent model discrimination and calibration.

In the external validation sample, model 1 demonstrated excellent discrimination overall (ROC AUC = 0.841, 95% CI = 0.813 to 0.870) (Figure 1,B). Predictive accuracy of the model remained high when the analysis was limited to women, men, whites, or nonwhites (including Hispanics), with ROC AUCs ranging from 0.828 to 0.849 (Table 3). Model 1 performed well in the PLCO control sample ever-smokers (ROC AUC = 0.807, 95% CI = 0.794 to 0.820) but less so in control subject never-smokers (ROC AUC = 0.662, 95% CI = 0.598 to 0.725). When model 1 was applied to the external validation sample smokers and never-smokers separately, the respective ROC AUCs were 0.784 (95% CI = 0.753 to 0.815) and 0.597 (95% CI = 0.455 to 0.740).

Predictive Model in PLCO Ever-Smokers

Lung cancer is uncommon in never-smokers, and it is expected that lung cancer prediction models will most often be applied to smokers. A model intended for use in smokers that is developed in smokers may outperform a model developed in a general population. Model 2 was designed as a lung cancer risk prediction model based on PLCO control arm ever-smokers only (Table 3). The unadjusted probability of developing lung cancer increased with age, pack-years, and smoking duration, but decreased with quit-time (Figure 2,A–D). In model 2, significant nonlinear components were observed for age, pack-years, and quit-time (Wald $P = .001$, $<.001$, and $= .01$, respectively). Smoking duration and quit-time were highly negatively correlated ($r = -.87$). In the presence of quit-time, duration no longer demonstrated a significant nonlinear

Table 1. Distribution of study variables by lung cancer status and overall in the model development sample*

Variable	No lung cancer† (n = 69922)	Lung cancer† (n = 1040)	*P*	Total‡ (N = 70962)
Sociodemographic				
Age, mean (SD), y	62.57 (5.35)	64.80 (5.09)	<.001§	62.60 (5.36)
Sex, No. (%)				
Women	34994 (98.85)	408 (1.15)		35402 (49.89)
Men	34928 (98.22)	632 (1.78)	<.001‖	35560 (50.11)
Race/ethnicity, No. (%)				
White, non-Hispanic	61687 (98.52)	924 (1.48)	.017‖	62611 (88.27)
Black, non-Hispanic	3606 (98.07)	71 (1.93)		3677 (5.18)
Hispanic	1336 (98.89)	15 (1.11)		1351 (1.90)
Asian	2675 (99.11)	24 (0.89)		2699 (3.81)
Pacific Islander	413 (99.04)	4 (0.96)		417 (0.59)
American Indian	172 (98.85)	2 (1.15)		174 (0.25)
Race–sex, No. (%)				
Black men	1590 (97.73)	37 (2.27)	.009‖	1627 (2.29)
Others	68332 (97.73)	1003 (1.45)		69335 (97.71)
Education, No. (%)				
High school and below	21132 (98.03)	424 (1.97)	<.001‖	21556 (30.49)
Greater than high school	48519 (98.75)	613 (1.25)		49132 (69.51)
Medical history				
Family history of lung cancer, No. (%)				
Absent	60398 (98.67)	815 (1.33)	<.001‖	61213 (87.00)
Present	8935 (97.67)	213 (2.33)		9148 (13.00)
COPD, No. (%)				
History absent	60952 (98.75)	773 (1.25)	<.001‖	61725 (92.81)
History present	4589 (95.92)	195 (4.08)		4784 (7.19)
Chest x-rays in past 3 y, No. (%)				
No	31113 (98.97)	323 (1.03)	<.001¶	31436 (46.10)
Yes, on one occasion	22899 (98.41)	370 (1.59)		23269 (34.12)
Yes, more than one occasion	13186 (97.74)	305 (2.26)		13491 (19.78)
Body mass index, mean (SD), kg/m^2	27.34 (5.01)	26.39 (4.51)	<.001§	27.32 (5.01)
Smoking history				
Smoking status, No. (%)				
Never-smoker	32612 (99.77)	76 (0.23)	<.001¶	32688 (46.08)
Former smoker	30152 (98.32)	516 (1.68)		30668 (43.23)
Current smoker	7138 (94.09)	448 (5.91)		7586 (10.69)
Smoking duration, mean (SD), y				
Former smokers	23.84 (12.48)	35.02 (11.18)	<.001§	24.03 (12.54)
Current smokers	42.02 (7.46)	45.68 (6.72)	<.001§	42.23 (7.47)
Quit-time in former smokers,# mean (SD), y	20.35 (12.02)	12.86 (10.65)	<.001§	20.23 (12.04)
Pack-years smoked,** mean (SD), y				
Former smokers	25.37 (24.91)	47.69 (30.90)	<.001§	25.75 (25.19)
Current smokers	40.10 (25.68)	52.40 (26.80)	<.001§	40.81 (25.90)

* Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial control arm. COPD = chronic obstructive pulmonary disease.

† Percentage values in parentheses are row percents for variable levels by lung cancer status. The row percents of lung cancer represent the cumulative incidence of lung cancer for that exposure (variable) level.

‡ Percentage values in parentheses are column percents for variable totals.

§ *P* value by two-sided Student *t* test.

‖ *P* value by two-sided Fisher exact test.

¶ *P* value by two-sided Wilcoxon rank-sum test.

Time in years since ever-smokers permanently quit smoking.

** Pack-years is the average number of packages of cigarettes smoked per day times the number of years smoked.

effect (Wald *P* = .44), and when modeled together the effect of quit-time dominated duration.

When sex, race/ethnicity, or black men *vs* other sex–race groups pooled were added to model 2, they did not contribute substantially to the model. The respective likelihood ratio test *P* values were .79, .20, and .26. Furthermore, adding black men *vs* other sex–race groups pooled did not contribute to predictive discrimination. The ROC AUCs for model 2 with and without the variable for black men were 0.809 (95% CI = 0.796 to 0.822) and 0.809 (95% CI = 0.796 to 0.822) (*P* = .46 for difference in AUCs), respectively.

For model 2 performance, the ROC AUC was 0.809 (95% CI = 0.796 to 0.822; bootstrap optimism-corrected *c* = 0.805), and the optimism corrected calibration slope was 0.979. The mean and 90th percentile absolute errors were 0.0014 and 0.0029, respectively (Table 3 and Figure 1,C). These statistics indicate excellent discrimination and calibration for model 2.

Table 2. Distribution of study variables by lung cancer status and overall for the external validation sample*

Variable	No lung cancer† (n = 43 994)	Lung cancer† (n = 213)	*P*	Total‡ (N = 44 207)
Sociodemographic				
Age, mean (SD), y	71.01 (5.89)	71.81 (5.67)	.047§	71.01 (5.89)
Sex, No. (%)				
Women	23 012 (99.60)	93 (0.40)	.013‖	23 105 (52.27)
Men	20 982 (99.43)	120 (0.57)		21 102 (47.73)
Race/ethnicity, No. (%)				
White, non-Hispanic	39 973 (99.52)	193 (0.48)	.787‖	40 166 (90.89)
Black, non-Hispanic	1438 (99.58)	6 (0.42)		1444 (3.27)
Hispanic	711 (99.44)	4 (0.56)		715 (1.62)
Asian	1561 (99.49)	8 (0.51)		1569 (3.55)
Pacific Islander	198 (99.00)	2 (1.00)		200 (0.45)
American Indian	99 (100.00)	0 (0.00)		99 (0.22)
Education, No. (%)				
High school and below	12 318 (99.40)	74 (0.60)	.032‖	12 392 (28.07)
Greater than high school	31 621 (99.56)	139 (0.44)		31 760 (71.93)
Medical history				
Family history of lung cancer, No. (%)				
Absent	38 151 (99.57)	163 (0.43)	<.001‖	38 314 (87.27)
Present	5541 (99.14)	48 (0.86)		5589 (12.73)
COPD, No. (%)				
History absent	41 591 (99.55)	189 (0.45)	.002‖	41 780 (94.86)
History present	2242 (99.03)	22 (0.97)		2264 (5.14)
Chest x-rays in past 3 y, No. (%)				
No	20 470 (99.55)	92 (0.45)	.088¶	20 562 (48.70)
Yes, on one occasion	13 960 (99.49)	71 (0.51)		14 031 (33.23)
Yes, more than one occasion	7583 (99.38)	47 (0.62)		7630 (18.07)
Body mass index, mean (SD), kg/m^2	27.40 (5.03)	26.12 (4.54)	<.001§	27.40 (5.03)
Smoking history				
Smoking status, No. (%)				
Never-smoker	21 614 (99.90)	21 (0.10)	<.001¶	21 635 (48.96)
Former smoker	18 698 (99.44)	106 (0.56)		18 804 (42.55)
Current smoker	3666 (97.71)	86 (2.29)		3752 (8.49)
Smoking duration, mean (SD), y				
Former smokers	24.16 (13.51)	34.82 (11.78)	<.001§	24.23 (13.53)
Current smokers	48.07 (8.53)	52.01 (5.52)	<.001§	48.17 (8.49)
Quit-time in former smokers,# mean (SD) , y	28.54 (11.96)	21.15 (11.42)	<.001§	28.50 (11.97)
Pack-years smoked,** mean (SD), y				
Former smokers	32.12 (28.02)	50.27 (32.20)	<.001§	32.23 (28.08)
Current smokers	57.32 (30.08)	69.67 (31.92)	<.001§	57.62 (30.18)

* Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial intervention arm participants who completed the Supplemental Questionnaire. All variables in this table come from the Supplemental Questionnaire, except for education, chronic obstructive pulmonary disease (COPD), and chest x-ray in the past 3 years, which come from the Baseline Questionnaire.

† Percentage values in parentheses are row percents for variable levels by lung cancer status. The row percents of lung cancer represent the cumulative incidence of lung cancer for that exposure (variable) level.

‡ Percentage values in parentheses are column percents for variable totals.

§ *P* value by two-sided Student *t* test.

‖ *P* value by two-sided Fisher exact test.

¶ *P* value by two-sided Wilcoxon rank-sum test.

Time in years since ever-smokers permanently quit smoking.

** Pack-years is the average number of packages of cigarettes smoked per day times the number of years smoked.

In external validation in the PLCO intervention arm ever-smokers, model 2 demonstrated high discrimination overall (ROC AUC = 0.784, 95% CI = 0.745 to 0.824; Figure 1,D). Predictive accuracy of the model remained high when the analysis was limited to women, men, whites, or nonwhites (including Hispanics), with ROC AUC ranging from 0.778 to 0.876 (Table 3).

ROC AUCs were compared between the full model 2 (Table 3), the nested model containing only the four smoking variables, and the nested model containing only pack-years smoked. The respective ROC AUCs were 0.809 (95% CI = 0.796 to 0.822), 0.790 (95% CI = 0.776 to 0.804), and 0.738 (95% CI = 0.723 to 0.752). The difference in ROC AUC between the full model 2 and the model with smoking variables only was highly statistically significant (−0.0197, 95% CI = −0.0262 to −0.0132 $P < .001$) as was the difference in ROC AUC between the model with all smoking variables *vs* the model with pack-years only (−0.0500, 95% CI = −0.0602 to −0.0399, $P < .001$). Thus, nonsmoking and non-pack-year smoking variables contributed to prediction.

