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 8,760 participants ages 45 to 79 years, recruited between 1998 and 2008, was followed annually through the hospital episode statistics until January 31, 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 the bootstrap method. During mean follow-up of 8.7 years, 237 participants developed lung cancer. Age [hazard ratio (HR), 1.04; 95% confidence interval (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), chronic obstructive pulmonary disease (HR, 2.43; 95% CI, 1.79–3.30), prior diagnosis of malignant tumor (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 \( \chi^2 = 7.58, P = 0.371 \)). 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, 0.829–0.873). The LLPI risk model may assist in identifying individuals at high risk of developing lung cancer in population-based screening programs. Cancer Prev Res, 8(6); 570–5. ©2015 AACR.
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 nonmalignant 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 (<60 years), or late (>60 years)]. Previous history of cancer (except melanoma) was coded as “yes” or “no.”

Information on prior history of nonmalignant lung diseases such as asthma, COPD, pneumonia, and tuberculosis was 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 the χ² test or the Fisher 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 = 8,523) at the end of the follow-up period. The 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 unscaled and scaled Schoenfeld residuals (20). Follow-up duration was used as the time axis in the model. For participants who developed lung cancer, the follow-up time started from baseline visit, until the date of diagnosis of lung cancer. For participants without incident lung cancer, the follow-up duration was counted from the baseline visit until death, withdrawal, loss to follow-up, or January 31, 2013. Five percent of the 8,760 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 ≤ 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 reentered in the final multivariable model, with adjustment for the remaining significant covariates to ensure that no omitted covariate significantly reduced the log likelihood χ² 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 the jackknife 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 Grennesby–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) and SAS statistical software version 9.3 (SAS Institute).

Results

Overall, 8,760 participants ages 45 to 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 a family history of lung cancer and prior diagnosis of malignant tumor.

In univariate analysis, age, marital status, smoking duration, prior diagnosis of pneumonia, asthma, COPD, emphysema, bronchitis, prior diagnosis of malignant tumor, 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; 95% CI, 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; 95% CI, 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; 95% CI, 2.03–3.72). Prior diagnosis of malignant tumor 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 socioeconomic 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-year intervals 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 number 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 tumor (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; Fig. 1) and 0.849 (95% CI, 0.829–0.870) by internal validation with bootstrap resampling and correction for optimism. The Grønnesby–Borgan goodness-of-fit test demonstrated overall good calibration \( \chi^2 7.58, P = 0.371 \). 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 at an average of 8.7 years of follow-up is defined by the formula:

\[
P = 1 - S_0(t)^\exp\left(\sum_{i=1}^{n} \beta_i X_i \right)
\]

where \( S_0(t) \) is the baseline survival at mean value 8.7 years, \( \beta_i \) is the estimated regression coefficient, \( X_i \) is the value of the covariate, \( X_i \) is the corresponding mean for continuous covariates or proportion for categorical covariates, and \( P_i \) is the number of covariates.

Based on the Cox model, the probability of developing lung cancer at an average of 8.7 years was calculated. This can be illustrated by considering a man diagnosed with lung cancer at the age 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

\[
\sum_{i=1}^{n} \beta_i X_i = 0.035(60) + 0.391(30) + 0.520(0) + 0.071(0) = 4.371.
\]

\[
P = 1 - S_0(t)^\exp\left(4.371\right) = 0.060 = 6.0%.
\]

*Mutually adjusted baseline survival probability at 8.7 years \( S_0(t) = 0.9728386 \).
risk of developing lung cancer for three hypothetical individuals with diverse risk profile. First, the absolute risk of a woman ages 65 with 37 years smoking history and a history of COPD diagnosis with a late-onset family history of lung cancer (ages <60 at diagnosis) and no other risk factors is calculated as $P = 1 - S_0(t)^{exp\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} b_i X_i\right)} = 9.9\%$ (see details of the formula under Table 2). Second, the absolute risk of a man ages 67 without a smoking history but with an early-onset family history of lung cancer (ages <60 at diagnosis) and a prior diagnosis of cancer is 6%. Third, the absolute risk of a man ages 73 who has smoked for 59 years and also had an early-onset (ages <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 tumor, 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, 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 programs.

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 and colleagues 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 (PLCO) cancer screening used prospective data to build two risk prediction models for the general population (model 1) and a subcohort of ever-smokers (model 2; ref. 34). The bootstrap optimism-corrected ROC AUC were 0.857 and 0.805, respectively. External validation of the models produced a C-statistic of 0.871 (95% CI, 0.867–0.876; ref. 35). Hoggart and colleagues used prospective data from the European Prospective Investigation into Cancer (EPIC) to build a predictive model for lung cancer.

