VOLUME 35 · NUMBER 8 · MARCH 10, 2017

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study

David C. Muller, Mattias Johansson, and Paul Brennan

Author affiliations and support information (if applicable) appear at the end of this article.

Corresponding author: David C. Muller, Department of Epidemiology and Biostatistics, Imperial College London, St Mary's Campus, London W2 1PG, United Kingdom; e-mail: david.muller@imperial.ac.uk.

© 2017 by American Society of Clinical Oncology

0732-183X/17/3508w-861w/$20.00

Published at jco.org on January 17, 2017.

A B S T R A C T

Purpose

Several lung cancer risk prediction models have been developed, but none to date have assessed the predictive ability of lung function in a population-based cohort. We sought to develop and internally validate a model incorporating lung function using data from the UK Biobank prospective cohort study.

Methods

This analysis included 502,321 participants without a previous diagnosis of lung cancer, predominantly between 40 and 70 years of age. We used flexible parametric survival models to estimate the 2-year probability of lung cancer, accounting for the competing risk of death. Models included predictors previously shown to be associated with lung cancer risk, including sex, variables related to smoking history and nicotine addiction, medical history, family history of lung cancer, and lung function (forced expiratory volume in 1 second [FEV1]).

Results

During accumulated follow-up of 1,469,518 person-years, there were 738 lung cancer diagnoses. A model incorporating all predictors had excellent discrimination (concordance (c)-statistic [95% CI] = 0.85 [0.82 to 0.87]). Internal validation suggested that the model will discriminate well when applied to new data (optimism-corrected c-statistic = 0.84). The full model, including FEV1, also had modestly superior discriminatory power than one that was designed solely on the basis of questionnaire variables (c-statistic = 0.84 [0.82 to 0.86]; optimism-corrected c-statistic = 0.83; $p_{FEV1} = 3.4 \times 10^{-13}$). The full model had better discrimination than standard lung cancer screening eligibility criteria (c-statistic = 0.66 [0.64 to 0.69]).

Conclusion

A risk prediction model that includes lung function has strong predictive ability, which could improve eligibility criteria for lung cancer screening programs.

J Clin Oncol 35:861-869. © 2017 by American Society of Clinical Oncology

INTRODUCTION

As the result of common incidence and poor prognosis, lung cancer is responsible for more than 1.5 million deaths worldwide each year.[1] In the United Kingdom alone, in 2011, there were more than 43,400 new cases of lung cancer diagnosed, and 35,184 deaths from lung cancer.[2] Lung cancer survival rates are poor overall, but prognosis varies strongly by disease stage at diagnosis. In the United Kingdom, the 1-year relative survival rate is approximately 15% for patients with stage IV disease, but approximately 70% for patients with stage I disease.[3] Currently, the majority of lung cancers are diagnosed at stage III or IV.[3] Taken together, these facts indicate that an increase in the proportion of lung cancer cases diagnosed at an earlier stage could have a profound effect on overall lung cancer survival rates.

It has been shown that screening for lung cancer may be beneficial. The US National Lung Screening Trial (NLST), a randomized study of more than 50,000 people, showed that screening using low-dose computed tomography (CT) reduced lung cancer mortality rates by 20%.[4] Baseline results from the UK Lung Cancer Screening Trial were recently published,[5] and follow-up results for lung cancer mortality from this and several other European screening trials are eagerly awaited.[6] Despite the promise of CT

ASSOCIATED CONTENT

See accompanying Editorial on page 818

Appendix

DOI: 10.1200/JCO.2016.69.2467

DOI: 10.1200/JCO.2016.69.2467

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

screening, accurate identification of the appropriate population to be screened is essential to maximizing screening program efficacy.

The NLST enrolled participants between the ages of 55 and 75 years who were ostensibly at high risk of lung cancer on the basis of their smoking behavior, having accumulated a smoking history of at least 30 pack-years, as well as being current smokers or having quit within the last 15 years. Even within this high-risk population, it has been shown that the 20% of participants at lowest risk of lung cancer in the NLST accounted for only 1% of the lung cancer deaths prevented by CT screening.[7] This underlines the importance of having a precisely defined population to be screened. Limiting screening to those at sufficiently high risk, who are most likely to benefit from screening, will maximize the benefit in terms of early detection, and minimize the harms associated with screening-detected false-positives among those at lower risk.

The identification of individuals at high risk requires an accurate and powerful risk prediction model. Several models for lung cancer have been developed, largely on the basis of strong risk factors such as tobacco smoking behaviors, family history, and occupational exposures.[8-17] In this study, we developed, internally validated, and evaluated a risk prediction model using data from the UK Biobank prospective cohort study. We focused on established risk factors for lung cancer that are routinely available or easily ascertained by general practitioners during a standard consultation. These risk factors included smoking behaviors (including number of cigarettes smoked per day or time since smoking cessation)[18-23] and nicotine addiction[24]; personal medical history, including previous cancer diagnoses, emphysema and bronchitis,[25] and allergic conditions[26,27]; as well as family history of lung cancer.[28-32] In addition to established risk factors, we sought to evaluate the predictive performance of a measure of lung function, the forced expiratory volume in 1 second (FEV1).

METHODS

The UK Biobank

The UK Biobank is a prospective cohort study of more than 500,000 people who were between 37 and 73 years of age at the time of recruitment, which occurred between 2006 and 2010. At baseline assessment, participants completed a questionnaire that included items assessing general health and medical history, lifestyle and diet, and family history of disease. Participants also underwent a physical assessment that included measurements of blood pressure, pulse rate, height, weight, body fat, vision, fitness, grip strength, bone density, and lung function. Follow-up for cancer incidence and death was conducted via linkage to cancer and death registries. Participants were followed up from the date of baseline attendance until the date of diagnosis of an invasive primary lung cancer, date of death, or January 1, 2012, whichever occurred first. All participants provided written consent, and the study protocol was approved by the North West Multicenter Research Ethics Committee in the United Kingdom.