Table 3. Logistic regression lung cancer prediction models prepared in all of the PLCO control arms (model 1) and in smokers only (model 2)*

Variable	Model 1† All PLCO control arm (N = 61999), OR (95% CI)	*P*	Model 2‡ Smokers only in PLCO control arm (N = 33049), OR (95% CI)	*P*
Age, per year				
Age spline 1	1.212 (1.110 to 1.322)	<.001	1.245 (1.130 to 1.372)	<.001
Age spline 2	0.732 (0.551 to 0.972)	.031	0.705 (0.505 to 0.984)	.040
Age spline 3	1.884 (0.917 to 3.869)	.085	2.205 (0.860 to 5.651)	.100
Education, per 1 of 7 levels change	0.930 (0.890 to 0.971)	.001	0.928 (0.887 to 0.971)	.001
BMI, per 1 unit change	0.970 (0.955 to 0.985)	<.001	0.972 (0.956 to 0.988)	.001
Family history of lung cancer, yes *vs* no	1.564 (1.323 to 1.848)	<.001	1.561 (1.313 to 1.856)	<.001
COPD, yes *vs* no	1.380 (1.153 to 1.651)	<.001	1.374 (1.145 to 1.648)	.001
Chest x-ray in past 3 y, per 1 of 3 levels	1.114 (1.020 to 1.217)	.017	1.117 (1.019 to 1.225)	.019
Pack-years smoked, per 1 pack-year				
PKYR spline 1	1.108 (1.073 to 1.144)	<.001	1.059 (1.044 to 1.074)	<.001
PKYR spline 2	0.500 (0.392 to 0.636)	<.001	0.949 (0.935 to 0.964)	<.001
Smoking duration, linear, per 1 y			1.012 (0.995 to 1.029)	.171
Smoking duration, splines, per 1 y				
Duration spline 1	0.986 (0.949 to 1.025)	.480		
Duration spline 2	1.127 (1.019 to 1.246)	.020		
Smoking quit-time in smokers, per 1 y				
Quit-time spline 1			0.945 (0.918 to 0.974)	<.001
Quit-time spline 2			1.047 (1.011 to 1.085)	.010
Smoking status				
Never/former	Baseline	<.001	Baseline	.010
Current	1.721 (1.426 to 2.077)		1.356 (1.077 to 1.708)	
Model performance statistics				
Hosmer–Lemeshow goodness of fit		.274		.416
Nagelkerke's R^2 (BOC)	0.199 (0.195)§		0.152 (0.147)§	
ROC AUC/*c* statistic (95% CI) (BOC)	0.859 (95% CI = 0.8476 to 0.8707) (0.857)§		0.809 (95% CI = 0.7957 to 0.8219) (0.805)§	
Calibration line	Slope (BOC) = 0.987§ Intercept (BOC) = −0.042§ Mean absolute error = 0.0009 0.9 quantile of absolute error = 0.0025		Slope (BOC) = 0.979§ Intercept (BOC) = −0.061§ Mean absolute error = 0.0014 0.9 quantile of absolute error = 0.0029	
External validation	ROC AUC (95% CI)		ROC AUC (95% CI)	
All validation sample	0.841 (0.813 to 0.870), n = 36363		0.784 (0.745 to 0.824), n = 15169	
Women	0.828 (0.781 to 0.876), n = 18988		0.779 (0.711 to 0.847), n = 6121	
Men	0.849 (0.815 to 0.883), n = 17375		0.789 (0.743 to 0.835), n = 9048	
Whites	0.843 (0.813 to 0.872; n = 33116)		0.778 (0.736 to 0.819), n = 13 900	
Nonwhites, including Hispanics	0.829 (0.719 to 0.939), n = 3247		0.876 (0.809 to 0.943), n = 1269	

* BMI = body mass index; BOC = bootstrap optimism corrected; CI = confidence interval; PKYR = pack-years smoked; PLCO = Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial; ROC AUC = receiver operator characteristic area under the curve.

† Splines for age, pack-years smoked, and smoking duration in model 1 are based on all PLCO control subjects. Knots for age were at 55, 60, 65, and 72 years. Knots for pack-years were at 0, 2.25, and 49 pack-years. Knots for smoking duration were at 0, 6, and 41 years.

‡ Splines for age, pack-years smoked, and quit-time in model 2 are based on the distribution of these variables in smokers only. Knots for age were at 55, 60, 64, and 72 years. Knots for pack-years were at 3.25, 23.25, and 63 pack-years. Knots for quit-time were at 0, 15, and 35 years.

§ Bootstrap optimism corrected estimate of model performance based on 200 bootstrap resamplings.

The predictive performance of a quadratic model analogous to model 2 (model 2b) but which uses quadratic terms in place of restricted cubic splines and allows comparison of the two modeling procedures was slightly below that of model 2 (external validation sample ROC AUC 0.779 *vs* 0.784; Table 4), and it is not as well calibrated according to the Hosmer–Lemeshow goodness-of-fit test (P = .008). Cox models (models 1c and 2c, Table 5) were prepared with the same predictor variables as in models 1 and 2.

