Published in final edited form as:
Cancer Prev Res (Phila). 2008 September ; 1(4): 250–254. doi:10.1158/1940-6207.CAPR-08-0060.

An Expanded Risk Prediction Model for Lung Cancer

Margaret R. Spitz[1], **Carol J. