LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence

Michael W. Marcus [1], Ying Chen [1], Olaide Y. Raji [1], Stephen W. Duffy [2], John K. Field [1]

[1] Roy Castle Lung Cancer Research Programme, the University of Liverpool Cancer Research Centre, Institute of Translational Medicine, the University of Liverpool. Liverpool L3 9TA, UK.

[2] Wolfson Institute of Preventive Medicine, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ

Corresponding author:

Dr. Michael W. Marcus
Roy Castle Lung Cancer Research Programme,
The University of Liverpool Cancer Research Centre
Department of Molecular and Clinical Cancer Medicine,
Institute of Translational Medicine,
The University of Liverpool
200 London Road
Liverpool L3 9TA
UK
Phone: +44 151 794 8956, Fax: +44 151 794 8989, email: m.w.marcus@liv.ac.uk

Running Title: The LLPi Risk model

Keywords: Lung cancer, population-based cohort, risk model

Financial Support: The Liverpool Lung project was principally funded by the Roy Castle Lung Cancer Foundation, UK. Michael W. Marcus is funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no.HEALTH-F2-2010-258677 (CURELUNG project) and grant agreement no.258868 (LCAOS project).

COI: The authors declare that they have no conflicts of interest.

Abstract

Identification of high risk individuals will facilitate early diagnosis, reduce overall costs and also improve the current poor survival from lung cancer. The Liverpool Lung Project prospective cohort of 8760 participants aged 45-79 years, recruited between 1998 and 2008, were followed annually through the hospital episode statistics until 31 January 2013. Cox proportional hazards models were used to identify risk predictors of lung cancer incidence. C-statistic was used to assess the discriminatory accuracy of the models. Models were internally validated using bootstrap method. During mean follow-up of 8.7 years, 237 participants developed lung cancer. Age (hazard ratio [HR] 1.04; 95%CI, 1.02-1.06), male gender (HR 1.48; 95%CI:1.10-1.98), smoking duration (HR 1.04; 95%CI 1.03-1.05), COPD (HR 2.43; 95%CI 1.79-3.30), prior diagnosis of malignant tumour (HR 2.84; 95%CI 2.08-3.89)and early onset of family history of lung cancer (HR 1.68;95%CI 1.04-2.72) were associated with the incidence of lung cancer. The LLPi risk model had a good calibration (goodness-of-fit χ^2 7.58, *P*=0.371). The apparent C-statistic was 0.852(95%CI 0.831-0.873) and the optimism-corrected bootstrap resampling C-statistic was 0.849 (95% CI of 0.829–0.873). The LLPi risk model may assist in identifying individuals at high risk of developing lung cancer in population-based screening programmes.

Introduction

Lung cancer is the leading cause of cancer-related death worldwide with mortality rate exceeding that of prostate, breast and colon cancer combined (1). It has one of the poorest survival outcomes of any cancer because a vast majority of patients are diagnosed at an advanced stage when surgical resection or other treatment options are less effective (2). The National Lung Screening Trial (NLST) showed that lung cancer screening with low-dose computed tomography (LDCT) reduced lung cancer mortality by 20% (3). In practice, the success of any lung cancer screening program will depend on successful identification of individuals at high risk. Risk prediction models have been recognised as a method of identifying individuals at high risk of developing lung cancer (4, 5). Identification of individuals at high risk will facilitate early diagnosis, reduce overall costs and also improve the current poor survival from lung cancer.

To date, most risk prediction models for lung cancer were developed in case control studies (6-9). Case-control studies are proficient in studying dynamic populations where follow-up is difficult, are usually less expensive and are less time consuming, but may be plagued by biases and cannot study the incidence of a disease (10-13). Cohort studies though time consuming and expensive, offer the methodological advantage of direct calculation of incidence rate and can demonstrate the temporal sequence between exposure and outcome (14, 15).

We previously developed the LLP risk model from the LLP case-control study and validated it in three independent populations (16). Covariates such as prior history of respiratory diseases and prior diagnosis of malignant diseases have been reported as risk factors for lung cancer (16, 17). Clinical covariates in the previous case-control model were based on un-validated self-reported questionnaire responses (18). However, in the newly developed model reported in this article, clinical covariates were confirmed in the Hospital Episode Statistics (HES) database. HES is the national statistical data warehouse for the National Health Service (NHS) that includes clinical and administrative information

about the care provided to NHS patients who live, or are treated in England (19). The aim of the current study was to use baseline data from the LLP population-based cohort to develop and validate a risk prediction model for lung cancer incident.

Materials and Methods

Study Population

This study was performed as part of the Liverpool Lung Project. The objectives, methods, rationale, and study design have been described previously (18). In short, 8760 randomly selected healthy subjects aged 45-79 years were recruited between 1998 and 2008 and followed annually for lung cancer and mortality outcomes through the Office for National Statistics (ONS), public health England (the North West Cancer Intelligence Service), and hospital case-note review until 31 January 2013.

Data Collection and Extraction of Risk Factors

A standardised questionnaire was used to collect self-reported information on demographic and socioeconomic economic characteristics, medical history, family history of cancer, history of tobacco consumption and lifetime occupational history. Information on age, gender, smoking duration, marital status, education level, family history of lung cancer, prior history of other cancers and prior history of non-malignant lung disease such as asthma, chronic obstructive pulmonary disease (COPD), pneumonia, bronchitis, emphysema, tuberculosis and exposure to asbestos was extracted from the questionnaire. Smoking duration was measured in years; ever smoker was defined as someone who had smoked at least 100 cigarettes in their lifetime and a current smoker was defined as a participant who reported smoking within 12 months of the date of the interview. Marital status was reported as single, married, living together, widowed, divorced/separated or other. Education was classified as high school and below and greater than high school and recorded as “yes” or “no”. Family history of

lung cancer included age at onset in a first-degree relative (none, early [<6o years], or late [>60 years]). Previous history of cancer (except melanoma) was coded as "yes" or "no". Information on prior history of non-malignant lung diseases such as asthma, chronic obstructive pulmonary disease (COPD), pneumonia and tuberculosis were coded "yes" or "no" and diagnosis was corroborated in the HES database prior to their recruitment into the study. Asbestos exposure was determined based on the documentation of participant's employment history in asbestos-related occupation or industry and was coded "yes" or "no". The study protocol was approved by the Liverpool Research Ethic Committee and all research participants provided written, informed consent in accordance with the Declaration of Helsinki.

Statistical analyses

Descriptive statistics were obtained and compared by using Chi-square test or Fisher's exact test for categorical variables and t-tests for normally distributed variables, between participants who developed lung cancer(n=237) and those who did not (n=8523) at the end of the follow-up period. Cox proportional hazards model was used to develop a multivariable model for lung cancer risk and to compute hazard ratios (HR) and corresponding confidence intervals (CI) after testing the proportional hazards assumption with the un-scaled and scaled Schoenfeld residuals (20). Follow-up duration was used as the time axis in the model. For participants that developed lung cancer, follow-up time started from baseline visit, till the date of diagnosis of lung cancer. For participants without incident lung cancer, the follow-up duration was counted from the baseline visit till death, withdrawal, loss to follow-up or 31 January 2013. Five percent of the 8760 participants had missing data; we therefore performed complete case analysis because the characteristics of participants with missing data were largely similar to those of participants with complete observed data. Martingale residuals were used to determine the best function of all covariates (21). The multivariable model was built in two phases.

First, all covariates with $P \leq 0.10$ in the univariate analyses were considered for inclusion in the multivariable model. Second, backward selection procedure with ($P<0.05$) was used to choose the covariates in the final multivariable model (22). Covariates eliminated were re-entered in the final multivariable model with adjustment for the remaining significant covariates to ensure that no omitted covariate significantly reduced the log likelihood chi-square of the model (23). The performance of the multivariable model was quantified by Harrell's concordance (C) statistics which is analogous to the receiver operating characteristics (ROC) curve for binary data with confidence interval estimates using jacknife resampling method (24). A C-statistic can be interpreted as the probability that the model predicts a higher risk of lung cancer for those who actually developed lung cancer compared with those that did not develop lung cancer over the follow-up time (25). Model calibration was evaluated using the omnibus Grønnesby-Borgan goodness-of-fit test. Bootstrapping techniques were used for internal validation of the model (26, 27) and bootstrap samples were drawn 200 times with replacement. Regression models were created in each bootstrap sample and tested on the original sample to obtain stable estimates of the optimism of the model, i.e. how much the model performance was expected to decrease when applied in new datasets (28-30). All analyses were performed using STATA® version 13.1 (StataCorp LP, College Station, Texas) and SAS® statistical software version 9.3 (SAS Institute, Cary, North Carolina).

Results

Overall, 8760 participants aged 45-79 recruited between 1998 and 2008 were included in this study and 237 participants (2.7%) developed lung cancer during a mean follow-up of 8.7 years. Table 1 depicts the baseline characteristics of the study population stratified by incident lung cancer status. Lung cancer incidence was higher in older participants, men and also in participants with longer smoking duration, participants with prior history of pneumonia, asthma, tuberculosis, COPD,

emphysema and bronchitis. Furthermore, the percentage of lung cancer incidence was higher in participants with; family history of lung cancer and prior diagnosis of malignant tumour.

In univariate analysis, age, male gender, smoking duration, prior diagnosis of pneumonia, asthma, COPD, emphysema, bronchitis, prior diagnosis of malignant tumour and family history of lung cancer were significantly associated with the risk of developing lung cancer. In addition, we also performed analyses that adjusted all potential confounders in Table 1 for smoking duration. There was a significant increase in risk with age both before (HR=1.08, 1.06-1.09) and after adjusting for smoking duration (HR=1.05, 95% CI 1.03-1.07). Male gender was significantly associated with the incidence of lung cancer both before (HR 1.55, 95%CI 1.20-2.01) and after (HR=1.51, 95%CI 1.16-1.97) adjusting for smoking. Participants with a prior history of asthma had a significant increase in risk before adjustment (HR=1.53, 95%CI 1.12-2.09) but not after (HR=1.25, 0.91-1.71). Participants with a prior history of COPD had a significant increase in risk both before (HR=5.13, 95%CI 3.85-6.84) and after adjustment (HR=2.75, 2.03-3.72). Prior diagnosis of malignant tumour increased risk before (HR=4.18, 95%CI 3.15-5.55) and after adjustment (HR=3.09, 95%CI 2.33-4.11). Family history of lung cancer with early onset (<60 years) increased risk with marginal significantly before (HR=1.56, 95%CI 0.98-2.48) but not after adjustment (HR=1.38, 95%CI 0.87-2.20) for smoking. Education, a measure of socio-economic status was protective against lung cancer before (HR=0.58, 95%CI 0.43-.0.79) for low education (high school and below) and (HR=0.28, 95%CI 0.16-.0.49) for higher education (greater than high school) while only higher education was protective against lung cancer after adjusting for smoking (HR=0.53, 95%CI 0.30-0.95). No significant effect was observed on marital status, occupational exposure to asbestos, pneumonia, bronchitis, emphysema and tuberculosis on lung cancer risk after adjustment for smoking duration. Ever smokers (HR=24.1, 95%CI 10.74-54.3) were at higher risk than never smokers. Fitting smoking duration as a continuous covariate and in 10- and 20 years interval revealed a steady increase in lung cancer risk. There was also a steady increase in risk with increasing smoking pack-years and the average amount smoked, although in neither case was it

as large as that with smoking duration. A significant dose-response effect was also observed for the amount of cigarettes (*P*<0.001), smoking duration (*P*<0.001) and smoking pack-years (*P*<0.001).

Table 2 presents the final multivariate Cox-regression model. Age (HR= 1.04, 95% CI 1.02-1.06), male gender (HR=1.48, 95%CI 1.10-1.98), Smoking duration (HR= 1.04, 95% CI 1.03-1.05), COPD (HR= 2.43, 95%CI 1.79-3.30), prior diagnosis of malignant tumour (HR=2.84, 95%CI 2.07-3.89) and family history of early onset of lung cancer (HR=1.68, 95%CI 1.04-2.72) significantly increased the risk of developing lung cancer. The LLPi model for incident lung cancer had good discrimination C-statistic 0.852(95%CI 0.832-0.873) (Figure 1) and 0.849 (95% CI of 0.829–0.870) by internal validation with bootstrap re-sampling and correction for optimism. The Grønnesby-Borgan goodness-of-fit test demonstrated overall good calibration χ^2 7.58, *P*=0.371. A standard test of the proportional hazard (PH) assumption for the model for each covariate using scaled Schoenfeld residuals showed that the PH was not violated.

Using the equation under Table 2, the absolute risk of lung cancer within a mean follow-up period of 8.7 years was calculated. The LLPi risk model was used to estimate the absolute risk of developing lung cancer for three hypothetical individuals with diverse risk profile. First, the absolute risk of a woman aged 65 with 37 years smoking history and a history of COPD diagnosis with a late onset family history of lung cancer (aged >60 at diagnosis) and no other risk factors is calculated as $\hat{P} = 1 - S_0(t)^{\exp\left(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i\right)}$ = 9.9% (see details of the formula under Table 2). Second, the absolute risk of a man aged 67 without a smoking history but with an early onset family history of lung cancer (aged <60 at diagnosis) and a prior diagnosis of cancer is 6%. Third, the absolute risk of a man aged 73 who has smoked for 59 years and also had an early onset (aged <60 at diagnosis) family history of lung cancer is 29%.

Discussions

In this study, we developed and internally validated the LLPi risk model for incident lung cancer in the LLP population cohort. Age, male gender, smoking duration, prior history of COPD, prior diagnosis of malignant tumour and family history of early onset of lung cancer (< 60 years) were significant risk factors. The C-statistic of the LLPi risk model and the bias corrected bootstrap resampling were very similar: 0.852 (95%CI 0.832-0.873) and 0.849 (95% CI of 0.829–0.870) respectively. The average difference known as the degree of optimism was 0.003 which indicates that the LLPi can discriminate well between patients with lung cancer and population controls. A model such as the LLPi with a high discrimination and good calibration is expedient for counselling and population-based screening programmes.