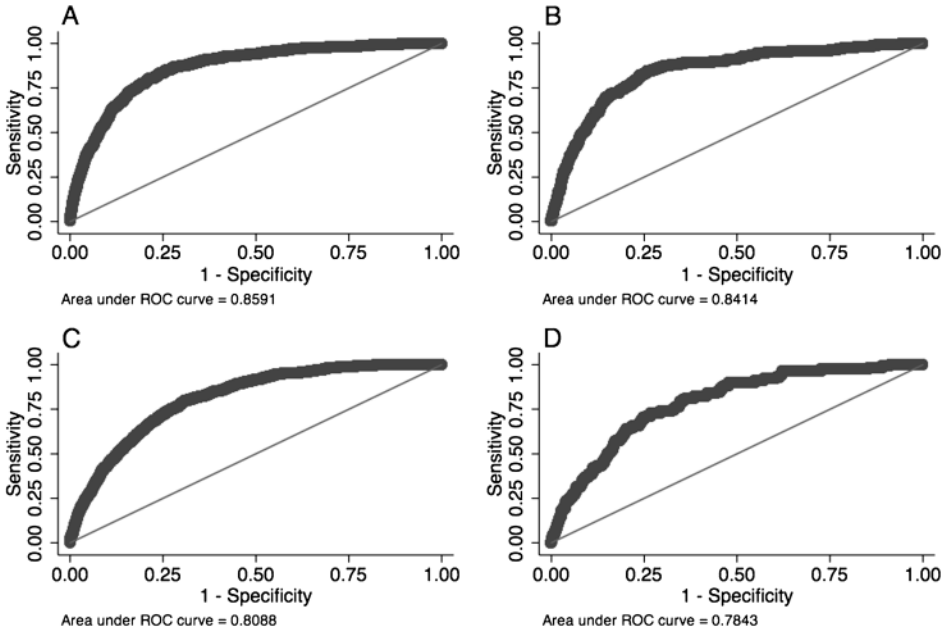

Figure 1. Receiver operator characteristic (ROC) plots for models 1 and 2. **A**) Model 1, developed in and applied to all PLCO control subjects; **B**) model 1 applied to external validation dataset (PLCO intervention arm); **C**) model 2 developed in and applied to smokers in the PLCO control arm; **D**) model 2 applied to smokers in external validation set (PLCO intervention arm). PLCO = Prostate, Lung, Colorectal, Ovarian cancer screening trial.

The hazard ratios were similar in direction, magnitude, and statistical significance to the odds ratios produced by the corresponding logistic regression models (Table 5), and the Cox model *c* statistics were close to the ROC AUCs from the logistic models. The Cox model performed slightly better than the logistic model, most likely because it uses accurate time-to-follow-up data and deals effectively with subjects lost to follow-up.

Discussion

Prediction models 1 and 2 demonstrated high discrimination and calibration. When both models were applied to the external validation sample, the ROC AUCs declined but remained high overall. The ROC AUCs also remained high in external validation for subsamples of women, men, whites, and nonwhites separately. In addition to model 2 variables, sex, race/ethnicity, and black men *vs* other race–sex groups pooled did not appear to be important predictors. Etzel et al. (13) developed a lung cancer risk prediction model in African Americans because existing models had been developed in whites, levels of risk are different for risk factors that African Americans share with whites, and unique group-specific risk factors exist for African Americans. (13) Our models appear to work equally well in whites and nonwhites.

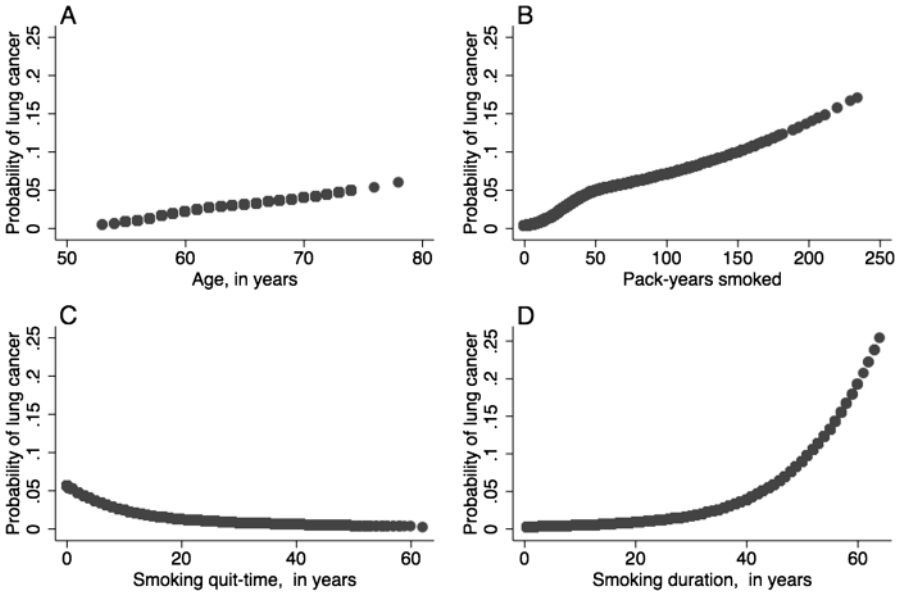

Figure 2. Estimated probabilities of developing lung cancer in smokers for four predictors in unadjusted logistic regression models. A) Age. B) Pack-years smoked. C) Smoking quit-time. D) Smoking duration. The nonlinear relationships between these predictor variables and lung cancer risk were estimated using restricted cubic splines. Splines for age, pack-years smoked, quit-time and smoking duration were prepared with knot placement based on the percentile distributions of these variables in smokers only. Knots for age were at 55, 60, 64, and 72 years. Knots for pack-years were at 3.25, 23.25 and 63 pack-years. Knots for quit-time were at 0, 15, and 35 years. Knots for duration were at 8, 28, and 45 years. The y-axis in each of the four figures is the same scale to allow comparison of the relative impacts of each predictor on lung cancer risk.

