A Risk Model for Prediction of Lung Cancer

Margaret R. Spitz, Waun Ki Hong, Christopher I. Amos, Xifeng Wu, Matthew B. Schabath, Qiong Dong, Sanjay Shete, Carol J. Etzel

Background Reliable risk prediction tools for estimating individual probability of lung cancer have important public health implications. We constructed and validated a comprehensive clinical tool for lung cancer risk prediction by smoking status.

Methods Epidemiologic data from 1851 lung cancer patients and 2001 matched control subjects were randomly divided into separate training (75% of the data) and validation (25% of the data) sets for never, former, and current smokers, and multivariable models were constructed from the training sets. The discriminatory ability of the models was assessed in the validation sets by examining the areas under the receiver operating characteristic curves and with concordance statistics. Absolute 1-year risks of lung cancer were computed using national incidence and mortality data. An ordinal risk index was constructed for each smoking status category by summing the odds ratios from the multivariable regression analyses for each risk factor.

Results All variables that had a statistically significant association with lung cancer (environmental tobacco smoke, family history of cancer, dust exposure, prior respiratory disease, and smoking history variables) have strong biologically plausible etiologic roles in the disease. The concordance statistics in the validation sets for the never, former, and current smoker models were 0.57, 0.63, and 0.58, respectively. The computed 1-year absolute risk of lung cancer for a hypothetical male current smoker with an estimated relative risk close to 9 was 8.68%. The ordinal risk index performed well in that true-positive rates in the designated high-risk categories were 69% and 70% for current and former smokers, respectively.

Conclusions If confirmed in other studies, this risk assessment procedure could use easily obtained clinical information to identify individuals who may benefit from increased screening surveillance for lung cancer. Although the concordance statistics were modest, they are consistent with those from other risk prediction models.

J Natl Cancer Inst 2007;99:715–26

Approximately 85% of all lung cancers occur in current or former cigarette smokers (1). Peto et al. (2) has computed the cumulative risk of lung cancer in long-term cigarette smokers by age 75 to be approximately 16% for men and 9.5% for women. These findings pose two challenges for estimating lung cancer risk. First, how do we identify those ever cigarette smokers who have the highest risk of developing lung cancer? Second, what are the risk factors for the 15% of lung cancers that occur in lifetime never smokers?

The potential public health benefits of individualized estimates of the probability of developing lung cancer are large. Prevention of even 10% of annual deaths from lung cancer would save more than 16000 lives, more than all the annual deaths in the United States from ovarian cancer or from brain cancers (3). High-risk individuals could undergo a program of screening surveillance that might not be appropriate for lower risk individuals and may also consider chemoprevention interventions. Moreover, as Bach et al. (4) have pointed out, risk prediction tools could be incorporated into the design of smaller, more powerful, and "smarter" prevention trials by enriching the number of observed events.

Risk prediction is most well developed in the context of cardiovascular diseases, for which prediction models use a combination of variables (blood pressure, smoking history, lipid levels, and family history of heart disease) to assess an individual's risk of heart disease (5). Risk data from the Framingham Heart Study have been used to construct the Framingham Coronary Risk Prediction Model and to formulate guidelines for cholesterol-lowering therapy (6). As Grundy et al. (6) note, the Framingham risk scores can both motivate and reassure the patient; they also illustrate the cumulative nature of multiple risk factors.

The National Cancer Institute in its 2006 budget proposal cited risk prediction as an area of extraordinary opportunity (7). The

Affiliations of authors: Department of Epidemiology (MRS, CIA, XW, QD, SS, CJE) and Division of Cancer Medicine (WKH), The University of Texas M. D. Anderson Cancer Center, Houston, TX; The University of Texas School of Public Health, Houston, TX (MBS).

Correspondence to: Margaret R. Spitz, MD, MPH, Department of Epidemiology-Unit 1340, The University of Texas M. D. Anderson Cancer Center, PO Box 301439, Houston, TX 77230-1439 (e-mail: mspitz@mdanderson.org).

See "Notes" following "References."

DOI: 10.1093/jnci/djk153

CONTEXT AND CAVEATS

Background

Risk prediction models for cancer could be valuable for identifying individuals who may benefit from preventive treatments or increased surveillance or who are good candidates to participate in clinical trials. Existing models for lung cancer prediction focus mainly on smokers.

Study design

Predictive models were developed for never, former, and current smokers using a portion of the data from a case–control study of lung cancer. The models were validated using the rest of the data.

Contribution

The models could predict the development of lung cancer with modest discriminatory accuracy, similar to that of other cancer prediction models. The statistically significant variables in the models (including history of exposure to environmental tobacco smoke, family history of cancer, dust and asbestos exposure, history of respiratory diseases, and smoking history) can be assessed by patient interview.

Implications

The models can be used to compute absolute risks of lung cancer, and risks can be presented using an easy-to-understand ordinal risk index that may be helpful for risk communication.

Limitations

The models may not be sufficiently discriminatory to allow accurate risk assessment at the individual level. In addition, the models were developed in a single population and need to be validated in independent populations.

best known and most widely applied cancer risk prediction model is that developed by Gail et al. (8), which uses a woman's current age and panel of risk factors to estimate her risk of breast cancer over a defined period, taking into account information on relative risks (RRs), baseline hazard rate, and competing risks. Newer modifications of the Gail model with increased numbers of risk factors (9) or addition of mammographic density (10) to the model have produced modest improvements in discriminatory power. Women found to be at high risk based on their Gail model scores are encouraged to undergo screening or genetic evaluation and are eligible to be enrolled in chemoprevention trials (e.g., the Study of Tamoxifen and Raloxifene). The Gail model has been shown to reliably predict risk at the population level, but its discriminatory accuracy at the individual level is modest (11).

Other statistical models have been developed to estimate an individual's risk of developing breast cancer (10,12,13), colorectal cancer (14–16), melanoma (17,18), ovarian cancer (19), and prostate cancer (20). Relatively few models have been developed to estimate lung cancer risk. Previous lung cancer risk prediction models (2,4,21) have tended to concentrate on smoking characteristics, sex, and age. Bach et al. (4) used smoking history data from a large randomized trial of retinol and carotene in heavy smokers and asbestos-exposed individuals to generate a lung cancer risk prediction model that is applicable to smokers between 50 and 75 years of age, who are or were heavy smokers (10–60 cigarettes per day for 25–60 years) and who had quit no more than 20 years previously.

To extend the work of Bach et al. (4) and to include additional risk factors beyond smoking history and asbestos exposure, we used epidemiologic data from a large case–control study of lung cancer to construct and validate a risk prediction tool for lung cancer. We divided the data into training sets to guide model development and validation sets to assess the prediction of risk. Matching variables were not included in the analysis, and because the study design matched on smoking status, all model building and analytic approaches were stratified by this important predictor. We constructed multivariable models separately for never, former, and current smokers, incorporating into each model variables that exhibited statistically significant main effects. We computed the absolute risk of lung cancer in the presence of competing causes of death. Finally, ordinal risk indices were constructed from the statistically significant risk factors that were included in the final models.

Methods

Study Population and Epidemiologic Data

The recruitment of case patients and control subjects for an ongoing molecular epidemiology study of lung cancer has been described previously (22,23). Briefly, lung cancer patients have been recruited from the Thoracic Center at The University of Texas M. D. Anderson Cancer Center since July 1995. The case patients are all newly diagnosed patients presenting with histologically confirmed lung cancer and are enrolled before initiation of chemo- or radiation therapy. There are no age, sex, ethnicity, or disease stage restrictions on recruitment, but emphasis has been placed on enrolling subsets of special interest, including minority patients, younger (<50 years of age) patients, and lifetime never smokers. Healthy control subjects without a prior history of cancer (except nonmelanoma skin cancer) are recruited from the Kelsey–Seybold clinics, the largest private multispecialty physician group in the Houston metropolitan area, which includes a network of 23 clinics and more than 300 physicians. Control subjects are frequency matched to the case patients by age (±5 years), sex, ethnicity, and smoking status (never, former, or current). All study participants provide written informed consent, and trained M. D. Anderson interviewers administer an epidemiologic questionnaire to study participants.