To date, several risk prediction models have been developed in population-based cohort data with variable discrimination presented as ROC AUC or C-statistic (Table 3). Bach et al. used prospective cohort data for smokers in the Carotene and Retinol Efficacy Trial (CARET) to develop their model that included age, gender, smoking history and exposure to asbestos as predictors (31). The model was externally validated in the alpha-tocopherol, beta-carotene cancer prevention (ATBC) study control arm with a C-statistic of 0.69 (32,33). The Prostate, Lung, Colorectal and Ovarian cancer screening (PLCO) used prospective data to build two risk prediction models for the general population (model 1) and a sub-cohort of ever-smokers (model 2) (34). The bootstrap optimism corrected ROC AUC were 0.857 and 0.805 respectively. External validation of the models was performed in the PLCO intervention arm with AUC ROC of 0.841 and 0.784 for models 1 and 2 respectively (34). Park et al., developed a risk model in a large population-based cohort study of Korean men with a C-statistic of 0.864 (95%CI 0.860-0.868) (34). External validation of their model produced a C-statistic of 0.871 (95%CI 0.867-0.876) (35). Hoggart et al. used prospective data from the European Prospective Investigation into Cancer (EPIC) to build a predictive model for lung cancer (36). Using smoking information alone, their model had an AUC ROC of 0.843 (95% CI 0.810-0.875). External validation of

the Bach model using the same data gave an AUC ROC of 0.775 (95%CI 0.737-0.813) (36). In another study, Tammemagi et al. modified the PLCO model 1 and 2 to ensure applicability to National Lung Screening Trial (NLST) (37). The newly developed model $PLCO_{M2012}$ utilised data from PLCO control groups of individuals who had ever smoked and validated the model in the PLCO intervention group of smokers, NLST participants, and in the PLCO intervention group stratified according to whether or not they met the NLST criteria. The AUC of their model was 0.803 in the developmental dataset and 0.797 in the validation dataset (37).

Although the LLPi had a high C-statistic with magnitude comparable to the aforementioned models, its interpretation is not directly comparable with these models because the differences in distributions of covariates can affect performance statistics (38). However, the LLPi model will be more directly applicable for use in primary care setting because it included readily available, strong and plausible covariates that have been implicated in the aetiology of lung cancer from our own and numerous other case-control and cohort studies.

Previous lung diseases including COPD (emphysema, bronchitis), tuberculosis, pneumonia (39) and asthma (40) have been reported as risk factors for lung cancer. In our study, we also examined the association between COPD, emphysema, bronchitis, pneumonia, tuberculosis and asthma on lung cancer. COPD was the only significant covariate in our final multivariable model. The association between COPD and lung cancer we found in our study was consistent with the result from two earlier meta-analyses (17, 39). In contrast, emphysema, bronchitis, pneumonia, asthma and tuberculosis were not significantly associated with lung cancer in our final multivariate model. A plausible explanation for this observation is the small numbers of participants with these respiratory conditions in our study compared to the pooled analysis of 16, 17 and 39 studies for asthma, and other respiratory diseases in the aforementioned meta-analysis respectively. In addition, occupational exposure to asbestos was a risk factor for lung cancer in the LLP risk model (8). However, in this present study, occupational exposure to asbestos was not a risk factor for lung cancer. This

observation may be attributed to the fact that only one of the 237 individuals that developed lung cancer was exposed to occupational asbestos. Twenty three participants had missing data on occupational exposure to asbestos and 213 participants that also developed lung cancer were not exposed to asbestos (Table 1). Thus occupational exposure to asbestos cannot be included in the model because our study population does not have adequate sample of such individuals. Another plausible explanation for this observation is reporting bias and/or misclassification as there is no means of validating the asbestos exposure other than as reported by participant. This is not the case with the medical conditions that may be validated using HES.

Strengths of our study include: its prospective design; the large number of participants; long follow-up period; the population-based setting and that detailed information on the main risk factors (such as smoking and family history of lung cancer) was ascertained by closely supervised trained interviewers, using standardised questionnaires. In addition, information bias was prevented by complete documentation of lung cancer incidence (by the Office for National Statistics (ONS), the North West Cancer Intelligence Service, and hospital case-note review) and comorbidity data that were corroborated using the HES database. Furthermore, unlike in case-control studies, we were able to easily compute the absolute risk estimates for each combination of risk factors from Cox regression.