Table 4. Risk prediction model using quadratic terms in smokers analogous to model 2*

Variable	Model 2b† Smokers in PLCO control arm (N = 33 049), OR (95% CI)	*P*
Age, per year	1.885 (1.340 to 2.650)	<.001
Age squared	0.996 (0.993 to 0.998)	.002
Education, per 1 of 7 levels change	0.927 (0.885 to 0.970)	.001
BMI, per 1 unit change	0.971 (0.956 to 0.987)	<.001
Family history of lung cancer, yes *vs* no	1.560 (1.311 to 1.855)	<.001
COPD, yes *vs* no	1.370 (1.142 to 1.644)	.001
Chest x-ray in past 3 y, per 1 of 3 levels	1.122 (1.023 to 1.230)	.015
Pack-years smoked, per 1 pack-year	1.039 (1.031 to 1.048)	<.001
Pack-years squared	0.9998 (0.9997 to 0.9999)	<.001
Smoking duration, linear, per 1 y	1.013 (0.996 to 1.030)	.144
Smoking quit-time, per 1 y	0.941 (0.912 to 0.972)	<.001
Smoking quit-time squared	1.001 (1.000 to 1.002)	.011
Smoking status, current *vs* former	1.339 (1.067 to 1.682)	.012
Model performance statistics		
Hosmer–Lemeshow goodness of fit		.008
Nagelkerke's R^2	0.153 (0.152)	
ROC AUC/*c* statistic (BOC)	0.8099 (0.8077)	
Calibration line	Slope (BOC) = 0.988	
	Intercept (BOC) = −0.030	
	Mean absolute error = 0.0022	
	0.9 quantile of absolute error = 0.0047	
External validation	ROC AUC (95% CI)	
Validation sample, smokers only (n = 15 169)	0.779 (0.740 to 0.819)	

* BOC = bootstrap optimism corrected; BMI = body mass index; CI = confidence interval; PLCO = Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial; ROC AUC = receiver operator characteristic area under the curve.

† Bootstrap optimism corrected estimates of model performance based on 200 bootstrap resamplings.

Our estimates of ROC AUC corrected for overfitting by bootstrap were higher than those observed in the external validation sample, which is expected because it is difficult to bootstrap all phases of model development (42). For example, all phases of variable selection and evaluation were not bootstrapped, and the external validation sample differs in unmeasured ways from the development sample, and this was not accounted for in bootstrap validation.

To facilitate interpretation of model 2, a nomogram was prepared to allow estimation of the 9-year probability of lung cancer given an individual's specific risk factors. Individuals, patients, clinicians, and researchers can use this graphic tool to estimate lung cancer risk. The nomogram (Supplementary Figure 1, available online) converts specific variable level values into points, sums the points, and converts the overall point total for all predictor variables into the probability of lung cancer risk according to the logistic model 2.

To date, several lung cancer risk prediction models have been proposed. Most authors presented the ROC AUC or *c* statistic as the primary measure of predictive performance, but few have presented calibration data. Doll and Peto (6) and Prindiville et al. (7) described lung cancer risk prediction models but did not present predictive performances. Bach et al. (8) used prospective cohort data for smokers in the Carotene and Retinol Efficacy Trial to develop their prediction model. Their model predictors included age, sex, asbestos exposure, and smoking history. Cronin et al. (46) externally validated the Bach model (8) in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study control arm. The overall *c* statistic was 0.69. The initial Spitz et al. models (10) had cross-validated *c* statistics of 0.59, 0.63, and 0.65 in never, former, and current smokers, respectively. They attempted to improve the initial models by adding two markers of DNA repair capacity (12), which improved the ROC AUC from 0.67 to 0.70 in former smokers and from 0.68 to 0.73 in current smokers. The Etzel et al. (13) risk prediction model for blacks had a discrimination ability of 0.63 in external data. The risk prediction model produced by Cassidy et al. (11) included smoking duration, history of pneumonia, occupational exposure to asbestos, prior diagnosis of malignant tumor, and family history of lung cancer, and had an internally validated (cross-validation) ROC AUC of 0.70.

The predictive performance statistics of our models are not directly comparable with those of other models because differences in distributions of predictor variables can affect performance statistics (47). However, because our models include new predictor variables, including socioeconomic status (education), BMI, history of recent chest x-ray, and COPD; an increased number of smoking variables; and inclusion of nonlinear effects, they may have better predictive accuracy than older models. Socioeconomic status may function as a predictor of lung cancer because it is a marker of unmeasured environmental, occupational, or behavioral exposures. Higher BMI has been associated with reduced risk of lung cancer in several studies (16–25). Some research suggests that this association might have a biological basis. Both smoking-related DNA adducts measured in peripheral blood lymphocytes and oxidative DNA damage measured by levels of urinary 8-hydroxydeoxyguanosine appear to be inversely associated with BMI, adjusted for smoking (48–50). This suggests that lean individuals might be more vulnerable

Table 5. Cox regression predictive models prepared for lung cancer in all of the PLCO control subjects (model 1) and in smokers only (model 2)*

Variable	Model 1c†		Model 2c‡	
	All PLCO control arm (N = 61 986), HR (95% CI)	*P*	Smokers only in PLCO control arm (N = 33 039), HR (95% CI)	*P*
Age, per year				
Age spline 1	1.166 (1.071 to 1.271)	<.001	1.198 (1.090 to 1.317)	<.001
Age spline 2	0.785 (0.596 to 1.035)	.086	0.761 (0.550 to 1.052)	.098
Age spline 3	1.666 (0.829 to 3.348)	.152	1.907 (0.766 to 4.747)	.165
Education, per 1 of 7 levels change	0.923 (0.884 to 0.963)	.001	0.921 (0.881 to 0.962)	<.001
BMI, per 1 unit change	0.972 (0.957 to 0.987)	<.001	0.974 (0.959 to 0.990)	.001
Family history of lung cancer, yes *vs* no	1.561 (1.330 to 1.832)	<.001	1.555 (1.317 to 1.835)	<.001
COPD, yes *vs* no	1.446 (1.218 to 1.715)	<.001	1.442 (1.212 to 1.714)	<.001
Chest x-ray in past 3 y, per 1 of 3 levels	1.109 (1.017 to 1.208)	.019	1.111 (1.016 to 1.215)	.021
Pack-years smoked, per 1 pack-year				
PKYR spline 1	1.106 (1.072 to 1.141)	<.001	1.058 (1.043 to 1.073)	<.001
PKYR spline 2	0.505 (0.398 to 0.640)	<.001	0.950 (0.936 to 0.965)	<.001
Smoking duration, linear, per 1 y			1.014 (0.998 to 1.032)	.111
Smoking duration, splines, per 1 y				
Duration spline 1	0.987 (0.950 to 1.025)	.484		
Duration spline 2	1.132 (1.026 to 1.250)	.014		
Smoking quit-time in smokers, per 1 y				
Quit-time spline 1			0.946 (0.918 to 0.974)	<.001
Quit-time spline 2			1.048 (1.012 to 1.085)	.008
Smoking status				
Never/former	Baseline	<.001	Baseline	.011
Current	1.687 (1.406 to 2.024)		1.337 (1.070 to 1.670)	
Model performance statistics				
C statistic	0.8606		0.8101	

* These Cox models parallel the logistic models presented in Table 3. BMI = body mass index; CI = confidence interval; PKYR = pack-years; HR = hazard ratio; PLCO = Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial.