Data collected at the interview include demographic characteristics, smoking history, occupation, information about specific exposures at work or from hobbies, medical history, and family history of cancer in first-degree relatives. An individual who has never smoked or has smoked less than 100 cigarettes in his or her lifetime is defined as a never smoker. An individual who has smoked at least 100 cigarettes in his or her lifetime but quit smoking more than 12 months before lung cancer diagnosis (for case patients) or before the interview (for control subjects) is considered to be a former smoker. Current smokers include those currently smoking and "recent quitters," i.e., those who quit smoking less than 12 months before diagnosis (for case patients) or interview (for control subjects). Data on smoking history include smoking duration, number of cigarettes smoked per day, computed pack-years smoked, and age at smoking initiation (for all smokers) plus age at smoking cessation and computed years since cessation (for

former smokers). Exposure to second-hand smoke (environmental tobacco smoke, or ETS) is ascertained for never and former smokers and is defined as having been exposed to someone else's cigarette smoke at home or at work on a regular basis, i.e., daily or weekly, as well as on years of exposure to ETS.

Family history of cancer in first-degree relatives is obtained from participant-reported cancer histories in parents, siblings, and offspring. For each affected relative, we obtain information on year of birth, age at time of interview of the case or control subject, smoking status (never or ever), type of cancer, age at diagnosis, and year of death. For each unaffected relative, we obtain information on current age or age at death. Family histories of any cancer and of smoking-related cancers (defined as lung, upper aerodigestive tract, esophagus, renal, pancreas, bladder, cervix) are analyzed separately.

Participants are classified as positive for asbestos exposure if they report having been employed within a documented asbestos-related occupation or industry. Other self-reported exposures are classified as regular (8 hours a week) and/or prolonged (at least 1 year) exposure to a predefined list of chemicals (solvents, paint thinners, dry cleaning fluids, motor oils, gasoline, tar, hydrochloric acid, or bleach/cleansers); fumes (glues, paints, plastics, pesticides, car and truck exhaust, natural gas, or foam insulation); dusts (metal, concrete, sawdust, cotton, textile fibers, fiberglass, sand or dust, or dust storms); or a subcategory of wood dusts based on specific self-reported work exposures to wood dust, sawdust, or sanding dust. Participants are asked whether they have ever been diagnosed by a physician with emphysema, hay fever, or asthma.

All study participants are Texas residents. This analysis was limited to white non-Hispanic participants because there were inadequate numbers of nonwhite participants to perform smoking stratum–specific analyses. Through May 2006 (end date for this analysis), response rates among both the case patients and the control subjects have averaged 75%. This research has been approved by the Institutional Review Boards of the M. D. Anderson Cancer Center and the Kelsey–Seybold Clinics.

Statistical Analysis

The data for each smoking stratum were initially split at random into training sets (constituting 75% of the participants in the stratum) to guide the building of the smoking status–specific risk models and validation sets (constituting the remaining 25% of participants) to assess performance of each of the three models individually. Before building the models, we screened all variables (i.e., ETS exposure; physician-diagnosed emphysema, hay fever, or asthma; exposure to dusts, fumes, chemicals, asbestos, pesticides, or wood dusts; family history of cancer; age at smoking initiation, age at smoking cessation, number of years since smoking cessation, and pack-years of smoking) by univariate logistic regression in the training sets to examine their main effects by smoking status and by sex. Differences in the distribution of demographic variables (including sex and smoking status) between case patients and control subjects were evaluated by the two-sided chi-square test. For univariate analysis, several of the continuous variables were categorized as follows: age stopped smoking (<42 years, 42–53 years, and ≥54 years, based on the tertile distribution of age at cessation in former smoker control subjects in the training set) and pack-years of smoking (<28 pack-years, 28–41.9 pack-years, 42–57.4 pack-years, and ≥57.5 pack-years, based on the quartile distribution of pack-years among control current smokers in the training set). Differences in the distribution of continuous variables were evaluated using two-sided Student's t test. Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated as estimates of relative risk. Unless otherwise stated, all analyses were performed using Statistical Analysis System (SAS) software Version 9.1 (SAS Institute, Cary, NC).

Risk Model Building. Variables that were statistically significantly associated with lung cancer risk at the 5% level in univariate analysis in the three training sets were included in the multivariable logistic regression analyses for construction of the final risk models. In these analyses, we used the pack-year variable as the measure of smoking intensity for current and former smokers. There were no marked differences in risk estimates by sex, and we therefore report the relative risk estimates and three model validation results for males and females combined.

We used a backward selection procedure to choose the variables included in each of the final multivariable models. To further minimize the possibility of confounding effects due to high collinearity between predictor variables, we calculated a variance inflation factor for each variable (24). All variance inflation factor values were well below 10 (data not shown), indicating no collinearity between the final list of predictor variables within any of the models.

To test for the statistically significant contribution of interaction terms, we included each pairwise interaction term in the preliminary main-effects models and reran the logistic regression. No interaction terms were found to be statistically significant at the P less than .05 level in either the never-smoker or former-smoker models. One interaction term (emphysema and family history of smoking-related cancers) was statistically significant ($P = .03$) in the current-smoker model. However, given that multiple statistical tests were performed and that the interaction term did not substantially modify the Akaike Information Criterion (data not shown), we elected not to include this interaction in the final model.

Classification and Regression Tree Analysis. Due to sample size restrictions, we were not able to investigate all possible higher order interactions in our model building. Therefore, we used classification and regression tree (CART) analysis (25) to evaluate higher order (three-way and above) interactions in the training sets. We applied the recursive partitioning technique "rpart" package that was developed for Splus (Insightful Corporation, Seattle, WA) by Therneau and Atkinson (26) (http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm) to discriminate low- and high-risk subgroups. We grew each decision tree (one for each smoking status group) with the stipulation that each subsequent split yields two daughter nodes with at least 10 participants per node. An unconditional logistic regression model was fit at each recursive split to estimate the risk of lung cancer (as odds ratios with 95% confidence intervals) adjusted by age and sex; any branch that was not deemed to be statistically significant ($P<.05$) was pruned off the tree.

Assessment of Model Fit. We used a three-phase approach to model validation: First, each of the three models was evaluated by the Hosmer–Lemeshow goodness-of-fit test in the appropriate validation set. Next, using the validation set for each risk model, we calculated specificity and sensitivity of the resulting logistic regression model by constructing receiver operating characteristic curves and calculated the area under the curve (AUC) statistic to estimate each model's ability to discriminate between patients and control subjects. Approximate 95% confidence intervals for the AUCs were calculated using SPSS statistical software (SPSS Inc V12.0, Chicago, IL), assuming a binegative exponential distribution. An AUC of 0.5 indicates chance prediction (equivalent to a coin toss), whereas a statistic of 0.7 indicates good discrimination (27). Once the final variables for the three models had been validated in their respective validation sets, data from the training and validation sets were combined to calculate more precise overall risk estimates.

After calculating the risk estimates for the final models (which were based on the combined training and validation sets), we also evaluated final model discrimination by performing threefold cross validation, as follows (28). For each model, we randomly divided the combined data from the training and validation sets into three equally sized groups. Using two of the groups, we built a risk model (separately for never, former, and current smokers) using the final set of variables. For each risk model, we then used the remaining data group to calculate the concordance statistic (*C*) for each model, which, like the AUC, is an index of the model's ability to predict case patient or control subject status. We repeated this process three times and calculated the average *C*, namely $\overline{C}$ and its associated standard deviation, $S_{\overline{C}}$, across the cross-validations. We then calculated 95% confidence intervals as $C \pm 1.96S_{\overline{C}}$.

Risk Index. To create a simple and logical index for risk stratification, we generated a numerical score by assigning integer points based on the odds ratios from the logistic regression model for each risk factor that was statistically significantly associated with lung cancer in the respective multivariable model. The points were summed to compute a score that can be used to assign individuals to low-, medium-, or high-risk groups within their smoking status category. We used CART analysis on the training set to determine the cut points for low, medium, and high risk scores and then used these cut points to categorize subjects according to their risk. When necessary, we rounded to the nearest tenth of a decimal point for ease of scoring. These scoring metrics were developed in the training sets and validated in the validation sets. This approach provided an additional method to evaluate the discriminatory prediction accuracy of the models and also allowed us to convert a numeric risk estimate to a risk level assignment that lay users might find easier to interpret. We classified case patients and control subjects by their risk scores and calculated true-positive and true-negative rates for the high- and low-risk score categories, respectively.