Model Development and Assessment

We restricted our investigation to factors that have been consistently associated with lung cancer risk in prospective studies. In addition, these factors were limited to those that are routinely available or can be easily ascertained by general practitioners during a standard consultation. In particular, we investigated sex, variables related to smoking history and addiction to nicotine, personal medical history, family history of lung cancer, and lung function assessed by spirometry (ie, FEV1).

To calculate the probability of lung cancer in the presence of the competing risk of death, we separately modeled the hazard of lung cancer and the hazard of death from all causes. We used flexible parametric survival models on the cumulative hazard scale to estimate baseline hazards and hazard ratios.[33] Age was used as the timescale and was modeled by a restricted cubic spline with boundary knots at 40 and 70 years, and internal knots at 60 and 65 years. Separate models for lung cancer risk were fitted for never, former, and current smoker participants. The model for risk of death included sex, prebaseline cancer, and smoking status as covariates. Participants with missing data on any of the covariates were excluded from the primary analyses, and sensitivity analyses on the basis of multiply imputed data were conducted. We used restricted cubic splines to model potential nonlinear relationships with continuous variables. Nonproportional hazards were investigated by fitting interactions between covariates and the timescale. The 2-year absolute risk of lung cancer was estimated by the cumulative incidence function conditional on age at baseline attendance (Appendix, online only).

Model calibration was assessed by plotting the mean predicted probability against the mean observed probability of lung cancer by tenth of predicted risk. Model discrimination was assessed with the concordance statistic (c-statistic). Internal validation of model discrimination was assessed by calculating the bootstrap optimism-corrected c-statistic with 100 bootstrap replications,[34] as well as a development/validation split where the model was fitted to 80% of the data and evaluated on the remaining 20%.

We conducted sensitivity analyses, in which we excluded participants recorded as having extreme values for FEV1 ($>$ 10 L) or number of cigarettes smoked per day ($>$ 80), and the model parameters were not materially affected. All statistical analyses were conducted using Stata 12.1 (Stata Corporation, College Station, Texas) and R version 3.2.1.[35]

RESULTS

Of the 502,649 UK Biobank participants, 325 had a prebaseline diagnosis of lung cancer, and three were reported to have died on

Table 1. Cohort Follow-Up, Cancer Diagnoses, and Deaths Among UK Biobank Participants (n = 502,321) With No Prebaseline Diagnosis of Lung Cancer and Participants Included in the Risk Models (n = 443,535)

		Lung Cancers		Deaths		Incidence Rate Per 10,000*		
Variable	Person-Years	Total	In Models†	Total	In Models†	Overall	Women	Men
Smoking Status								
Never	999,350	161	136	1,995	1,616	8	7	8
Former	353,359	325	228	1,334	917	31	30	33
Current	116,808	252	182	627	413	95	77	108
Overall	1,469,518	738	546	3,956	2,946	20	16	24

*Age-standardized incidence rate for lung cancer, using the 2013 European Standard Population (European Commission, http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-13-028).
†Number of events included in models, after excluding individuals with missing data.

 © 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

the same day that they attended the study center. Among the remaining 502,321 participants, there were 738 incident lung cancer diagnoses and 3,956 deaths in a total of 1,469,518 person-years of follow-up. The majority of lung cancers were identified among participants who were current or former smokers (Table 1). Demographic characteristics and lung cancer risk factors are listed in Table 2. After excluding participants with missing covariate data, there were 443,535 participants included in the models, of whom 546 had a lung cancer diagnosis (Table 1).

Table 2. Distribution of Demographic Characteristics and Covariates

Variable	No Lung Cancer (n = 501,583)		Lung Cancer (n = 738)		Total (n = 502,321)	
	No.	%	No.	%	No.	%
Sex						
Female	272,983	54	330	45	273,313	54
Male	228,600	46	408	55	229,008	46
Age at recruitment, years						
37-49	117,871	23	27	4	117,898	23
50-54	76,273	15	54	7	76,327	15
55-59	90,667	18	122	17	90,789	18
60-64	121,168	24	241	33	121,409	24
65-73	95,604	19	294	40	95,898	19
Cigarette smoking status						
Never	342,146	68	161	22	342,307	68
Former	120,474	24	325	44	120,799	24
Current	38,963	8	252	34	39,215	8
Previous cancer diagnosis						
No	476,540	95	648	88	477,188	95
Yes	25,043	5	90	12	25,133	5
Family history of lung cancer						
No	439,581	88	582	79	440,163	88
Yes	62,002	12	156	21	62,158	12
History of asthma						
No	441,694	88	635	86	442,329	88
Yes	57,754	12	96	13	57,850	12
Missing data	2,135	0	7	1	2,142	0
History of allergy and/or eczema						
No	383,894	77	628	85	384,522	77
Yes	115,554	23	103	14	115,657	23
Missing data	2,135	0	7	1	2,142	0
History of emphysema and/or bronchitis						
No	491,213	98	655	89	491,868	98
Yes	8,235	2	76	10	8,311	2
Missing data	2,135	0	7	1	2,142	0
Maximum FEV1, liters						
0.02-2.28	112,614	22	289	39	112,903	22
2.29-2.75	113,574	23	141	19	113,715	23
2.76-3.33	112,484	22	91	12	112,575	22
3.34-22.3	114,402	23	66	9	114,468	23
Missing data	48,509	10	151	20	48,660	10
Cigarettes smoked per day (former and current smokers)						
0-9	17,978	11	37	6	18,015	11
10-14	29,757	19	72	12	29,829	19
15-19	25,856	16	71	12	25,927	16
20-29	56,023	35	232	40	56,255	35
30-140	20,589	13	121	21	20,710	13
Missing data	9,234	6	44	8	9,278	6
Age stopped smoking, years						
8-29	24,892	21	13	4	24,905	21
30-39	35,500	29	34	10	35,534	29
40-49	30,679	25	70	22	30,749	25
50-69	28,718	24	207	64	28,925	24
Missing data	685	1	1	0	686	1
Difficulty not smoking for 1 day, current smokers						
Very easy	2,294	6	1	0	2,295	6
Fairly easy	7,687	20	26	10	7,713	20
Fairly difficult	12,687	33	86	34	12,773	33
Very difficult	13,058	34	122	48	13,180	34
Missing data	3,237	8	17	7	3,254	8