Although the LLPi model demonstrated good discrimination and calibration, more work is needed to test the applicability of the model in diverse populations, including those from diverse geographic regions. Because of marked geographical differences in incidence rates, evaluation of the LLPi risk model in high and low risk areas is necessary. Advancement in high throughput methodologies and their application in molecular and genetic epidemiological studies have expanded the potential for 'omic'-based risk prediction (41). Genome-wide association studies have identified inherited susceptibility patterns for lung cancer at different loci (42-44) and several methylation (45-47) and microRNA biomarkers (48-51) associated with lung cancer have been identified. It is anticipated that many more DNA polymorphism and biomarkers will be developed. Due to fewer opportunities to

access biomaterials from large prospective cohort studies, most current biomarkers are used mainly for diagnosis, but their value in risk prediction have not been widely explored (41). We therefore recommend future studies to explore the contribution of genetic polymorphism, methylation and microRNA in combination with clinical and epidemiological factors on lung cancer risk.

In conclusion, we have developed and internally validated the LLPi risk model based on readily available, strong and plausible covariates that have been implicated in the aetiology of lung cancer from our own and numerous other case-control and cohort studies. The application of LLPi risk model in identifying individuals at high risk of developing lung cancer in population-based screening programmes needs further study.

References

1. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. *Clin Chest Med* 2011;32:605-44.
2. Qi WX, Wang Q, Jiang YL, Sun YJ, Tang LN, He AN, et al. Overall survival benefits for combining targeted therapy as second-line treatment for advanced non-small-cell-lung cancer: a meta-analsysis of published data. *PLoS One* 2013;8:e55637.
3. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med 2011*;365:395-409.
4. Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK. Lung cancer risk prediction: a tool for early detection. *Int J Cancer* 2007;120:1-6.
5. Field JK. Lung cancer risk models come of age. *Cancer Prev Res (Phila)* 2008;1:226-8.
6. Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, et al. A risk model for prediction of lung cancer. *J Natl Cancer Inst* 2007;99:715-26.
7. Etzel CJ, Kachroo S, Liu M, D'Amelio A, Dong Q, Cote ML, et al. Development and validation of a lung cancer risk prediction model for African-Americans. *Cancer Prev Res (Phila)* 2008;1:255-65.
8. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer* 2008;98:270-6.
9. Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, et al. An expanded risk prediction model for lung cancer. *Cancer Prev Res (Phila)* 2008;1:250-4.
10. Miettinen OS. Matching and design efficiency in retrospective studies. *Am J Epidemiol* 1970;91:111-8.
11. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Sci Publ* 1980:5-338.
12. Schulz KF, Grimes DA. Case-control studies: research in reverse. *Lancet* 2002;359:431-4.
13. Vandenbroucke JP, Pearce N. Case-control studies: basic concepts. *Int J Epidemiol* 2012;41:1480-9.
14. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. *Lancet* 2002;359:341-5.
15. Mann CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J* 2003;20:54-60.
16. Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, et al. Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. *Ann Intern Med* 2012;157:242-50.
17. Brenner DR, McLaughlin JR, Hung RJ. Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. *PloS one* 2011;6:e17479.
18. Field JK, Smith DL, Duffy S, Cassidy A. The Liverpool Lung Project research protocol. *Int J oncol* 2005;27:1633-45.
19. Clarke LC, Fraser SG. Hospital Episode Statistics and trends in ophthalmic surgery 1998-2004. *BMC Ophthalmol* 2006;4:37.
20. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;81:515-526.
21. Therneau TM, Grambsch PM. Martingale–based residuals for survival models. *Biometrika* 1990;77:147-60.
22. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996;49:907-16.

23. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assesing the increamental value of a new (bio)marker. *Heart* 2012;98:683-90.
24. Harrell FE, Califf RM, Prvor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;2472543-6.
25. Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 1982; 143**:**29-36.
26. Efron B. Bootstrap methods: Another look at the jackknife. *Ann. Statist* 1979; 7: 1–26.
27. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774-81.
28. 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* 1996;15:361-87.
29. Loughin TM, Koehler KJ. Bootstrapping regression parameters in multivariate survival analysis. *Lifetime Data Anal* 1997;3:157-77.
30. Schumacher M, Hollander N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building? *Stat Med* 1997;16:2813-27.
31. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003;95:470-8.
32. Cronin KA, Gail MH, Zou Z, Bach PB, Virtamo J, Albanes D. Validation of a model of lung cancer risk prediction among smokers *J Natl Cancer Inst* 2006;98:637-40.
33. D'Amello AM Jr, Cassidy A, Asomaning K, Raji OY, Duffy SW, Field JK, et al. Comparison of discriminatory power and accuracy of three lung cancer risk models. Br J Cancer 2010;103:423-9.
34. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation *J Natl Cancer Inst* 2011;103:1058-68.
35. Park S, Nam B-H, Yang H-R, Lee JA, Lim H, Han TJ, et al. Individualized risk prediction model for lung cancer in Koran men. *Plos One* 2013;8:e54823.
36. Hoggart C, Brennan P, Tjonneland A, Vogel U, Overvad K, Ostergaard JN, et al. A risk model for lung cancer incidence. *Cancer Prev Res (Phila)* 2012;5:834-46.
37. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. *N Engl J Med* 2013;368:728-36.
38. Whittemore AS. Evaluating health risk models. *Stat Med* 2010;29:2438-52.
39. Brenner DR, Boffetta P, Duell EJ, Bickeboller H, Rosenberger A, McCormack V, et al. Previous lung diseases and lung cancer risk: a pooled analysis from the International Lung Cancer Consortium. *Am J Epidemiol* 2012;176:573-85.
40. Rosenberger A, Bickeboller H, McCormack V, Brenner DR, Duell EJ, Tjonneland A, et al. Asthma and lung cancer risk: a systematic investigation by the International Lung Cancer Consortium. *Carcinogenesis* 2012;33:587-97.
41. Field JK, Chen Y, Marcus MW, McRonald FE, Raji OY, Duffy SW. The contribution of risk prediction models to early detection of lung cancer. J Surg oncol 2013;108:304-11.
42. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008;452:633-7.
43. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008;40:616-22.

44. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008;452:638-42.

45. Schmidt B, Liebenberg V, Dietrich D, Schlegel T, Kneip C, Seegebarth A, et al. SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates. *BMC cancer* 2010;10:600.
46. Nikolaidis G, Raji OY, Markopoulou S, Gosney JR, Bryan J, Warburton C, et al. DNA methylation biomarkers offer improved diagnostic efficiency in lung cancer. *Cancer Res* 2012;72:5692-701.
47. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, et al. A Prognostic DNA Methylation Signature for Stage I Non-Small-Cell Lung Cancer. J Clin Oncol 2013;31:4140-7.
48. Lin PY, Yu SL, Yang PC. MicroRNA in lung cancer. *B J Cancer* 2010;103:1144-8.
49. Zheng D, Haddadin S, Wang Y, Gu LQ, Perry MC, Freter CE, et al. Plasma microRNAs as novel biomarkers for early detection of lung cancer. *Int J Clin Exp Pathol* 2011;4:575-86.
50. Hennessey PT, Sanford T, Choudhary A, Mydlarz WW, Brown D, Adai AT, et al. Serum microRNA biomarkers for detection of non-small cell lung cancer. *PloS one* 2012;7:e32307.
51. Bediaga NG, Davies MP, Acha-Sagredo A, Hyde R, Raji OY, Page R, et al. A microRNA-based prediction algorithm for diagnosis of non-small lung cell carcinoma in minimal biopsy material. *B J Cancer* 2013;109:2404-11.

Table 1: Distribution of baseline characteristics stratified by lung cancer status

Characteristics	Incident lung cancer (n=237)	No –incident lung cancer (n=8523)	*P*-values
Mean Age (SD)	66.0 (7.2)	61.5 (8.5)	<0.001
Gender			0.002
Male	137 (57.8)	4052 (47.5)	
Female	100 (42.2)	4471 (52.5)	
Mean Smoking duration (SD)	39.9 (14.3)	18.9 (19.0)	<0.001
Prior diagnosis of pneumonia[a]			0.066
No	186 (83.0)	6971 (87.2)	
Yes	38 (17.0)	1022 (12.8)	
Prior diagnosis of asthma[a]			0.011
No	173 (77.2)	6687 (83.6)	
Yes	51 (22.8)	1310 (16.4)	
Prior diagnosis of tuberculosis[a]			0.674
No	216 (96.4)	7745 (96.9)	
Yes	8 (3.6)	246 (3.1)	
Prior diagnosis of COPD[a]			<0.001
No	90 (47.6)	5196 (82.6)	
Yes	99 (52.4)	1094 (17.4)	
Emphysema[a]			<0.001
No	203 (90.6)	7718 (96.5)	
Yes	21 (9.4)	277 (3.5)	
Bronchitis[a]			<0.001
No	135 (60.3)	5890 (73.7)	
Yes	89 (39.7)	2103 (26.3)	
Marital status[a]			0.017*
Single	17 (7.6)	725 (9.1)	
Married	133 (59.4)	5187 (65.0)	
Living together	3 (1.3)	202 (2.5)	
Widowed	44 (19.6)	958 (12.0)	
Divorced/separated	27 (12.1)	897 (11.2)	
other	0 (0.0)	15 (0.2)	
Occupational exposure to asbestos[a]			0.99
No	213 (99.5)	7877 (99.1)	
Yes	1 (0.5)	68 (0.9)	
Prior diagnosis of malignant tumor[a]			<0.001
No	154 (65.0)	7193 (89.9)	
Yes	70 (29.5)	805 (10.1)	
Education			<0.001
High school and below	201 (93.9)	6728 (84.3)	
Greater than high school	13 (6.1)	1256 (15.7)	
Family history of lung cancer			0.055
No	179 (75.5)	6940 (81.4)	
Early onset (<60 years)	20 (8.4)	484 (5.7)	
Late onset (≥60 years)	38 (16.0)	1099 (12.9)	

a= Numbers do not add up to total due to missing data; * P vales were derived from Fisher's Exact test.

Abbreviations: SD=Standard deviation; COPD= chronic obstructive pulmonary disease.

Table 2 Regression coefficients, hazard ratio (95% CI) and SE for covariates in the final model

Covariates	B-coefficient (SE)	HR (95%CI)	*P*-value
Age	0.036	1.04 (1.02-1.06)	<0.001
Gender (men vs. women)	0.391	1.48 (1.10-1.98)	0.009
Smoking duration	0.043	1.04 (1.03-1.05)	<0.001
COPD	0.890	2.43 (1.79-3.30)	<0.001
Prior diagnosis of malignant tumour	1.044	2.84(2.08-3.89)	<0.001
Family history of lung cancer			
None	Reference	Reference	
Early onset (<60 years)	0.521	1.68 (1.04-2.72)	0.034
Late onset ((≥60 years)	0.071	1.05(0.72-1.59)	0.722

Abbreviations: COPD= chronic obstructive pulmonary disease; HR= hazard ratio; SE = standard error

Based on the Cox model, the probability of developing lung cancer at an average of 8.7 years of follow-up is defined by the formula:

$$\hat{P} = 1 - S_0(t)^{\exp\left(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \overline{X}_i\right)}$$

$S_0(t)$ = the baseline survival at mean value 8.7 years;

β_i = the estimated regression coefficient,

X_i = the value of the covariate;

$\overline{X}_i$ = the corresponding mean for continuous covariates or proportion for categorical covariates;

P = the number of covariates.

Using the equation described above, the probability of developing lung cancer during the mean follow-up period of 8.7 years was calculated. This can be illustrated by considering a man diagnosed with lung cancer at the aged of 50, with 30 years smoking history and a known history of COPD and no other risk factors. The estimated risk based on the LLPi model is:

$\sum_{i=1}^{p} \beta_i X_i$ =0.036(50) +0.391(1) +0.043(30) +0.890(1) +1.044(0) +0.521(0) + 0.071(0) = 4.371

$\sum_{i=1}^{p} \beta_i \overline{X}_i$ =0.036(61.65) + 0.391(0.478) + 0.043(19.42) + 0.890(0.185) + 1.044(0.106) +

0.521(0.058) + 0.071(0.129) = 3.556

$\hat{P} = 1 - S_0(t)^{\exp\left(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \overline{X}_i\right)}$