Absolute 1-Year Risk of Lung Cancer. Estimates of absolute risk were developed based on the methods of Gail et al. (8) and Fears et al. (18). To calculate relative risk estimates, we multiplied the log-odds of the individual risk components from the final logistic model and denoted the relative risk as *r*. We estimated baseline hazards for male and female never, former, and current smokers separately as $h_{1ji} = v_{ji}(1-s_i)$, where v_{ji} is the age-specific incidence rate of lung cancer for men ($j = 1$) and women ($j = 2$) from the Surveillance, Epidemiology, and End Results (SEER) program for 2005 (29) adjusted for smoking status (never smokers [$i = 1$], former smokers [$i = 2$], and current smokers [$i = 3$]) and s_i is the attributable risk derived from the relative risk model as described in Fears et al. (18) for never smokers ($i = 1$), former smokers ($i = 2$), and current smokers ($i = 3$). We obtained the adjusted incidence rates in the following manner: Define I_j as the age-specific incidence rate of lung cancer for males ($j = 1$) and females ($j = 2$) from SEER data. This value is the ratio of the number of new cases (s) to the number of individuals at risk in the population (N), $I_j = s/\mathrm{N}$. Approximately 90% of all new male lung cancer cases and 80% of all new female lung cancer cases are ever smokers (30), and 23.2% of men and 19.2% of women are current smokers, 48.4% of men and 59.5% of women are never smokers (31), leaving 28.4% of men and 21.3% of women estimated as former smokers. Therefore, the adjusted incidence rate for male current smokers ($j = 1$, $i = 3$) can be estimated as the ratio of the number of new lung cancer cases who currently smoke to the number of current smokers in the population, which can be written as $v_{13} = 0.90s/0.232\mathrm{N} = (0.90/0.232)I_1$, where I_1 is the age-specific incidence rate for males. Hence, the adjusted incidence rate for any gender–smoking group combination can be estimated as $v_{ji} = c_{ji}I_j$, where c_{ji} is defined as the adjustment constant for each sex–smoking status group. Values for c_{ji} are given in Appendix Table 1. Furthermore, if a is the age in years and h_{2j} the mortality rate from other causes excluding lung cancer for males ($j = 1$) and females ($j = 2$) derived from National Center for Health Statistics (NCHS), 1999–2003 mortality rates (32) (Appendix Table 2), the absolute 1-year risk is estimated as

$$P(a, r, i, j) = \left(\frac{h_{1ji} r}{h_{1ji} r + h_{2j}} \right) \{1 - \exp[-(h_{1ji} r + h_{2j})]\}$$

We evaluated the utility of the absolute 1-year risk model for lung cancer by creating three different risk profile examples. We converted the odds ratios of the selected risk factors (Table 3) to relative risks using the formula RR = OR/(1 – P) + (P × OR). In general, P is defined as the incidence of the outcome of interest in the unexposed group. For our calculations, we define P as the age- and sex-specific incidence rates of lung cancer for white men and women obtained from SEER data. However, in all instances, we found that the odds ratio approximated the relative risk closely, and we have therefore used odds ratios for all subsequent estimations.

Results

Epidemiologic data from 1851 patients with lung cancer and 2001 control subjects, all of whom were non-Hispanic whites, were available for this analysis (Table 1). By study design, the case patients and control subjects were well matched by sex, although there was a slight excess of males. Case patients were, on average, 2 years older than control subjects but within the 5-year frequency matching criterion (Table 1). There were no statistically

Table 1. Distribution of study population by select variables

Variables	Case patients (N = 1851)	Control subjects (N = 2001)	P*
Sex, n (%)			
Male	975 (52.7)	1021 (51.0)	
Female	876 (47.3)	980 (49.0)	.306
Mean age (SD†), y	62.0 (11.2)	60.2 (10.7)	<.001
Smoking status, n (%)			
Never	330 (17.8)	379 (18.9)	
Former	784 (42.4)	884 (44.2)	
Current	737 (39.8)	738 (36.9)	.170
Mean pack-years smoked (SD)	51.9 (31.7)	44.6 (30.7)	<.001

* *P* value from the two-sided chi-square test (for categorical variables) and Student's *t* test (for continuous variables).

† SD = standard deviation.

significant differences in the distribution of case patients and control subjects by smoking status, a matching criterion, although there were more current smokers (39.8%) and fewer former smokers (42.4%) among the case patients than among the control subjects (36.9% and 44.2%, respectively). Not surprisingly, the case patients were also heavier smokers than the control subjects in terms of pack-years smoked (51.9 and 44.6, respectively); case patients had smoked cigarettes for an average of 36.1 years (±standard deviation [SD] 12.5), compared with 32.7 years (±SD 13.0) for the control subjects ($P<.001$). In addition, case patients were heavier daily smokers (mean cigarettes smoked per day = 28.1 [±SD 13.7]) than control subjects (26.4 [±SD 14.4]; $P\leq.001$; data not shown in Table).

To identify risk factors to include in the multivariable model, we performed univariate analyses by smoking status (Table 2). Among never smokers (330 case patients and 379 control subjects), exposure to ETS (OR = 1.77, 95% CI = 1.2 to 2.6) or dust (OR = 1.48, 95% CI = 1.0 to 2.1) and family history of any cancer in two or more first-degree relatives (OR = 1.96, 95% CI = 1.3 to 2.9) were all statistically significantly associated with lung cancer risk. Asthma was associated with a 1.43-fold increase in risk among never smokers, but this increase was not statistically significant (95% CI = 0.9 to 2.2). Among former smokers (784 case patients, 884 control subjects), risks were statistically significantly elevated with exposure to ETS (OR = 2.07, 95% CI = 1.3 to 3.2), dusts (OR = 1.64, 95% CI = 1.3 to 2.0), fumes (OR = 1.32, 95% CI = 1.1 to 1.6), and chemicals (OR = 1.25, 95% CI = 1.0 to 1.5); with a history of emphysema (OR = 2.99, 95% CI = 2.2 to 4.0); with a family history in two or more relatives of any cancer (OR = 1.84, 95% CI = 1.4 to 2.4) or smoking-related cancers (OR = 1.40, 95% CI = 1.1 to 1.7); and with a history of hay fever (OR = 0.72, 95% CI = 0.6 to 0.9). Risk factors among current smokers (737 case patients, 738 control subjects) were similar to

Table 2. Univariate analysis of lung cancer risk factors (as odds ratios with 95% confidence intervals) by smoking status*

Risk factor	Never smokers (330 case patients/ 379 control subjects)	Former smokers (784 case patients/ 884 control subjects)	Current smokers (737 case patients/ 738 control subjects)
Exposure			
ETS	**1.77 (1.2 to 2.6)**	**2.07 (1.3 to 3.2)**	NA
Emphysema	NA	**2.99 (2.2 to 4.0)**	**2.69 (2.0 to 3.6)**
Hay fever	0.90 (0.7 to 1.3)	**0.72 (0.6 to 0.9)**	**0.62 (0.5 to 0.8)**
Dusts	**1.48 (1.0 to 2.1)**	**1.64 (1.3 to 2.0)**	**1.67 (1.4 to 2.1)**
Fumes	1.02 (0.7 to 1.4)	**1.32 (1.1 to 1.6)**	**1.31 (1.1 to 1.6)**
Chemicals	1.00 (0.7 to 1.4)	**1.25 (1.0 to 1.5)**	**1.34 (1.1 to 1.7)**
Asbestos	0.86 (0.4 to 1.8)	1.25 (0.9 to 1.7)	**1.78 (1.3 to 2.4)**
Pesticides	1.52 (0.7 to 3.3)	1.22 (0.8 to 1.8	1.0 (0.7 to 1.4)
Wood dust	0.87 (0.5 to 1.7)	1.23 (0.9 to 1.7)	1.20 (0.9 to 1.6)
Asthma	1.43 (0.9 to 2.2)	1.23 (0.9 to 1.6)	1.01 (0.8 to 1.4)
Family history of cancer†			
0	1.0 (referent)	1.0 (referent)	1.0 (referent)
1	1.19 (0.8 to 1.7)	1.16 (0.9 to 1.5)	1.24 (1.0 to 1.6)
≥2	**1.96 (1.3 to 2.9)**	**1.84 (1.4 to 2.4)**	**1.68 (1.3 to 2.2)**
Family history of smoking-related cancer†			
0	1.0 (referent)	1.0 (referent)	1.0 (referent)
≥1	1.17 (0.8 to 1.7)	**1.40 (1.1 to 1.7)**	**1.58 (1.3 to 2.0)**
Age at smoking initiation (continuous)	NA	1.01 (0.99 to 1.0)	0.97 (0.95 to 1.0)
Age stopped smoking (>39 y versus ≤38 y)	NA	**1.57 (1.2 to 2.0)**	NA
Age at smoking cessation (continuous)	NA	**1.03 (1.02 to 1.04)**	NA
Years of cessation (continuous)	NA	0.99 (0.99 to 1.00)	NA
Pack-years smoked (continuous)	NA	**1.00 (1.00 to 1.01)**	**1.01 (1.01 to 1.02)**

* Numbers in bold type indicate statistically significant odds ratios. ETS = environmental tobacco smoke; NA = not applicable.