Abbreviation: FEV1, forced expiratory volume in 1 second.

© 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

Table 3. Hazard Ratios and 95% CIs for the Lung Cancer Risk Models

Predictor	Never Smokers		Former Smokers		Current Smokers	
	HR	95% CI	HR	95% CI	HR	95% CI
Male (*v* female)	1.89	1.25 to 2.87	1.61	1.18 to 2.21	1.52	1.07 to 2.14
Previous cancer diagnosis	2.36	1.43 to 3.89	2.18	1.50 to 3.18	1.40	0.81 to 2.43
Family history of lung cancer (first-degree relative)	0.82	0.48 to 1.41	1.61	1.19 to 2.18	1.50	1.07 to 2.12
History of emphysema and/or chronic bronchitis	2.66	0.97 to 7.27	1.57	1.00 to 2.47	1.90	1.18 to 3.05
History of hay fever, allergic rhinitis, or eczma	1.18	0.79 to 1.74	0.83	0.57 to 1.21	0.51	0.29 to 0.90
Maximum FEV1, liters	0.57	0.43 to 0.77	0.50	0.40 to 0.63	0.62	0.48 to 0.80
Age stopped smoking, years			1.05	1.04 to 1.07		
First spline basis of cigarettes per day (knots: 4, 15, 30)			1.10	1.04 to 1.16	1.03	0.97 to 1.08
Second spline basis of cigarettes per day (knots: 4, 15, 30)			1.00	1.00 to 1.00	1.00	1.00 to 1.00
Difficulty of not smoking for 1 day (easy *v* difficult)					0.48	0.29 to 0.77

Abbreviations: FEV1, forced expiratory volume in 1 second; HR, hazard ratio.

Table 3 lists the hazard ratios and 95% CIs for the lung cancer risk models that include FEV1. Maximum FEV1 was strongly inversely associated with lung cancer risk regardless of smoking status, and a likelihood ratio test of the FEV1 terms suggested that it contributed substantially to the model ($P = 3.4 \times 10^{-13}$). Similar estimates were obtained for all covariates when FEV1 was excluded from the models (Table 4). The analysis of multiply imputed data also yielded similar estimates (Table 5).

Figure 1 shows the observed 2-year risk of lung cancer (calculated using the Kaplan-Meier estimate) by tenth of model-predicted risk. The predicted and observed probabilities agreed well, suggesting that the model-predicted risks were well calibrated. The receiver operating characteristic curve for the fitted model is shown in Figure 2A, and the sensitivity and specificity of the model are plotted by the predicted probability cutoff in Figure 2B. When applied to all participants regardless of smoking status, the model was able to discriminate well between those who were diagnosed with lung cancer within 2 years and those who were not (c-statistic = 0.85 [0.82 to 0.87]). Discrimination was similarly good when applied only to current or former smokers (c-statistic = 0.82 [0.79 to 0.84]).

Internal validation suggested that the model will perform well when applied to new observations or data sets (optimism-corrected c-statistic = 0.84; development/validation split sample c-statistic = 0.83). Excluding FEV1 and using only demographic data and questionnaire-derived variables led to a model with slightly lower, but still substantial, discriminatory power when applied to the whole cohort (c-statistic = 0.84 [0.82 to 0.86]; optimism-corrected c-statistic = 0.83). Importantly, the inclusion of FEV1 also improved discrimination among never smokers (c-statistic = 0.74 [0.69 to 0.78] *v* 0.72 [0.67 to 0.76]; $P = .02$).

To assess the possible implications of adopting a model-based inclusion criterion for lung cancer screening, we compared the discrimination and sensitivity of a model-based strategy with that of the NLST inclusion criteria. The discrimination of the NLST criteria was substantially lower than the model-predicted probabilities when applied to the UK Biobank data, with a c-statistic of 0.66 (0.64 to 0.69) versus the optimism-corrected c-statistic of 0.84 for our full model. Instead of applying the NLST criteria, we could take a model-based approach to identifying participants at high risk on the basis of their predicted probability of lung cancer. In this approach, the sensitivity of selecting individuals for screening could be improved for a given specificity (Table 6). Applying the NLST criteria to the whole UK Biobank cohort would yield a specificity of 95% with a sensitivity of 37%. Setting the probability cut point to 0.354% to obtain the same specificity of 95% yields a sensitivity of 47% for the model-based strategy.