† Splines for age, pack-years smoked, and smoking duration in model 1c are based on all PLCO control subjects. Knots for age were at 55, 60, 65, and 72 years. Knots for pack-years were at 0, 2.25, and 49 pack-years. Knots for smoking duration were at 0, 6, and 41 years.

‡ Splines for age, pack-years smoked, and quit-time in model 2c are based on the distribution of these variables in smokers only. Knots for age were at 55, 60, 64, and 72 years. Knots for pack-years were at 3.25, 23.25, and 63 pack-years. Knots for quit-time were at 0, 15, and 35 years.

to smoking carcinogen-related DNA damage. Recent chest x-ray may be a marker of pulmonary disease or chronic inflammation and thus serve as a predictor of lung cancer. Smoking is the primary causal agent of lung cancer, and thus it is expected that models describing smoking exposure in greater detail would have improved predictive abilities. Furthermore, many associations in nature are nonlinear, and our models were improved by including nonlinear components. For example, model 2 with nonlinear terms had an ROC AUC of 0.809, and the same model with linear terms replacing the spline terms had an ROC AUC of 0.804, with the difference statistically significant (−0.005, 95% CI = −0.009 to −0.001; $P = .009$).

This study had several limitations. The PLCO study participants were aged 55–74 years at study entry and were on average of higher socioeconomic status than the general population and may exhibit a healthy volunteer effect (51), which may limit external generalizability. However, with the exception of educational level, the model predictors appear to have a biological basis that is expected to be independent of age and socioeconomic status. Data on several potentially useful predictors were unavailable for analysis, including exposure to radon, asbestos, secondhand smoke, occupational carcinogens, and history of adult pneumonia. Inclusion of these variables might have added to our risk prediction models. However, because predictors must have strong associations with lung cancer to have an impact on prediction (52), and these missing predictors have modest associations with lung cancer, their inclusion would have led to only small improvements in prediction. In addition, because our external validation sample came from the same referent population as our model development sample, our models may not perform as well when applied to other population samples.

This work also had several strengths. We used a prospective design, which does not have the methodological weaknesses of case–control studies (10,11,13), which cannot estimate incidence or absolute risk directly, and which is vulnerable to selection and recall biases. In some studies, the case patients and control subjects were matched on age (10,11), sex (11), and smoking status (10), which prevents effective assessment of these predictors because they have been forced to be similar by study design, and case patients and control subjects were not taken from the same referent population (10). One study chose "healthy" control subjects as the comparison group for case patients (10), which might lead to selection bias. For example, such sampling could lead to an exaggerated effect for emphysema, if individuals with emphysema were excluded from the control group, but not the case patient group. In contrast, the PLCO sampling was population based and

represents many different regions of the United States, resulting in improved internal and external validity.

We also used updated statistical methods, including modeling of nonlinear effects using restricted cubic splines and bootstrap correction of predictive performance optimism, which could improve the accuracy of predictive models. For example, Bach et al. (7) placed continuous smoking exposure data into four categories (7). The loss of information resulting from categorization of continuous data could lead to loss of predictive ability. In another analysis (10), selection of predictor variables for entry into the multivariable models was based on the criterion of *P* less than .05 in univariate analysis (10), which could result in important predictors being left out of models more often than when less stringent *P* are used. In the same study, final multivariable models were also restricted to predictors that had *P* less than .05, which could result in suboptimal predictive performance (41,42). Spitz et al. (10) found associations with lung cancer for some predictors only in subsets of their sample, and criteria for positivity changed from one group to another, which suggests that data exploration, use of optimal cut points, and multiple comparisons may have contributed to their findings. For example, Spitz et al. (10) reported that in former smokers, a family history of at least two of any cancers *vs* one or fewer was predictive, whereas in current smokers, a family history of at least one smoking-related cancer *vs* none was predictive (10). Such analytical practices are likely to lead to overfitting of models and lack of reproducibility (53).

Because lung cancer was a primary endpoint of the PLCO, and follow-up and monitoring for lung cancer were meticulously maintained, data quality was high. The PLCO is a large mature study with enough outcome events to allow estimation with precision and reduced tendency to overfit models. Because the PLCO trial was not restricted to high-risk individuals, modeling applicable to the general population was possible. External validation of the models provided a realistic sense of the predictive potential of the models.

In future research, it will be important to conduct additional external validations of our models in diverse samples. Although the current models demonstrated high predictive performance, the models can be improved. Genomewide association studies have identified inherited susceptibility variants for lung cancer at chromosomal loci 15q25 (54–56), 5p15 (57–60), and 6p21 (58). Future studies should investigate whether genetic polymorphism data contribute independent predictive information to models and whether they explain the predictive effect of family history of lung cancer. Numerous serum biomarkers associated with lung cancer (12,61–64) and pulmonary function (65) data need to be evaluated in risk prediction models.

In conclusion, our two lung cancer risk prediction models demonstrated high discrimination and calibration and are expected to be able to discriminate between high- and low-risk individuals. Other high-quality data sources in cohort settings should be used to validate and extend our findings.