† Number of first-degree relatives with cancer. Smoking-related cancers include renal cancer and cancers of the lung, upper aerodigestive tract, esophagus, pancreas, bladder, and cervix.

Table 3. Multivariable logistic model for lung cancer by smoking status*

Risk factor	Regression coefficient	*P*†	OR (95% CI)
Never smoker			
Intercept	−0.8806	<.001	
ETS (yes vs no)	0.5874	.0042	1.80 (1.20 to 2.69)
Family history (≥2 vs <2)‡	0.6954	<.001	2.00 (1.39 to 2.90)
Former smoker			
Intercept	−0.7606	<.001	
Emphysema (yes vs no)	0.9734	<.001	2.65 (1.95 to 3.60)
Dust exposure (yes vs no)	0.4654	<.001	1.59 (1.29 to 1.97)
Family history (≥2 vs <2)‡	0.4636	<.001	1.59 (1.28 to 1.98)
Age stopped smoking§			
<42 y	Referent		
42–53 y	0.2130	.1110	1.24 (0.95 to 1.61)
≥54 y	0.4080	.0018	1.50 (1.16 to 1.94)
	P for trend = .017		
Hay fever (no)	0.3711	.00e55	1.45 (1.12 to 1.88)
Current Smoker			
Intercept	−0.7173	<.001	
Emphysema (yes)	0.7561	<.001	2.13 (1.58 to 2.88)
Pack-years‖			
<28	Referent		
28–41.9	0.2219	.1932	1.25 (0.89 to 1.74)
42–57.4	0.3747	.0241	1.45 (1.05 to 2.01)
≥57.5	0.6151	<.001	1.85 (1.35 to 2.53)
	P for trend<.001		
Dust exposure (yes vs no)	0.3067	.0075	1.36 (1.09 to 1.70)
Asbestos exposure (yes vs no)	0.4109	.0127	1.51 (1.09 to 2.08)
Family history¶			
0	Referent		
≥1	0.3859	.0021	1.47 (1.15 to 1.88)
Hay fever (no)	0.4047	.0054	1.49 (1.13 to 1.99)

* Regression analysis was based on entire dataset (training and validation sets combined). OR = odds ratio; CI = confidence interval; ETS = environmental tobacco smoke.

† *P* value from Wald test.

‡ Number of first-degree relatives with any cancer.

§ Cut points based on the tertile of age at smoking cessation in control subjects in the training set.

‖ Cut points based on the quartile of current smoker pack-years in control subjects in the training set.

¶ Number of first-degree relatives with a smoking-related cancer (i.e., renal cancer and cancers of the lung, upper aerodigestive tract, esophagus, pancreas, bladder, and cervix).

those for former smokers, with statistically significant associations for emphysema (OR = 2.69, 95% CI = 2.0 to 3.6); exposure to dusts (OR = 1.67, 95% CI = 1.4 to 2.1), fumes (OR = 1.31, 95% CI = 1.1 to 1.6), chemicals (OR = 1.34, 95% CI = 1.1 to 1.7), or asbestos (OR = 1.78, 95% CI = 1.3 to 2.4); a family history in two or more relatives of any cancer (OR = 1.68, 95% CI = 1.3 to 2.2) or a smoking-related cancer (OR = 1.58, 95% CI = 1.3 to 2.0); and a history of hay fever (OR = 0.62, 95% CI = 0.5 to 0.8). Smoking variables (age at smoking cessation, for former smokers, and measures of smoking intensity, for both former and current smokers) were also statistically significantly associated with lung cancer risk.

Multivariable Risk Models

In the multivariable logistic regression analyses based on the combined training and validation datasets (Table 3), both exposure to ETS and family history of any cancer were statistically significantly associated with lung cancer in never smokers. Among former and current smokers, lung cancer was statistically significantly associated with exposure to dust, no prior history of hay fever (as the risk-conferring value of the variable), personal history of emphysema, family history of any cancer (for former smokers) or tobacco-related cancers (for current smokers), and smoking intensity (for current smokers) and age at smoking cessation (for former smokers). In addition, exposure to asbestos was statistically significantly associated with lung cancer in current smokers but not in former smokers.

We also constructed smoking status–specific risk models stratified by sex (data not shown) and found only a few differences in risk factors among men and women. Specifically, among former smokers, both age at smoking cessation and no prior hay fever were statistically significantly associated with lung cancer risk in men but not in women. Among current smokers, asbestos exposure was statistically significantly associated with lung cancer risk in men but not in women.

The same variables that were associated with lung cancer risk in the logistic regression analysis were also identified in the

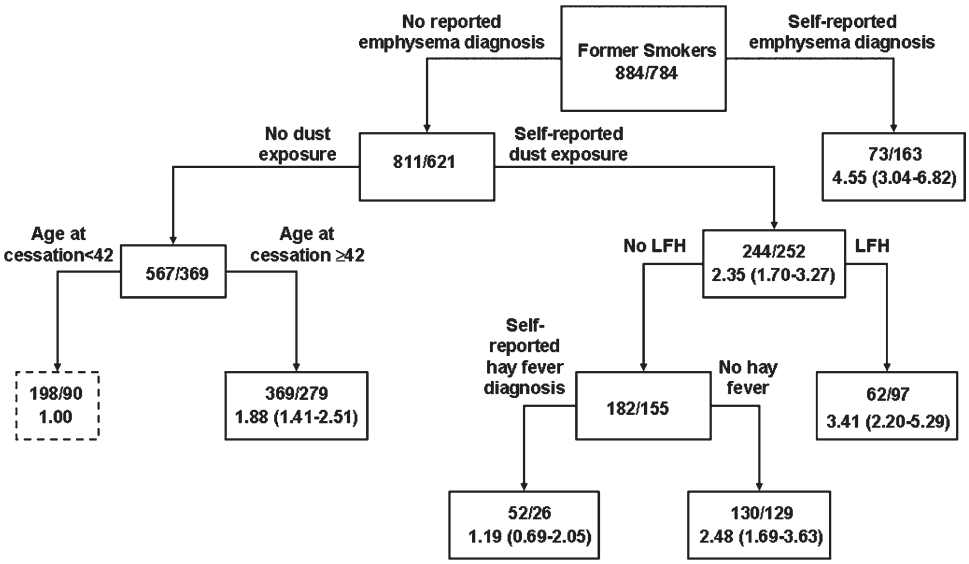

Fig. 1. Classification and regression tree analysis of risk predictors in former smokers. Nodes of the classification tree are formed by recursive splits of lung cancer case/control status by predictor variables. The numbers within each node indicate the number of control subjects/number of case patients. Within each terminal node, the odds ratio (with 95% confidence interval) of lung cancer (adjusted by age and sex) is shown for that node with respect to the reference node (**dotted** terminal node). LFH = lung cancer family history.

decision trees of the CART models (Figs. 1 and 2). No higher order interactions were evident from these models, but we did observe that, among never smokers, both exposure to ETS and family history of cancer were strongly associated with lung cancer risk (data not shown) as defined in our logistic regression modeling. Among former smokers, a history of emphysema was again the strongest risk factor (OR = 4.55, 95% CI = 3.0 to 6.8), followed by dust exposure in those without a history of emphysema (OR = 2.35, 95% CI = 1.7 to 3.3) (Fig. 1). For those former smokers without emphysema or dust exposure, later age at smoking cessation was associated with a 1.88-fold increase in risk (95% CI = 1.4 to 2.5), and the combination of dust exposure and family history of any cancer was associated with an OR of 3.41 (95% CI = 2.2 to 5.3). Among current smokers (Fig. 2), a history of emphysema was the strongest risk factor for lung cancer (OR = 4.20, 95% CI = 2.9 to 6.2), whereas smoking intensity (≥37 pack-years) was strongly associated with lung cancer risk among current smokers without emphysema (OR = 2.55, 95% CI = 1.9 to 3.5). A family history of smoking-related cancers was associated with lung cancer in subjects with heavier smoking histories (OR = 3.09, 95% CI = 2.0 to 4.7), as well as among those with lighter smoking histories and self-reported dust exposure (OR = 5.12, 95% CI = 2.6 to 10.3). Hay fever, which was identified in the backward selection procedures, appeared in a lower (albeit pruned off) branch of the tree, suggesting that the contribution to risk from hay fever was less than that for emphysema and smoking history in this analysis.