= 1- $0.9728386^{\exp(4.371-3.556)}$ =0.060= 6.0%

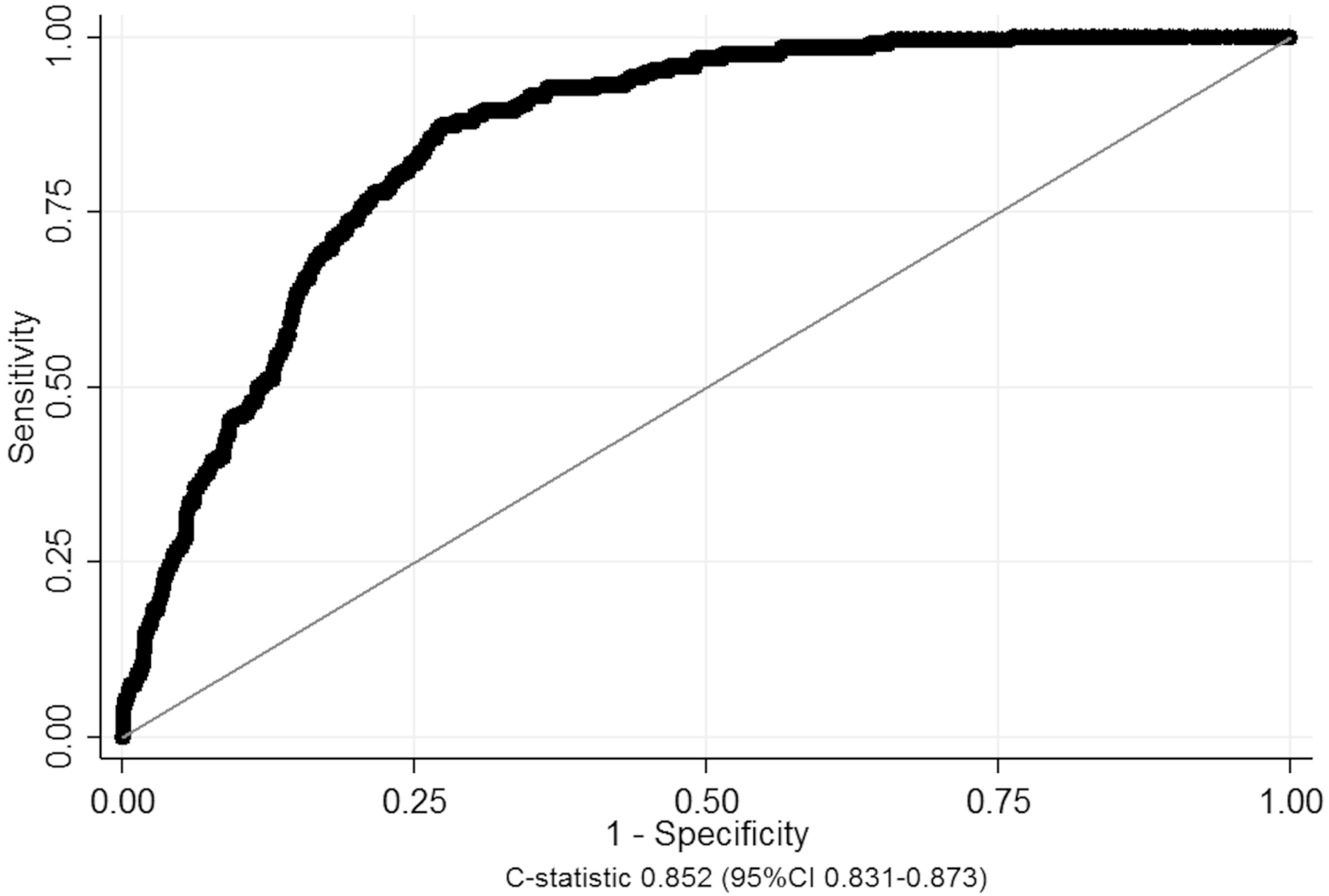
*Mutually adjusted baseline survival probability at 8.7 years $S_0(t)$=0.9728386

Table 3 Risk prediction model developed in prospective population-based cohort

Study characteristics	Bach (31)	PLCO (34)	$PLCO_{M2012}$ (37)	EPIC (36)	Park (35)
Population for modelling	14,254 heavy current- or former-smokers and 3,918 asbestos-exposed current- or former-smokers. Aged 45-69 at baseline. Population origin: United States.	70,962 cancer-free population-based individuals. Aged 55-74 at recruitment. Population origin: United States.	36,286 ever smoked population. Aged 55-74 at recruitment. Population origin: United States	169,035 former- and current-smokers from the general population (a random sample of 90% of each of the cases and controls as the training set, and the remaining 10% as validation set). Population origin: Europe.	1,309,144 men who underwent health examination. Aged 30-80 at baseline.) Population origin: South Korea.
Risk factors included in model	Age, sex, asbestos exposure, and smoking duration, duration of abstinence (if applicable) and average amount per day (while smoking).	Age, socioeconomic status (education), BMI, family history of lung cancer, COPD, recent chest X-ray and smoking status, pack-year, duration and time since quitting smoking (if applicable).	Age, race/ethnic group, education, BMI, COPD, personal history of cancer, family history of lung cancer and smoking-related factors (status, intensity, duration and quit time).	Age, smoking intensity (measured by average number of cigarettes smoked per day), age started smoking and smoking duration.	Age, smoking exposure (status and intensity), age at smoking initiation, BMI, physical activity and fasting glucose level.
Discriminatory power in modelling population	0.72	0.86 (for all subjects); 0.81 (for ever-smokers)	0.80	0.84	0.86
Discriminatory power in populations for external validation	0.69 (32); 0.66 (33)	0.84 (for all subjects); 0.78 (for ever-smokers only)	0.80	0.78	0.87
Strength	First lung cancer risk model.	Use of spline function in modelling; High discriminatory power.	Major classic risk factor included; high discriminatory power.	Large population; high discriminatory power.	Very large population; high discriminatory power.
Weakness	Moderate discriminatory power; smoker population only.	Healthy volunteer effect may limit external generalisation	Smoker population only.	Based on only age and smoking-related factors.	Men only; missing some other classical risk factors.

Abbreviations: PLCO = Prostate, Lung Colorectal and Ovarian cancer screening trial; EPIC = European Prospective Investigation into Cancer and Nutrition; BMI = Body Mass Index; COPD = Chronic Obstructive Pulmonary Disease

Figure 1 Performance of the LLPi risk model: C-statistic



AACR American Association for Cancer Research

Cancer Prevention Research

LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence

Michael W. Marcus, Ying Chen, Olaide Y. Raji, et al.

Cancer Prev Res Published OnlineFirst April 14, 2015.