References

1. Canadian Cancer Society/National Cancer Institute of Canada. *Canadian Cancer Statistics 2010.* Toronto, Canada: Canadian Cancer Society; 2010.
2. Jemal A, Siegel R, Xu J, et al. Cancer Statistics, 2010. *CA Cancer J Clin.* 2010;60(5):277–300.
3. Parkin DM, Bray F, Ferlay J, et al. Global cancer statistics, 2002. *CA Cancer J Clin.* 2005;55(2):74–108.
4. Aberle DR, Berg CD, Black WC, et al. The national lung screening trial: overview and study design. *Radiology.* 2011;258(1):243–253.
5. National Cancer Institute (U.S.). Lung cancer trial results show mortality benefit with low-dose CT. http://www.cancer.gov/newscenter/pressreleases/NLSTresultsRelease.
6. Doll R, Peto R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. *J Epidemiol Community Health.* 1978;32(4):303–313.
7. Prindiville SA, Byers T, Hirsch FR, et al. Sputum cytological atypia as a predictor of incident lung cancer in a cohort of heavy smokers with airflow obstruction. *Cancer Epidemiol Biomarkers Prev.* 2003;12(10):987–993.
8. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst.* 2003;95(6):470–478.
9. Cassidy A, Myles JP, Liloglou T, et al. Defining high-risk individuals in a population-based molecular-epidemiological study of lung cancer. *Int J Oncol.* 2006;28(5):1295–1301.
10. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. *J Natl Cancer Inst.* 2007;99(9):715–726.
11. Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer.* 2008;98(2):270–276.
12. Spitz MR, Etzel CJ, Dong Q, et al. An expanded risk prediction model for lung cancer. *Cancer Prev Res (Phila Pa).* 2008;1(4):250–254.
13. Etzel CJ, Kachroo S, Liu M, et al. Development and validation of a lung cancer risk prediction model for African-Americans. *Cancer Prev Res (Phila Pa).* 2008;1(4):255–265.
14. Alberg AJ, Ford JG, Samet JM. Epidemiology of lung cancer: ACCP evidence-based clinical practice guidelines. . (2nd edition). *Chest.* 2007;132(3) (suppl):29S–55S.
15. Tammemagi MC, Freedman MT, Pinsky PF, et al. Prediction of true positive lung cancers in individuals with abnormal suspicious chest radiographs: a prostate, lung, colorectal, and ovarian cancer screening trial study. *J Thorac Oncol.* 2009;4(6):710–721.
16. Nomura A, Heilbrun LK, Stemmermann GN. Body mass index as a predictor of cancer in men. *J Natl Cancer Inst.* 1985;74(2):319–323.
17. Hoffmans MD, Kromhout D, Coulander CD. Body Mass Index at the age of 18 and its effects on 32-year-mortality from coronary heart disease and cancer. A nested case-control study among the entire 1932 Dutch male birth cohort. *J Clin Epidemiol.* 1989;42(6):513–520.
18. Knekt P, Heliovaara M, Rissanen A, et al. Leanness and lung-cancer risk. *Int J Cancer.* 1991;49(2):208–213.
19. Kark JD, Yaari S, Rasooly I, et al. Are lean smokers at increased risk of lung cancer? The Israel Civil Servant Cancer Study. *Arch Intern Med.* 1995;155(22):2409–2416.
20. Singh PN, Lindsted KD. Body mass and 26-year risk of mortality from specific diseases among women who never smoked. *Epidemiology.* 1998;9(3):246–254.
21. Olson JE, Yang P, Schmitz K, et al. Differential association of body mass index and fat distribution with three major histologic types of lung cancer: evidence from a cohort of older women. *Am J Epidemiol.* 2002;156(7):606–615.
22. Kabat GC, Miller AB, Rohan TE. Body mass index and lung cancer risk in women. *Epidemiology.* 2007;18(5):607–612.
23. Kondo T, Hori Y, Yatsuya H, et al. Lung cancer mortality and body mass index in a Japanese cohort: findings from the Japan Collaborative Cohort Study (JACC Study). *Cancer Causes Control.* 2007;18(2):229–234.
24. Kabat GC, Kim M, Hunt JR, et al. Body mass index and waist circumference in relation to lung cancer risk in the women's health initiative. *Am J Epidemiol.* 2008.
25. Kollarova H, Machova L, Horakova D, et al. Is obesity a preventive factor for lung cancer? *Neoplasma.* 2008;55(1):71–73.
26. Shacter E, Weitzman SA. Chronic inflammation and cancer. *Oncology (Williston Park).* 2002;16(2):217–226. 229; discussion 230–232.
27. Schwartsburd PM. Chronic inflammation as inductor of pro-cancer microenvironment: pathogenesis of dysregulated feedback control. *Cancer Metastasis Rev.* 2003;22(1):95–102.
28. Baniyash M. Chronic inflammation, immunosuppression and cancer: new insights and outlook. *Semin Cancer Biol.* 2006;16(1):80–88.