Model Validation

We next used a three-phase validation process to assess the performance, in the validation sets, of the models developed in the training sets. As illustrated in Table 4, the risk models were well calibrated throughout the entire range of probabilities, as indicated by the non–statistically significant Hosmer–Lemeshow goodness-of-fit test statistics (0.777 for never smokers, 0.712 for former smokers, and 0.688 for current smokers). The AUC statistic

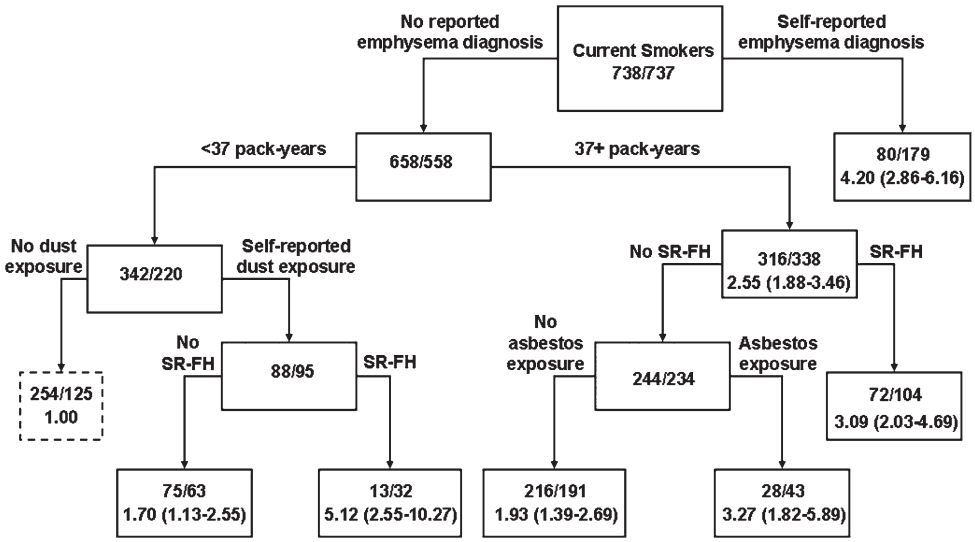

Fig. 2. Nodes of the classification tree are formed by recursive splits of lung cancer case/control status by predictor variables. The numbers within each node indicate the number of control subjects/number of case patients. Within each terminal node, the odds ratio (with 95% confidence interval) of lung cancer (adjusted by age and sex) is shown for that node with respect to the reference node (**dotted** terminal node). SR-FH = family history of smoking-related cancer.

Table 4. Model validation statistics*

Smoking category	*P* from Hosmer–Lemeshow goodness of fit†	AUC† (95% CI)	Concordance statistic‡ (95% CI)
Never smokers	.777	0.57 (0.47 to 0.66)	0.59 (0.51 to 0.67)
Former smokers	.712	0.63 (0.58 to 0.69)	0.63 (0.58 to 0.67)
Current smokers	.688	0.58 (0.52 to 0.64)	0.65 (0.60 to 0.69)

* AUC = area under the curve; CI = confidence interval.

† Derived from validation set.

‡ Derived from threefold cross-validation for combined dataset.

obtained from the validation sets (Table 4) was low for never smokers (AUC = 0.57, 95% CI = 0.47 to 0.66) and current smokers (AUC = 0.58, 95% CI = 0.52 to 0.64) and slightly higher for former smokers (AUC = 0.63, 95% CI = 0.58 to 0.69). The resulting concordance statistics, calculated by threefold cross-validation of the combined dataset, were 0.59, 0.63, and 0.65 for never, current, and former smokers, respectively, indicating that the models performed reasonably well in discriminating between case patients and control subjects (Table 4).

Given the lack of consistency in the literature on the association between hay fever and lung cancer risk, we reran our models excluding hay fever as a covariate and obtained similar results. We also evaluated the fit of the Bach et al. (4) model using our own data in ever smokers (i.e., never smokers were not used to develop the Bach data). The resulting AUC was only 0.57 (95% CI = 0.56 to 0.59), indicating that adding clinical and epidemiologic variables improves the risk prediction.

Estimation of Absolute 1-Year Risk for Lung Cancer

We next used the lung cancer risk model to estimate 1-year absolute risks for lung cancer for three hypothetical individuals at high, moderate, and low risks. The first individual is a 75-year-old white man, a current smoker with a 58–pack-year smoking history (i.e., he smoked approximately 1 pack a day for 60 years), a history of both emphysema and hay fever, two first-degree relatives diagnosed with a smoking-related cancer, and prior asbestos exposure. Based on the model, his estimated relative risk of lung cancer compared with a man of similar age but without these risk factors is 8.75, as derived from multiplying the component risk estimates for each risk factor ($r = 1.85 \times 2.13 \times 1.47 \times 1.51$; see Table 3). The baseline hazard is obtained from the sex/smoking status adjustment constant (3.88 in this case; Appendix Table 1), age-specific SEER (29) lung cancer incidence rates for white men 75–79 years old (564.36 per 100 000; Appendix Table 2), and the attributable risk for men, as derived from the model for current smokers (0.51404), as $h_{113} = 3.88 \times 564.36/100\,000 \times (1 - 0.51404)$. On the basis of NCHS data (32), we estimated the mortality rate from causes other than lung cancer among white men 75–79 years old as $h_{21} = 4836.4/100\,000$. The estimated 1-year absolute risk of lung cancer for this man is calculated as $p = (0.010641148 \times 8.75)/(0.010641148 \times 8.75 + 0.048364) \times \{1 - \exp[-(0.010641148 \times 8.75 + 0.048364)]\} = .0868236$. This risk (8.68%) is more than 15 times that of the age-specific SEER incidence rate for lung cancer in white men (0.56%).

For a white female former smoker, aged 66, who quit smoking at age 54 and had a history of dust exposure but no family history of cancer and no prior hay fever, the estimated relative risk of lung cancer, based on our model is 3.458 ($r = 1.50 \times 1.59 \times 1.45$) times that of a white woman of the same age without those risk factors. The lung cancer incidence rate from SEER (29) for white women 65–69 years old is 246.85/100 000 (Appendix Table 2), the adjustment constant for a female former smoker is 3.76 (Appendix Table 1), and the attributable risk for former smokers from our model is 0.45352, hence the baseline hazard $h_{122} = 3.76 \times 0.002468457 \times (1 - 0.45352)$. The mortality rate from NCHS (32) for white women 65–69 years old from causes other than lung cancer is $h_{22} = 1197/100\,000$. The estimated 1-year absolute risk for lung cancer for this woman is $p = (0.00507210 \times 3.458)/(0.00507210 \times 3.458 + 0.01197) \times \{1 - \exp[-(0.\ 00507210 \times 3.458 + 0.01197)]\} = 0.017284$. This risk (1.70%) is seven times higher than the age-specific SEER (29) incidence of lung cancer for white women (0.24%).

A third example is a white male never smoker, aged 45 years, with no exposure to ETS and no family history of cancer. His estimated relative risk based on our model for never smoker is 1 ($r = 1 \times 1$). The lung cancer incidence rate from SEER (29) for men 45–49 years old is 25.49/100 000 (Appendix Table 2), and the adjustment constant for a male never smoker is 0.21 (Appendix Table 1). The attributable risk for never smokers from our model is 0.4751. The baseline hazard is $h_{111} = 0.21 \times 0.000254856 \times (1 - 0.4751)$. The mortality rate from NCHS (32) for men 45–49 years old from causes other than lung cancer is $h_{21} = 400.7/100\,000$. The estimated 1-year absolute risk for lung cancer for the man is $p = (0.000028108 \times 1)/(0.000028108 \times 1 + 0.004007) \times \{1 - \exp[-(0.000028108 \times 1 + 0.004007)]\} = 0.000028052$. This estimated risk is approximately one-tenth that of the annual SEER (29) age-specific lung cancer incidence rate for white men (0.0028% versus 0.025%, respectively).