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

<table>
<thead>
<tr>
<th>Study characteristics</th>
<th>Bach (31)</th>
<th>PLCO (34)</th>
<th>PLCO 2012 (37)</th>
<th>EPIC (36)</th>
<th>Park (35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population for modeling</td>
<td>44,254 heavy current or former smokers and 3,918 asbestos-exposed current or former smokers. Ages 45 to 69 at baseline. Population origin: United States.</td>
<td>70,962 cancer-free population-based individuals. Ages 55 to 74 at recruitment. Population origin: United States.</td>
<td>36,286 ever smoked population. Ages 55 to 74 at recruitment. Population origin: United States.</td>
<td>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 the validation set). Population origin: Europe.</td>
<td>1,309,144 men who underwent health examination. Ages 30 to 80 at baseline. Population origin: South Korea.</td>
</tr>
<tr>
<td>Risk factors included in model</td>
<td>Age, sex, asbestos exposure, and smoking duration, duration of abstinence (if applicable), and average amount per day (while smoking).</td>
<td>Age, socioeconomic status (education), BMI, family history of lung cancer, COPD, and smoking-related factors.</td>
<td>Age, race/ethnic group, education, BMI, COPD, personal history of cancer, smoking-related factors (status, intensity, duration, and quit time).</td>
<td>Age, smoking intensity (measured by average number of cigarettes smoked per day), age started smoking, and smoking duration.</td>
<td>Age, smoking exposure (status and intensity), age at smoking initiation, BMI, physical activity, and fasting glucose level.</td>
</tr>
<tr>
<td>Discriminatory power in modeling population</td>
<td>0.72</td>
<td>0.86 (for all subjects); 0.81 (for ever-smokers)</td>
<td>0.80</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>Discriminatory power in populations for external validation</td>
<td>0.69 (32); 0.66 (33)</td>
<td>0.84 (for all subjects); 0.78 (for ever-smokers only)</td>
<td>0.80</td>
<td>0.78</td>
<td>0.87</td>
</tr>
<tr>
<td>Strength</td>
<td>First lung cancer risk model. Use of spline function in modeling; high discriminatory power. Healthy volunteer effect may limit external generalization</td>
<td>Major classic risk factor included; high discriminatory power. Smoker population only. Based on only age and smoking-related factors.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weakness</td>
<td>Moderate discriminatory power; smoker population only.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: BMI, body mass index.
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; ref. 36). In another study, Tammemagi and colleagues modified the PLCO models 1 and 2 to ensure applicability to NLST (37). The newly developed model PLCO2012 utilized 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 etiology 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 among 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 with 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 1 of the 237 individuals who developed lung cancer was exposed to occupational asbestos. Twenty-three participants had missing data on occupational exposure to asbestos, and 213 participants who 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 participants. 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 standardized questionnaires. In addition, information bias was prevented by complete documentation of lung cancer incidence (by the ONS, the North West Cancer Intelligence 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 has 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 etiology of lung cancer from our own and numerous other case–control and cohort studies. The application of the LLPi risk model in identifying individuals at high risk of developing lung cancer in population-based screening programs needs further study.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors’ Contributions

Conception and design: M.W. Marcus, Y. Chen, O.Y. Raji, J.K. Field

Development of methodology: M.W. Marcus, Y. Chen, O.Y. Raji, S.W. Duffy, J.K. Field

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.K. Field

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M.W. Marcus, O.Y. Raji

Writing, review, and/or revision of the manuscript: M.W. Marcus, O.Y. Raji, S.W. Duffy, J.K. Field

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.W. Marcus, J.K. Field

Study supervision: M.W. Marcus, J.K. Field

Grant Support

The Liverpool Lung project was principally funded by the Roy Castle Lung Cancer Foundation, UK. M.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 (CAOS project).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 1, 2014; revised March 16, 2015; accepted April 5, 2015; published OnlineFirst April 14, 2015.
LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence

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


Updated version
Access the most recent version of this article at:
doi:10.1158/1940-6207.CAPR-14-0438

Cited articles
This article cites 50 articles, 8 of which you can access for free at:
http://cancerpreventionresearch.aacrjournals.org/content/8/6/570.full#ref-list-1

Citing articles
This article has been cited by 9 HighWire-hosted articles. Access the articles at:
http://cancerpreventionresearch.aacrjournals.org/content/8/6/570.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://cancerpreventionresearch.aacrjournals.org/content/8/6/570.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.