29. Malkinson AM, Bauer A, Meyer A, et al. Experimental evidence from an animal model of adenocarcinoma that chronic inflammation enhances lung cancer risk. *Chest.* 2000;117(5) (suppl 1):228S.
30. Blanco D, Vicent S, Fraga MF, et al. Molecular analysis of a multistep lung cancer model induced by chronic inflammation reveals epigenetic regulation of p16 and activation of the DNA damage response pathway. *Neoplasia.* 2007;9(10):840–852.
31. Walser T, Cui X, Yanagawa J, et al. Smoking and lung cancer: the role of inflammation. *Proc Am Thorac Soc.* 2008;5(8):811–815.
32. Lee G, Walser TC, Dubinett SM. Chronic inflammation, chronic obstructive pulmonary disease, and lung cancer. *Curr Opin Pulm Med.* 2009; 15(4):303–307.
33. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials.* 2000;21(6) ((suppl):273S–309S.
34. Gohagan JK, Prorok PC, Hayes RB, et al. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials.* 2000;21(6) (suppl):251S–272S.
35. Oken MM, Marcus PM, Hu P, et al. Baseline chest radiograph for lung cancer detection in the randomized Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial. *J Natl Cancer Inst.* 2005;97(24):1832–1839.
36. Parkin DM, Moss SM. Lung cancer screening: improved survival but no reduction in deaths–the role of "overdiagnosis". *Cancer.* 2000;89(11) (suppl):2369–2376.
37. Marcus PM, Bergstralh EJ, Zweig MH, et al. Extended lung cancer incidence follow-up in the Mayo Lung Project and overdiagnosis. *J Natl Cancer Inst.* 2006;98(11):748–756.
38. Schwartz AG, Swanson GM. Lung carcinoma in African Americans and whites. A population-based study in metropolitan Detroit, Michigan. *Cancer.* 1997;79(1):45–52.
39. Haiman CA, Stram DO, Wilkens LR, et al. Ethnic and racial differences in the smoking-related risk of lung cancer. *N Engl J Med.* 2006;354(4):333–342.
40. Edwards BK, Ward E, Kohler BA, et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer.* 2010;116(3):544–573.
41. Harrell FE. *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis.*. New York, NY: Springer; 2001.
42. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.*. New York, NY: Springer; 2009.
43. Hosmer DW Jr., Lemeshow S. *Applied Logistic Regression.* 2nd ed. New York, NY: John Wiley & Sons, Inc; 1999.
44. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics.* 2005;6(2):227–239.
45. R Development Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2009.
46. Cronin KA, Gail MH, Zou Z, et al. Validation of a model of lung cancer risk prediction among smokers. *J Natl Cancer Inst.* 2006;98(9):637–640.
47. Whittemore AS. Evaluating health risk models. *Stat Med.* 2010;29(23): 2438–2452.
48. Godschalk RW, Feldker DE, Borm PJ, et al. Body mass index modulates aromatic DNA adduct levels and their persistence in smokers. *Cancer Epidemiol Biomarkers Prev.* 2002;11(8):790–793.
49. Mizoue T, Kasai H, Kubo T, et al. Leanness, smoking, and enhanced oxidative DNA damage. *Cancer Epidemiol Biomarkers Prev.* 2006;15(3): 582–585.
50. Mizoue T, Tokunaga S, Kasai H, et al. Body mass index and oxidative DNA damage: a longitudinal study. *Cancer Sci.* 2007;98(8):1254–1258.
51. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am J Epidemiol.* 2007;165(8):874–881.
52. Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol.* 2004;159(9):882–890.
53. Altman DG, Lausen B, Sauerbrei W, et al. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst.* 1994;86(11):829–835.
54. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature.* 2008;452(7187):633–637.
55. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 2008;40(5):616–622.
56. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature.* 2008;452(7187):638–642.
57. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet.* 2008;40(12):1404–1406.
58. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet.* 2008;40(12):1407–1409.
59. Rafnar T, Sulem P, Stacey SN, et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet.* 2009; 41(2):221–227.
60. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet.* 2009;85(5):679–691.
61. Chaturvedi AK, Caporaso NE, Katki HA, et al. C-reactive protein and risk of lung cancer. *J Clin Oncol.* 2010;28(16):2719–2726.
62. Yee J, Sadar MD, Sin DD, et al. Connective tissue-activating peptide III: a novel blood biomarker for early lung cancer detection. *J Clin Oncol.* 2009;27(17):2787–2792.
63. Chaturvedi AK, Gaydos CA, Agreda P, et al. Chlamydia pneumoniae infection and risk for lung cancer. *Cancer Epidemiol Biomarkers Prev.* 2010;19(6):1498–1505.
64. Church TR, Anderson KE, Caporaso NE, et al. A prospectively measured serum biomarker for a tobacco-specific carcinogen and lung cancer in smokers. *Cancer Epidemiol Biomarkers Prev.* 2009;18(1):260–266.
65. Tammemagi MC, Lam SC, McWilliams AM, et al. Incremental value of pulmonary function and sputum DNA image cytometry in lung cancer risk prediction. *Cancer prevention research* 2011;4(4):552–61.

Funding

This research was supported by contracts from the Division of Cancer Prevention, National Cancer Institute, National Institute of Health, Department of Health and Human Services. The funders did not have any involvement in the design of this ancillary study; the collection, analysis, and interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Notes

This report describes an ancillary study in the National Cancer Institute's Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial, which is a federally funded registered clinical trial (ClinicalTrials.gov Identifier: NCT00002540).The authors thank the Screening Center investigators and staff of the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. Most importantly, we acknowledge the study participants for their contributions to making this study possible. We thank Professor Frank E. Harrell, Jr, for helpful guidance regarding risk prediction modeling and interpretation. Dr C. Martin Tammemagi's involvement in the PLCO has been through his affiliation with Henry Ford Health System in Detroit, Michigan.

Affiliations of authors: Department of Community Health Sciences, Brock University, Ontario, Canada (CMT); National Cancer Institute, National Institutes of Health, Bethesda, MD (PFP); Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD (NEC); Division of Pulmonary & Critical Care Medicine, Henry Ford Health System, Detroit, MI (PAK); Department of Hematology/Oncology, Marshfield Clinic, Marshfield, WI (WGH); Department of Environmental Health Sciences, University of Minnesota, Minneapolis, MN (TRC); Information Management Services, Inc, Rockville, MD (TLR, JC); Hubert H. Humphrey Cancer Center, North Memorial Health Care, Robbinsdale, MN (MMO); Early Detection Research Group, Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, Bethesda, MD (CDB); Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, National Institute of Health, Bethesda, MD (PCP).