Ordinal Risk Index

The clinical utility of a risk prediction tool lies in its value for decision making and ease of use at the individual level. Therefore, to facilitate the use of the model, we developed a way to compute ordinal risk indices from odds ratios derived from the multivariable regression analyses for the statistically significant risk factors from each model (Table 3). In the absence of a particular risk factor, a baseline integer of 1 is assigned to the score. For ordinal variables (pack-years, years since cessation), the lowest category is assigned an integer of 1. We evaluated both additive and multiplicative scoring approaches; the results were nearly identical, and we present only the additive model for simplicity (Table 5). Based on CART analysis, we established three levels of risk for each smoking

Table 5. Assignment of case patients and control subjects to risk score categories

	Training set		Validation set		Combined	
Score*	**Case patients**	**Control subjects**	**Case patients**	**Control subjects**	**Case patients**	**Control subjects**
Former smoker, n (%)						
<5.9 (low)	78 (13.8)	186 (28.1)	74 (38.5)	144 (65.5)	188 (24.9)	368 (41.8)
5.9–6.9 (medium)	289 (51.2)	372 (56.3)	75 (39.1)	60 (27.3)	372 (49.2)	430 (48.8)
≥7 (high)	197 (34.9)	103 (15.6)	43 (22.4)	16 (7.3)	196 (25.9)	83 (9.4)
Current smoker, n (%)						
<6.9 (low)	101 (18.9)	233 (42.3)	35 (19.4)	66 (35.9)	136 (19.0)	286 (38.9)
6.9–7.9 (medium)	235 (43.9)	241 (43.7)	117 (65.0)	105 (57.1)	371 (51.9)	354 (48.2)
≥8 (high)	199 (37.2)	77 (14.0)	28 (15.6)	13 (7.1)	208 (29.1)	95 (12.9)

* Risk scores are calculated by summing the odds ratios from the multivariable model in Table 3 for any reported risk factor. The cut points for three categories of risk for each smoking status category were defined by classification and regression tree analysis.

category. Because of the small number of never smokers in the validation set, the results in Table 5 are presented for former and current smokers only in both the test and validation sets and for the combined datasets. The percentage of control subjects in the high-risk category for the combined dataset was 9.4% for former smokers and 12.9% for current smokers (Table 5). Among former smokers, the true-negative rates for the lowest risk subgroup designation were 66% (95% CI = 59% to 72%) for the validation set and 66% (95% CI = 62% to 70%) in the combined analysis. For current smokers, the true-negative rates in the low-risk categories were 65% (95% CI = 55% to 75%) and 68% (95% CI = 63% to 72%) for the validation and combined sets, respectively. The true-positive rates for the high-risk group were 73% (95% CI = 60% to 84%) and 70% (95% CI = 65% to 76%) for former smokers in the validation and combined sets, respectively. Among current smokers, the true-positive rates were 68% (95% CI = 52% to 82%) and 69% (95% CI = 63% to 74%) in the validation and combined sets, respectively. We also compared the true-positive and true-negative rates for the training, validation, and combined datasets; in all instances, the 95% confidence intervals were overlapping, indicating concordance in the results among the different sets (data not shown).

We used the risk scenarios cited above to illustrate the usefulness of the easy-to-compute risk indices in Table 5 to classify subjects into three risk groups (low, intermediate, or high) for each smoking stratum–specific model. From the first example above, the current smoker's risk score of 8.96 (=1.85 + 2.13 + 1.47 + 1.51 + 1 + 1) would place him in the high-risk group for current smokers. In the second example, the former smoker's risk score of 6.54 (=1.50 + 1.59 + 1.45 + 1 + 1) would classify her in the intermediate risk group of former smokers.

Discussion

We used existing data from a large, ongoing lung cancer case–control study to develop and internally validate separate risk prediction models for never, former, and current smokers. The models are derived from a large case–control study, and, in addition to smoking variables, they also incorporate other epidemiologic and clinical risk factors. Moreover, we constructed independent training and validation sets to avoid any potential for overfitting of the models. For never smokers, the best model included exposure on a regular basis to ETS and family history of cancer in two or more first-degree relatives. For former smokers, the best model included emphysema, no prior hay fever, dust exposure, and family history of cancer in two or more first-degree relatives. For current smokers, the best model also included asbestos exposure, and the family history variable was limited to one or more first-degree relatives with a smoking-related cancer. We computed absolute risks of lung cancer and presented calculations for varying risk profiles, demonstrating that a currently smoking man with an estimated relative risk close to 9 had an 8.68% 1-year absolute risk of lung cancer, compared with a 0.56% annual incidence rate for his age group. Such a degree of risk would justify more intensive surveillance and counseling. A similar approach could be used to compute 5- and 10-year absolute risks. Finally, we have presented a simplified numerical score that can be easily used in the clinical setting.

Bach et al. (33) demonstrated that their model (4) performed very well in the validation analysis of three different cohorts. Their model incorporates smoking intensity (number of years smoked, number of cigarettes smoked per day, years since quitting), sex, and asbestos exposure. However, they also cautioned that their model is based on a cohort that was assembled and enrolled in the late 1980s and that it has not been sufficiently validated in women.

All the variables in our models have strong, biologically plausible etiologic roles in lung cancer that are supported by published findings from our own and numerous other case–control and cohort studies of lung cancer. Several dusts and fibers are classified by the International Agency for Research on Cancer (IARC) as human carcinogens (34), and we have previously reported that dust exposure is statistically significantly associated with lung cancer risk in a subset of the population analyzed in the current study (35). Inhalation of particulate irritants such as cigarette smoking or dust causes lung inflammation, which is characterized by tissue destruction, altered vasculature, airway remodeling, and impaired wound healing (36). An inflammatory microenvironment, with a continual cycle of injury and repair and generation of both reactive oxygen and nitrogen species, promotes genotypic and phenotypic changes that lead to malignancy (37). Previous studies (38–41), including our own (42), have reported increased lung cancer risks associated with a prior diagnosis of emphysema, although not all studies found statistically significant associations. Prospective studies have also shown that lung function tests predict future lung cancer risk (reviewed in 43).

Another risk factor that was identified in this analysis was no prior history of hay fever. We previously reported the existence of an inverse association between prior history of hay fever and lung cancer risk (42). However, the association between hay fever and lung cancer is still considered controversial, and there are two distinct and contradictory hypotheses for the association (44). Studies demonstrating an inverse association suggest that the enhanced immune surveillance that characterizes hay fever results in stimulated immune systems that are better at detecting and destroying malignant cells (44–48). Conversely, other studies suggest that chronic immune stimulation leads to random pro-oncogenic mutations in actively dividing stem cells and an increased risk of cancer (44,49). A review of the literature on atopy and cancer risk (50) suggests that more studies (n = 11) have reported inverse associations than have reported no association (n = 3) or increased risk (n = 2). Inverse associations between eczema and other allergic skin conditions and lung cancer risk have also been noted (46,47,51,52). Even though the weight of evidence is in favor of an inverse association, we reran our analyses excluding hay fever and noted a lower AUC in the model without hay fever compared to the model we presented with hay fever. Therefore, although the inclusion of hay fever may be controversial, we find it to be useful in lung cancer risk prediction.

A family history of cancer or lung cancer was also a statistically significant predictor of lung cancer risk in this study. A number of other studies (53–63) have shown that a first-degree family history of lung cancer is associated with lung cancer risk. This association could be explained by shared genes, shared smoking patterns, or both. In a previous analysis of data from the same population as analyzed here, we reported an increased risk of lung and other smoking-related cancers among first-degree relatives of lung cancer patients after adjustment for smoking behaviors in both patients and relatives (64), thereby providing evidence for the contribution of both exposure and genetic susceptibility to risk.

All variables included in these predictive models are easily ascertained by a health care provider. Risk prediction models have two applications: to help design prevention trials and to discriminate among individuals of different risks. For example, a prevention trial that enrolls only high-risk smokers or former smokers could achieve statistical power equivalent to a trial that enrolls all smokers or former smokers, but requiring a larger trial size or longer duration. As far as individual risk prediction goes, for any smoker or former smoker, there is substantial interindividual variability in susceptibility to tobacco carcinogenesis. In this context, the discriminatory ability of a risk prediction model (as quantified by calculating the concordance statistic) is most important for clinical decision making (65). Our concordance statistics were modest, in the upper 0.6 range, although they are in line with those of other prediction models. For example, Cronin et al. (66) assessed the validity of the Bach et al. (4) model using the placebo arm of another chemoprevention trial with different entry criteria and surveillance regimen and found that the overall concordance index was 0.69, with age-specific concordance indices ranging from 0.57 to 0.77. Likewise, validation of the Gail model showed similarly modest predictive accuracy, with a concordance index of 0.67, 95% CI = 0.65 to 0.68 (13). Chen et al. (67) reported an age-specific concordance for the Gail model of only 0.596, compared with 0.643 with addition of mammographic density to improve the discriminatory power of the model. Another recent breast validation study reported concordance statistics of 0.631 and 0.624 for premenopausal and postmenopausal women, respectively (68). The discriminatory accuracy for models to predict melanoma has ranged from 0.62 (17) to 0.70 for women 50 years and older to 0.80 for men aged 20–49 years (18). An ovarian cancer risk model had a concordance statistic of 0.60 (19). As Cronin et al. (66) point out, the relatively low discriminatory ability of these models reflects the inherent challenges in predicting risk, even when there are well-established and quantifiable risk factors such as with lung cancer.