Similarly, when we consider only those participants with a history of smoking, the NLST criteria yield a specificity of 83% and a sensitivity of 50%. The model-based strategy would improve the sensitivity to 64% for the same specificity. The distribution of predicted probabilities by NLST eligibility and lung cancer case status at 2-year follow-up is plotted in Figure 3. A vertical line is drawn at 0.354%, which would yield the same specificity as the

Table 4. Hazard Ratios and 95% CIs for the Lung Cancer Risk Models Excluding Forced Expiratory Volume in 1 Second

Predictor	Never Smokers		Former Smokers		Current Smokers	
	HR	95% CI	HR	95% CI	HR	95% CI
Male (*v* female)	1.17	0.83 to 1.64	0.97	0.74 to 1.27	1.07	0.80 to 1.45
Previous cancer diagnosis	2.37	1.43 to 3.90	2.23	1.53 to 3.24	1.41	0.81 to 2.44
Family history of lung cancer (first-degree relative)	0.82	0.48 to 1.40	1.60	1.18 to 2.17	1.51	1.08 to 2.13
History of emphysema/chronic bronchitis	3.30	1.22 to 8.94	2.29	1.49 to 3.53	2.36	1.49 to 3.74
History of hay fever, allergic rhinitis, or eczema	1.19	0.80 to 1.77	0.82	0.56 to 1.20	0.51	0.29 to 0.89
Age stopped smoking, years			1.06	1.05 to 1.08		
First spline basis of cigarettes per day (knots: 4, 15, 30)			1.11	1.04 to 1.17	1.03	0.98 to 1.09
Second spline basis of cigarettes per day (knots: 4, 15, 30)			1.00	1.00 to 1.00	1.00	1.00 to 1.00
Difficulty of not smoking for 1 day (easy *v* difficult)					0.46	0.28 to 0.74

Abbreviation: HR, hazard ratio.

 © 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

Table 5. Hazard Ratios and 95% CIs for the Lung Cancer Risk Models Fitted to Multiply Imputed Data

Predictor	Never Smokers		Former Smokers		Current Smokers	
	HR	95% CI	HR	95% CI	HR	95% CI
Male (*v* female)	1.82	1.22 to 2.70	1.69	1.28 to 2.23	1.42	1.04 to 1.94
Previous cancer diagnosis	2.51	1.61 to 3.93	2.06	1.50 to 2.81	1.56	1.00 to 2.45
Family history of lung cancer (first-degree relative)	0.73	0.44 to 1.23	1.67	1.30 to 2.15	1.41	1.05 to 1.90
History of emphysema and/or chronic bronchitis	2.25	0.91 to 5.55	1.43	0.99 to 2.05	1.72	1.17 to 2.54
History of hay fever, allergic rhinitis, or eczema	1.15	0.79 to 1.66	0.84	0.61 to 1.16	0.59	0.38 to 0.92
Maximum FEV1, liters	0.58	0.43 to 0.78	0.55	0.45 to 0.68	0.63	0.50 to 0.80
Age stopped smoking, years			1.06	1.05 to 1.07		
First spline basis of cigarettes per day (knots: 4, 15, 30)			1.09	1.04 to 1.15	1.02	0.97 to 1.06
Second spline basis of cigarettes per day (knots: 4, 15, 30)			1.00	1.00 to 1.00	1.00	1.00 to 1.00
Difficulty of not smoking for 1 day (easy *v* difficult)					0.49	0.32 to 0.75

Abbreviations: FEV1, forced expiratory volume in 1 second; HR, hazard ratio.

NLST criteria when applied to these data. Thirty-eight percent of current smokers and 10% of former smokers have predicted risks exceeding this threshold, but only 87 ($< 0.1\%$) of never smokers in our sample exceed the threshold.

DISCUSSION

We found that a model designed solely on the basis of demographic data, family and personal medical history, and smoking behavior information strongly discriminated between incident lung cancer cases and noncases. The addition of FEV1 as a measure of lung function provided a modest increase in discriminatory power (c-statistic of the final model = 0.85). When the model was applied only to ever smokers, the discriminative ability remained high (c-statistic = 0.82). Internal validation of model discrimination indicated that the model is likely to perform well when applied to new individuals.

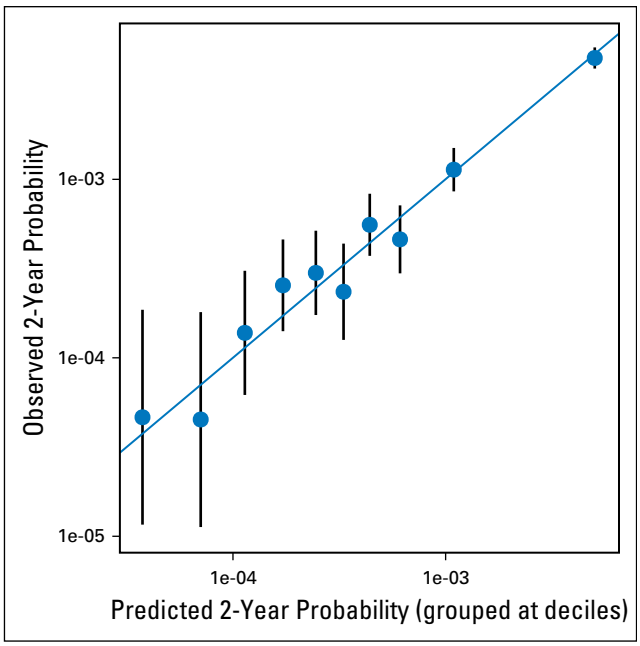


Fig 1. Calibration of the prediction model—the observed 2-year probability of lung cancer (Kaplan-Meier estimates and 95% CIs by tenths of model-predicted probability.

Strengths and Weaknesses

Our study has several strengths. To our knowledge, this is the first study to directly assess the ability of lung function, as measured by spirometry, to predict lung cancer in a population-based cohort. The prospective nature of the UK Biobank cohort also offers several advantages. The possibility of differential reporting biases is minimal, because all participants completed the baseline assessment before any diagnosis of lung cancer. Furthermore, we were able to directly model lung cancer incidence rates and overall mortality rates within one of the largest cohorts worldwide in a continuous fashion.