This study has several limitations. Our prediction tool is based on relative risk estimates that were derived from a single, albeit large, case–control study, in which the case patients were recruited from a single tertiary cancer center and the control group was not population based. We were restricted to this design because our case–control study mandates enrollment of patients before therapy is initiated. Likewise, we acknowledge the tradeoff we have made between classical epidemiologic rigor and feasibility in the selection of control subjects. Nevertheless, because M. D. Anderson Cancer Center serves as a referral center for many cancer patients from the Kelsey–Seybold system, the case patients are likely to come from a population base similar to that of the control subjects, especially because all participants had to be residents of Texas. Another limitation is the fact that we only used data from non-Hispanic whites to construct the models. Therefore, the models may not be applicable to other ethnic groups.

Other potential limitations include recall and reporting bias, especially of prior medical conditions. Any misclassification of self-reported physician-diagnosed conditions could be a concern because we did not validate the medical conditions. However, some of the important risk factors for lung cancer, such as inflammatory processes, are not well known, and thus, differential recall between patients and control subjects is unlikely. Furthermore, our prevalence data for hay fever in control subjects are consistent with national statistics. Selection bias is a possibility; it could be argued that participants who declined to participate had different prevalence rates of lung diseases.

Finally, our study population was matched on age and smoking status, and so the overwhelming contribution of age and smoking to lung cancer risk were somewhat masked. However, we have attempted to incorporate smoking status into our absolute risk estimates by adjusting baseline incidence rates to account for smoking status. External validation of our models in independent populations remains an important next step.

The purpose of this analysis was to create a parsimonious model for assessing lung cancer risk with a minimal number of risk predictors that is realistic to use in clinical practice and to validate the model in an independent sample from the same population. In our experience, patients are agreeable to completing health questionnaires, either self-administered or administered by personal interview. We plan next to incorporate pathway-based gene variation data in the model to acknowledge the important contribution of genetic susceptibility to lung cancer risk. Adding such data is likely to further improve the sensitivity and specificity of the models, although incorporating genetic data may not be practicable for community-based settings.

Appendix Table 1. Adjustment constants* (c_{ji}) to estimate smoking status–specific incidence rates

Sex	Never smoker	Former smoker	Current smoker
Male	0.21	3.17	3.88
Female	0.34	†3.76	4.17

* Adjustment constants (c_{ji}, j = 1 male, j = 2 female; i = 1 never smoker, i = 2 former smoker, i = 3 current smoker) computed based on the following prevalence estimates: According to data from the Office on Smoking and Health, 2004, 90% of all new male lung cancer cases and 80% of all new female lung cancer cases occur in ever smokers (30); in 2005, 23.2% of adult men and 19.2% of adult women were current smokers, 48.4% of men and 59.5% of women were never smokers, and therefore 28.4% of men and 21.3% of women were former smokers (31).

† Therefore, for a female former smoker, the constant is derived from the ratio of the proportion of all new lung cancer cases in ever-smoking women (0.80) to the proportion of women former smokers in the population at risk (0.213), i.e., c_{22} = 0.80/0.213 = 3.76.

Appendix Table 2. Lung cancer incidence rates/100 000 and overall mortality/100 000 (excluding lung cancer) by age and sex (whites only)

	Men		Women	
Age (y)	Incidence*	Mortality†	Incidence*	Mortality†
40–44	10.78	275.1	11.03	153.20
45–49	25.49	400.7	23.19	218.80
50–54	56.60	560.0	45.51	313.40
55–59	116.58	786.9	93.93	479.10
60–64	221.18	1210.2	164.9	762.90
65–69	346.77	1855.1	246.85	1197.00
70–74	478.10	2947.4	318.69	1968.30
75–79	564.36	4836.4	344.67	3306.10
80–84	532.36	7980.7	308.28	5761.20
≥85	498.44	15559.4	266.72	14016.2

* Lung cancer incidence data derived from Surveillance, Epidemiology, and End Results, 2005 (29).

† Mortality data from all causes excluding lung cancer derived from the National Center for Health Statistics for whites only for 1999–2003 (32).

References

(1) Shopland DR, Eyre HJ, Pechacek TF. Smoking-attributable cancer mortality in 1991: is lung cancer now the leading cause of death among smokers in the United States? J Natl Cancer Inst 1991;83:1142–8.

(2) Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. Br Med J 2000;321:323–9.

(3) American Cancer Society. Cancer facts and figures, 2006. Atlanta (GA): American Cancer Society; 2007.

(4) 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.

(5) Chambless LE, Dobson AJ, Patterson CC, Raines B. On the use of a logistic risk score in predicting risk of coronary heart disease. Stat Med 1990;9:385–96.

(6) Grundy SM, Balady GJ, Criqui MH, Fletcher G, Greenland P, Hiratzka LF, et al. Primary prevention of coronary heart disease: guidance from Framingham. Circulation 1998;97:1876–87.

(7) National Cancer Institute. The nation's investment in cancer research. A plan and budget proposal for the fiscal year 2006. Available at: http://plan.cancer.gov. [Last accessed: July 2006.]

(8) Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879–86.

(9) Rockhill B, Byrne C, Rosner B, Louie MM, Colditz G. Breast cancer risk prediction with a log-incidence model: evaluation of accuracy. J Clin Epidemiol 2003;56:856–61.

(10) Tice JA, Cummings SR, Ziv E, Kerlikowske K. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. Breast Cancer Res Treat 2005;94:115–22.

(11) Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 2001;93:358–66.

(12) Taplin SH, Thompson RS, Schnitzer F, Anderman C, Immanuel V. Revisions in the risk-based Breast Cancer Screening Program at Group Health Cooperative. Cancer 1991;67:2400.

(13) Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004;23:1111–30.

(14) Wijnen JT, Vasen HFA, Khan M, Zwinderman AH, Van Der Klift H, Mulder A, et al. Clinical findings with implications for genetic testing in families with clustering of colorectal cancer. N Engl J Med 2006;339:511–8.

(15) Selvachandran SN, Hodder RJ, Ballal MS, Joes P, Cade D. Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. Lancet 2002;360:278–83.

(16) Imperiale TF, Wagner DR, Lin CY, Larkin GN, Rogge JD, Ransohoff DF. Using risk for advanced proximal colonic neoplasia to tailor endoscopic screening for colorectal cancer. Ann Intern Med 2003;139:959–65.

(17) Cho E, Rosner BA, Feskanich D, Colditz GA. Risk factors and individual probabilities of melanoma for whites. J Clin Oncol 2005;23:2669–75.

(18) Fears TR, Guerry D IV, Pfeiffer RM, Sagebiel RW, Elder DE, Halpern A, et al. Identifying individuals at high risk of melanoma: a practical predictor of absolute risk. J Clin Oncol 2006;24:3590–95.

(19) Hartge P, Whittemore AS, Itnyre J, McGowan L, Cramer D. Rates and risks of ovarian cancer in subgroups of white women in the United States. The collaborative Ovarian Cancer Group. Obstet Gynecol 1994;84:760–4.

(20) Eastham JA, May R, Robertson JL, Sartor O, Kattan MW. Development of a nomogram that predicts the probability of a positive prostate biopsy in men with an abnormal digital rectal examination and a prostate-specific antigen between 0 and 4 ng/ml. Urology 1999;54:703–13.

(21) Colditz GA, Atwood KA, Emmons K, Monson RR, Willett WC, Trichopoulos D, Hunter DJ. Harvard report on cancer prevention volume 4: Harvard cancer risk index. Risk Working Group, Harvard Center for Cancer Prevention. Cancer Causes Control 2000;11:477–88.

(22) Wu X, Zhao H, Amos CI, Shete S, Makan N, Hong WK, et al. p53 genotypes and haplotypes associated with lung cancer susceptibility and ethnicity. J Natl Cancer Inst 2002;94:681–90.

(23) Hudmon KS, Honn SE, Jiang H, Chamberlain RM, Xiang W, Ferry G, et al. Identifying and recruiting healthy control subjects from a managed care organization: a methodology for molecular epidemiological case-control studies of cancer. Cancer Epidemiol Biomarkers Prev 1997;6:565–71.

(24) Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models. New York (NY): WCB McGraw-Hill; 1996.

(25) Breiman L, Friedman JN, Olsen RA, Stone CJ. Classification and regression trees. Monterey (CA): Wadsworth and Brooks/Cole; 1984.

(26) Therneau TM, Atkinson B. Technical report series no. 61, an introduction to recursive partitioning using the RPART routines. Rochester (MN): Department of Health Science Research, Mayo Clinic; 1997.

(27) Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. Crit Care 2004;8:508–12.

(28) Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B 1974;36:111–47.

(29) SEER-Scientific Systems. National Cancer Institute, 2005. Available at: http://www.seer.cancer.gov. [Last accessed: July 2006.]

(30) U.S. Department of Health and Human Services. The health consequences of smoking: a report of the Surgeon General. Rockville (MD): U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2004.

(31) Centers for Disease Control. Cigarette smoking among adults—United States, 2004. Morb Mortal Wkly Rep 2005;54:1121–4.

(32) Worktable 210R. Death rates for 113 selected causes, alcohol-induced causes, drug-induced causes and injury by firearms, by 5-year age groups, race, and sex: United States, 2003. National Center for Health Statistics. Available at: http://www.cdc.gov/nchs/datawh/statab/unpubd/mortabs/gmwk210_10.htm. [Last accessed: July 2006.]
(33) Bach PB, Elkin EB, Pastorino U, Kattan MW, Mushlin AI, Begg CB, et al. Benchmarking lung cancer mortality rates in current and former smokers. Chest 2004;126:1742–9.
(34) International Agency for Research on Cancer. Wood dust and formaldehyde. IARC monographs on the evaluation of carcinogenic risks to humans. Vol 62. Lyon (France): IARC; 1997.
(35) Barcenas CH, Delclos GL, El-Zein R, Tortolero-Luna G, Whitehead LW, Spitz MR. Wood dust exposure and the association with lung cancer risk. Am J Ind Med 2005;47:349–57.
(36) Martey CA, Pollock SJ, Turner CK, O'Reilly KMA, Baglole CJ, Phipps RP, et al. Cigarette smoke induces cyclooxygenase-2 and microsomal prostaglandin E2 synthase in human lung fibroblasts: implications for lung inflammation and cancer. Am J Physiol Lung Cell Mol Physiol 2004;287:981–91.
(37) Ames BN, Shigenaga MK, Gold LS. DNA lesions, inducible DNA repair, and cell division: three key factors in mutagenesis and carcinogenesis. Environ Health Perspect 1993;101(Suppl):535–44.
(38) Mayne ST, Buenconsejo J, Janerich DT. Previous lung disease and risk of lung cancer among men and women nonsmokers. Am J Epidemiol 1999;149:13–20.
(39) Alavanja MCR, Brownson RC, Boice JD Jr, Hock E. Preexisting lung disease and lung cancer among nonsmoking women. Am J Epidemiol 1992;136:623–32.
(40) Wang SY, Hu YL, Wu YL, Li X, Chi GB, Chen Y, et al. A comparative study of the risk factors for lung cancer in Guangdong, China. Lung Cancer 1996;1:S99–105.
(41) Brenner AV, Wang Z, Kleinerman RA, Wang L, Zhang S, Metayer C, et al. Previous pulmonary diseases and risk of lung cancer in Gansu Province, China. Int J Epidemiol 2001;1:118–24.
(42) Schabath MB, Delclos GL, Martynowicz MM, Greisinger AJ, Lu C, Wu X, et al. Opposing effects of emphysema, hay fever, and select genetic variants for lung cancer risk. Am J Epidemiol 2005;161:412–22.
(43) Littman AJ, Thornquist MD, White E, Jackson LA, Goodman GE, Vaughan TL. Prior lung disease and risk of lung cancer in a large prospective study. Cancer Causes Control 2004;15:819–27.
(44) Talbot-Smith A, Fritschi L, Divitini ML, Mallon DF, Knuiman MW. Allergy, atopy, and cancer: a prospective study of the 1981 Busselton cohort. Am J Epidemiol 2003;157:606–12.
(45) Cockcroft DW, Klein GJ, Donevan RE, Copland GM. Is there a negative correlation between malignancy and respiratory atopy? Ann Allergy 1979;43:345–7.
(46) Vena JE, Bona JR, Byers TE, Middleton E Jr, Swanson MK, Graham S. Allergy-related diseases and cancer: an inverse association. Am J Epidemiol 1985;122:66–74.
(47) Gabriel R, Dudley BM, Alexander WD. Lung cancer and allergy. Br J Clin Pract 1972;6:202–4.
(48) McDuffie HH. Atopy and primary lung cancer. Histology and sex distribution. Chest 1991;99:404–7.
(49) McWhorter WP. Allergy and risk of cancer. A prospective study using NHANESI followup data. Cancer 1988;62:451–5.
(50) Wang H, Diepgen TL. Is atopy a protective or a risk factor for cancer? A review of epidemiological studies. Allergy 2005;60:1098–111.
(51) Castaing M, Youngson J, Zaridze D, Szeszenia-Dabrowska N, Rudnai P, Lissowska J, et al. Is the risk of lung cancer reduced among eczema patients? Am J Epidemiol 2005;162:542–7.
(52) McDuffie HH, Cockcroft DW, Talebi Z, Klaassen DJ, Dosman JA. Lower prevalence of positive atopic skin tests in lung cancer patients. Chest 1988;93:241–6.
(53) Tokuhata GK, Lilienfeld AM. Familial aggregation of lung cancer in humans. J Natl Cancer Inst 1963;30:289–312.
(54) Ooi WL, Elston RC, Chen VW, Bailey-Wilson JE, Rothschild H. Increased familial risk for lung cancer. J Natl Cancer Inst 1986;76:217–222.
(55) Samet J, Humble C, Pathak D. Personal and family history of respiratory disease and lung cancer risk. Am Rev Respir Dis 1986;134:466–70.
(56) Wu AH, Yu MC, Thomas DC, Pike MC, Henderson BE. Personal and family history of lung disease as risk factors for adenocarcinoma of the lung. Cancer Res 1988;48:7279–84.
(57) Shaw GL, Falk RT, Pickle LW, Mason TJ, Buffler PA. Lung cancer risk associated with cancer in relatives. J Clin Epidemiol 1991;44:429–37.
(58) Osann KE. Lung cancer in women: the importance of smoking, family history of cancer, and medical history of respiratory diseases. Cancer Res 1991;51:4893–7.
(59) Schwartz AG, Yang P, Swanson GM. Familial risk of lung cancer among nonsmokers and their relatives. Am J Epidemiol 1996;144:554–62.
(60) Wu AH, Fontham ET, Reynolds P, Greenberg RS, Buffler P, Liff J, et al. Family history of cancer and risk of lung cancer among lifetime nonsmoking women in the United States. Am J Epidemiol 1996;143:535–42.
(61) Mayne ST, Buenconsejo J, Janerich DT. Familial cancer history and lung cancer risk in United States nonsmoking men and women. Cancer Epidemiol Biomarkers Prev 1999;8:1065–9.
(62) Bromen K, Pohlabeln H, Ingeborg J, Ahrens W, Jockel K-H. Aggregation of lung cancer in families: results from a population-based case-control study in Germany. Am J Epidemiol 2000;152:497–505.
(63) Brownson RC, Alavanja MCR, Caporaso N, Berger E, Chang JC. Family history of cancer and risk of lung cancer in lifetime non-smokers and long-term ex-smokers. Int J Epidemiol 1997;26:256–63.
(64) Etzel CJ, Amos CI, Spitz MR. Risk for smoking-related cancer among relatives of lung cancer patients. Cancer Res 2003;63:8531–5.
(65) Freedman AN, Seinara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. J Natl Cancer Inst 2005;97:715–23.
(66) 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.
(67) Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, et al. Projecting absolute invasive breast cancer risk in white women: a model that includes mammographic density. J Natl Cancer Inst 2006;98:1215–26.
(68) Barlow W, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 2006;98:1204–14.

Notes

This study was supported by the Flight Attendant Medical Research Institute and Public Health Service grants CA55769, CA070907, CA093592, and CA016672 from the National Cancer Institute, National Foundation for Cancer Research, and DAMD17-02-1-0706 (TARGET), a grant from the Department of Defense to Dr W. K. Hong. The authors had full responsibility for the design of the study, the analysis and interpretation of the data, the decision to submit the study for publication, and the writing of the manuscript.

We thank Drs Joe Ensor (Department of Biostatistics, M. D. Anderson Cancer Center) and Eric A. Engels (Division of Cancer Epidemiology and Genetics, National Cancer Institute) for their valuable comments and suggestions.

Manuscript received October 5, 2006; revised February 13, 2007; accepted March 9, 2007.