In addition, because the model only requires information that is readily available to general practitioners, it would be straightforward to implement on a wide scale in a primary care setting. This is of crucial importance if individual risk assessment becomes an integral part of any population screening programs for lung cancer. One potential practical limitation to implementing our model in practice is the requirement for the assessment of FEV1. Spirometry can be performed by physicians or nurses at most general practice surgeries in the United Kingdom, as well as local specialist clinics and hospitals.[36] The use of spirometry is routine in the diagnosis and management of chronic obstructive pulmonary disease, for which its use is encoded in the National Institute for Health and Care Excellence guidelines.[37] Thus although the addition of FEV1 adds some complexity to the risk assessment, the added burden is minimal.

The principal weakness of our study is that we were unable to externally validate the risk prediction model. The results of the internal validation provide promising evidence that the model will perform well when applied to new individuals. However, in practice, performance may vary when applied to new data, especially in populations with different average risks. Also, the relatively short follow-up to date of some participants in the UK Biobank means that we cannot evaluate estimates of absolute risk over a long-term follow-up.

Existing Lung Cancer Risk Prediction Models

Several lung cancer risk prediction models have been developed in various settings (described in a recent review by Tammemägi[38]). The Spitz et al[10] and LLP[9] models were developed using data from case-control studies combined with population

© 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

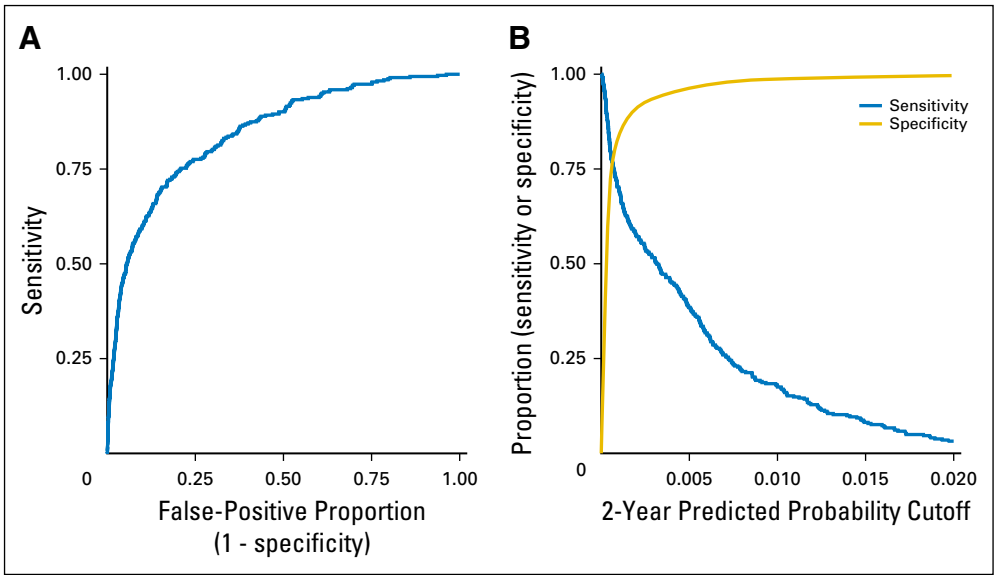


Fig 2. Discrimination of the prediction model. (A) Receiver operating characteristic curve (concordance [c]-statistic of fitted model = 0.85 [0.82 to 0.87]; optimism-corrected c-statistic = 0.84). (B) Sensitivity and specificity as a function of the predicted probability cutoff.

cancer incidence and overall mortality statistics. Similar to our model, those described by Bach et al,[8] Kovalchik et al,[7] Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)[12] (as well as the updated $PLCO_{m2012}$[15] and $PLCO_{all2014}$[39]), European Prospective Investigation into Cancer and Nutrition,[14] Park et al,[16] and Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence (LLPi),[17] were developed using data from prospective studies, either cohorts or randomized trials. The Bach et al, Kovalchik et al, $PLCO_{m2012}$, and European Prospective Investigation into Cancer and Nutrition models were developed using data from participants who had ever smoked, whereas the $PLCO_{all2014}$, Park et al, and LLPi models also included data from participants who had never smoked.

To our knowledge, the only previous study to assess lung function directly was conducted in British Columbia, Canada, and included 2,596 "high-risk" individuals, defined as having a smoking history of at least 20 pack-years.[13] This study also found that the inclusion of lung function improved model discrimination (AUC 0.767 [0.725 to 0.809]) compared with a base model (AUC 0.718 [0.671 to 0.765]). It is difficult to directly compare the discriminative ability of models developed using data from different target populations with varying baseline risks. For instance, it is difficult to compare statistics from populations with differing age distributions, or from a population of ever smokers versus a population that also includes people who have never smoked.

Table 6. Comparison of Sensitivities of the National Lung Screening Trial Criteria and the Model-Based Strategy for Identifying Individuals at High Risk in the UK Biobank Cohort

	Specificity (%)*		Sensitivity (%)	
Category	NLST	Model	NLST	Model
Current and former smokers	83	83	50	64
All participants	95	95	37	47

Abbreviation: NLST, National Lung Screening Trial.
*For the model-based strategy, the probability cut point was selected to obtain the same specificity as the NLST criteria.

Nevertheless, each of the models developed using data from prospective studies has similar discriminative ability, with apparent c-statistics between 0.72 and 0.86. Models such as the Park et al, $PLCO_{all2014}$, and LLPi, which were applied to samples including never smokers, tend to have higher discriminative power as a result of the predictably low lung cancer incidence among never smokers. The predictive power of our model was similarly high when applied to all participants, and remained substantial when applied only to participants who had ever smoked (c-statistic = 0.82).

Potential for Targeted Screening

There is great potential for a risk prediction model such as ours to improve the performance of population screening programs for lung cancer. The NLST demonstrated that CT screening could reduce lung cancer death rates by 20% in a population of current or former heavy smokers.[4] It was subsequently shown by Kovalchik et al[7] that the 20% of participants at lowest predicted risk experienced essentially no benefit from screening, accounting for only 1% of the prevented deaths from lung cancer. In a recent paper, Katki et al[40] applied an updated version of the Kovalchik et al model to NLST and PLCO data. They estimated that a model-based inclusion criterion would lead to a 17% reduction in the number of people needed to screen to prevent one death. Similarly, Tammemägi et al compared an inclusion criterion on the basis of the $PLCO_{m2012}$ model with the US Preventive Services Taskforce guidelines (similar to the NLST criteria). When these criteria were applied to ever-smokers in the PLCO intervention arm, the model-based approach led to improved sensitivity (80% *v* 71%) and specificity (66% *v* 63%).[39]

Screening trials in Europe have typically used similar inclusion criteria to those of the NLST, the exception being the United Kingdom Lung Cancer Screening trial, which has used the LLPi risk prediction model to identify those at high risk before random assignment.[6] We envisage that our model, with the inclusion of an assessment of lung function, may further improve such targeted screening approaches. This could particularly be the case if FEV1

 © 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

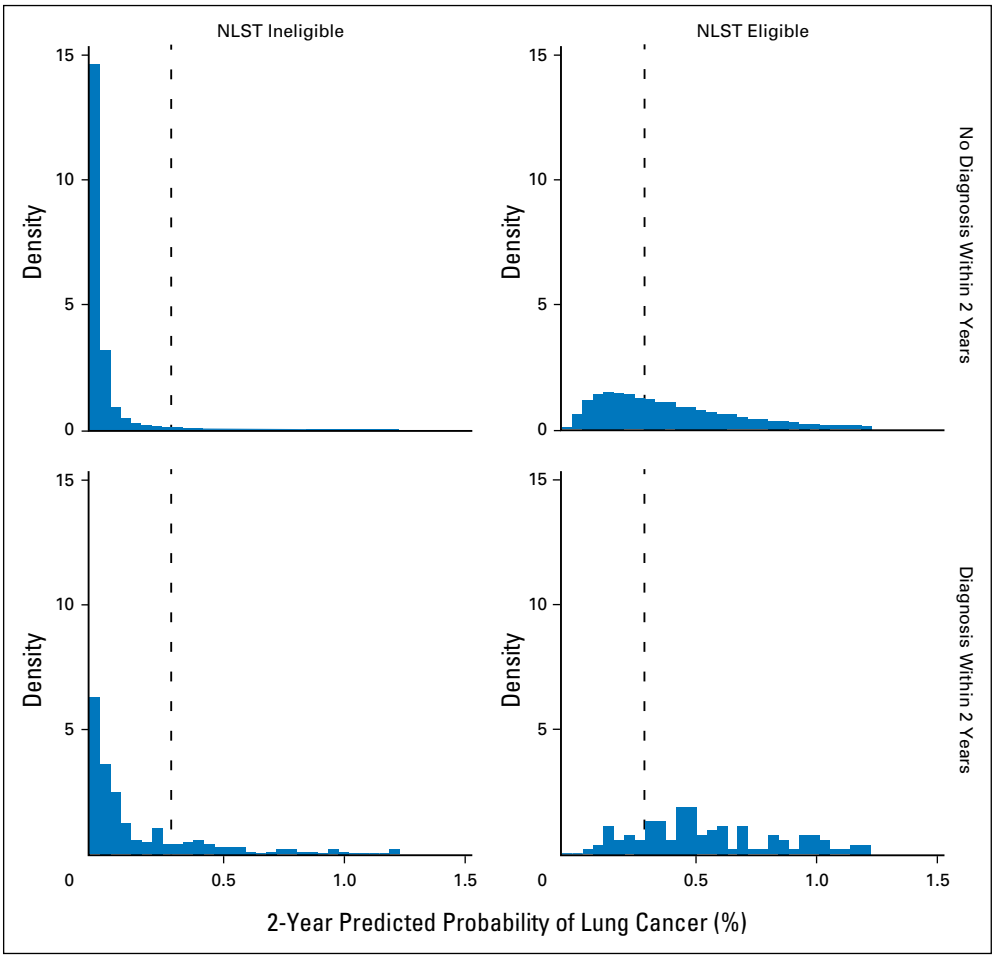


Fig 3. Distribution of predicted probabilities by National Lung Screening Trial (NLST) eligibility. The dashed line indicates a probability cutoff of 0.354%, which yields specificity of 95% to match the specificity of the NLST criteria when applied to these data.

can provide additional power to detect never smokers at sufficiently high risk to warrant screening.

Future Research in Lung Cancer Risk Prediction

We have demonstrated that a relatively simple model, designed on the basis of readily available individual characteristics and a measure of lung function, is capable of discriminating well between those at high and low risk of lung cancer. However, there remains substantial room for improvement. One potentially fruitful avenue would be to incorporate genetic and molecular predictors of lung cancer in the model. Genome-wide association studies have identified multiple risk loci for lung cancer.[41-43] Although taken individually, these genetic variants are typically weakly associated with risk, inclusion of a panel of multiple variants might improve predictive power.

Circulating microRNAs have also been identified as potential markers of lung cancer risk, but these have not yet been evaluated in large prospective studies.[44,45] Circulating metabolites of nicotine, such as cotinine, are likely to improve model discrimination among current smokers. Circulating cotinine is strongly associated with lung cancer risk,[46] and could be used as a more sensitive assessment of smoking intensity than self-reported number of cigarettes smoked per day. Further, it has recently been shown that both prosurfactant protein B[47] and diacetylspermine[48] can provide additional predictive power. Comprehensive models that incorporate demographic data and smoking information, lung function, and molecular risk markers thus have great potential to improve lung cancer risk stratification. However, any potential gains from the inclusion of molecular markers must be carefully weighed against the increased cost and inconvenience of conducting more invasive blood tests on a routine basis.

In summary, we developed and internally validated a lung cancer risk prediction model incorporating a direct assessment of lung function via spirometry, using data from a large prospective cohort of UK residents. The model was capable of discriminating well between those at high and low risk of lung cancer, and the model-predicted probabilities were well calibrated. Using this model to establish eligibility for screening could lead to substantially greater sensitivity than the NLST screening criteria in the UK population. The model is based on covariates that are readily available or easily ascertained in a primary care setting. Our model could thus be easily implemented as an aid to physicians and patients who may be considering undergoing screening for lung cancer.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at jco.org.

© 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Administrative support: Mattias Johansson
Collection and assembly of data: David C. Muller
Data analysis and interpretation: David C. Muller
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, et al: Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136:E359-E386, 2015
2. Cancer Research UK: Cancer statistics for the UK. http://www.cancerresearchuk.org/cancer-info/cancerstats/
3. Cancer Research UK: Lung cancer survival statistics. http://www.cancerresearchuk.org/cancer-info/cancerstats/types/lung/survival/#stage
4. Aberle DR, Adams AM, Berg CD, et al: Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395-409, 2011
5. Field JK, Duffy SW, Baldwin DR, et al: The UK Lung Cancer Screening Trial: A pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer. Health Technol Assess 20:1-146, 2016
6. Field JK, van Klaveren R, Pedersen JH, et al: European randomized lung cancer screening trials: Post NLST. J Surg Oncol 108:280-286, 2013
7. Kovalchik SA, Tammemägi M, Berg CD, et al: Targeting of low-dose CT screening according to the risk of lung-cancer death. N Engl J Med 369:245-254, 2013
8. Bach PB, Kattan MW, Thornquist MD, et al: Variations in lung cancer risk among smokers. J Natl Cancer Inst 95:470-478, 2003
9. Cassidy A, Myles JP, van Tongeren M, et al: The LLP risk model: An individual risk prediction model for lung cancer. Br J Cancer 98:270-276, 2008
10. Spitz MR, Hong WK, Amos CI, et al: A risk model for prediction of lung cancer. J Natl Cancer Inst 99:715-726, 2007
11. Spitz MR, Etzel CJ, Dong Q, et al: An expanded risk prediction model for lung cancer. Cancer Prev Res (Phila) 1:250-254, 2008 (Phila Pa)
12. Tammemägi CM, Pinsky PF, Caporaso NE, et al: Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation. J Natl Cancer Inst 103:1058-1068, 2011
13. Tammemägi MC, Lam SC, McWilliams AM, et al: Incremental value of pulmonary function and sputum DNA image cytometry in lung cancer risk prediction. Cancer Prev Res (Phila) 4:552-561, 2011 (Phila Pa)
14. Hoggart C, Brennan P, Tjonneland A, et al: A risk model for lung cancer incidence. Cancer Prev Res (Phila) 5:834-846, 2012 (Phila Pa)
15. Tammemägi MC, Katki HA, Hocking WG, et al: Selection criteria for lung-cancer screening. N Engl J Med 368:728-736, 2013
16. Park S, Nam B-H, Yang H-R, et al: Individualized risk prediction model for lung cancer in Korean men. PLoS One 8:e54823, 2013
17. Marcus MW, Chen Y, Raji OY, et al: LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence. Cancer Prev Res (Phila) 8:570-575, 2015 (Phila Pa)
18. Ebbert JO, Yang P, Vachon CM, et al: Lung cancer risk reduction after smoking cessation: Observations from a prospective cohort of women. J Clin Oncol 21:921-926, 2003
19. Crispo A, Brennan P, Jöckel K-H, et al: The cumulative risk of lung cancer among current, ex- and never-smokers in European men. Br J Cancer 91:1280-1286, 2004
20. Brennan P, Crispo A, Zaridze D, et al: High cumulative risk of lung cancer death among smokers and nonsmokers in Central and Eastern Europe. Am J Epidemiol 164:1233-1241, 2006
21. Agudo A, Bonet C, Travier N, et al: Impact of cigarette smoking on cancer risk in the European prospective investigation into cancer and nutrition study. J Clin Oncol 30:4550-4557, 2012
22. Peto R, Darby S, Deo H, et al: Smoking, smoking cessation, and lung cancer in the UK since 1950: Combination of national statistics with two case-control studies. BMJ 321:323-329, 2000
23. Pinsky PF, Zhu CS, Kramer BS: Lung cancer risk by years since quitting in 30+ pack year smokers. J Med Screen 22:151-157, 2015
24. Gu F, Wacholder S, Kovalchik S, et al: Time to smoke first morning cigarette and lung cancer in a case-control study. J Natl Cancer Inst 106:dju118, 2014
25. Denholm R, Schüz J, Straif K, et al: Is previous respiratory disease a risk factor for lung cancer? Am J Respir Crit Care Med 190:549-559, 2014
26. Hemminki K, Försti A, Fallah M, et al: Risk of cancer in patients with medically diagnosed hay fever or allergic rhinitis. Int J Cancer 135:2397-2403, 2014
27. El-Zein M, Parent M-E, Siemiatycki J, et al: History of allergic diseases and lung cancer risk. Ann Allergy Asthma Immunol 112:230-236, 2014
28. Samet JM, Humble CG, Pathak DR: Personal and family history of respiratory disease and lung cancer risk. Am Rev Respir Dis 134:466-470, 1986
29. Wu AH, Yu MC, Thomas DC, et al: Personal and family history of lung disease as risk factors for adenocarcinoma of the lung. Cancer Res 48:7279-7284, 1988
30. Wu AH, Fontham ET, Reynolds P, et al: Family history of cancer and risk of lung cancer among lifetime nonsmoking women in the United States. Am J Epidemiol 143:535-542, 1996
31. Osann KE: Lung cancer in women: The importance of smoking, family history of cancer, and medical history of respiratory disease. Cancer Res 51:4893-4897, 1991
32. Nitadori J, Inoue M, Iwasaki M, et al: Association between lung cancer incidence and family history of lung cancer: Data from a large-scale population-based cohort study, the JPHC study. Chest 130:968-975, 2006
33. Royston P, Parmar MKB: Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med 21:2175-2197, 2002
34. Harrell FE Jr, Lee KL, Mark DB: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361-387, 1996
35. The R Foundation: The R Project for Statistical Computing. http://www.R-project.org/
36. NHS Choices: Spirometry. http://www.nhs.uk/Conditions/spirometry/Pages/Introduction.aspx
37. National Institute for Health and Care Excellence: Chronic obstructive pulmonary disease in over 16s: Diagnosis and management. https://www.nice.org.uk/guidance/cg101
38. Tammemägi MC: Application of risk prediction models to lung cancer screening: A review. J Thorac Imaging 30:88-100, 2015
39. Tammemägi MC, Church TR, Hocking WG, et al: Evaluation of the lung cancer risks at which to screen ever- and never-smokers: Screening rules applied to the PLCO and NLST cohorts. PLoS Med 11:e1001764, 2014
40. Katki HA, Kovalchik SA, Berg CD, et al: Development and validation of risk models to select ever-smokers for CT lung cancer screening. JAMA 315:2300-2311, 2016
41. Hung RJ, McKay JD, Gaborieau V, et al: A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452:633-637, 2008
42. 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 40:616-622, 2008
43. Brennan P, Hainaut P, Boffetta P: Genetics of lung-cancer susceptibility. Lancet Oncol 12:399-408, 2011
44. Montani F, Marzi MJ, Dezi F, et al: miR-Test: A blood test for lung cancer early detection. J Natl Cancer Inst 107:djv063, 2015
45. Wozniak MB, Scelo G, Muller DC, et al: Circulating microRNAs as non-invasive biomarkers for early detection of non-small-cell lung cancer. PLoS One 10:e0125026, 2015
46. Timofeeva MN, McKay JD, Smith GD, et al: Genetic polymorphisms in 15q25 and 19q13 loci, cotinine levels, and risk of lung cancer in EPIC. Cancer Epidemiol Biomarkers Prev 20:2250-2261, 2011
47. Sin DD, Tammemägi CM, Lam S, et al: Pro-surfactant protein B as a biomarker for lung cancer prediction. J Clin Oncol 31:4536-4543, 2013
48. Wikoff WR, Hanash S, DeFelice B, et al: Diacetylspermine is a novel prediagnostic serum biomarker for non-small-cell lung cancer and has additive performance with pro-surfactant protein B. J Clin Oncol 33:3880-3886, 2015

Affiliations

David C. Muller, **Mattias Johansson**, and **Paul Brennan**, International Agency for Research on Cancer, Lyon, France; **David C. Muller**, Imperial College London, London, United Kingdom.

 © 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

Support

UK Biobank was established by the Wellcome Trust medical charity, the Medical Research Council, the Department of Health, the Scottish Government, and the Northwest Regional Development Agency. UK Biobank has also had funding from the Welsh Assembly Government and the British Heart Foundation. Data collection was funded by UK Biobank. D.C.M. was funded by an International Agency for Research on Cancer–Australia Fellowship, supported by Cancer Council Australia.

Prior Presentation

Presented in part at the Annual Meeting of the American Association for Cancer Research, Philadelphia, PA, April 18-22, 2015.

Oncology Research Just Got Easier!



Visit the new ASCO Journals website today and take advantage of:

- Better Navigation and Readability
- Improved Search
- More Relevant Content—powered by TrendMD
- Altmetric Badges

Visit the new ASCO Journals website today at **ascopubs.org**.

© 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

David C. Muller
No relationship to disclose

Mattias Johansson
No relationship to disclose

Paul Brennan
No relationship to disclose

© 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.

Appendix

The 2-year absolute risk of lung cancer was estimated by the cumulative incidence function (CIF), conditional on age at baseline attendance t_0:

$$CIF(t_0 + 2|t_0) = \int_{t_0}^{t_0+2} h_1(u)\exp\left[-\int_{t_0}^{u} h_1(v) + h_2(v)dv\right]du$$

where $h_1(t)$ is the model-estimated hazard function for lung cancer, and $h_2(t)$ is the model-estimated hazard function for death.

An easy-to-use Stata code for calculating the model-predicted probability of lung cancer is available from GitHub, Inc. (https://github.com/dcmuller/ukbiobank_lca_model_predictions).

© 2017 by American Society of Clinical Oncology

Downloaded from ascopubs.org by Imperial College London on June 1, 2017 from 155.198.012.147
Copyright © 2017 American Society of Clinical Oncology. All rights reserved.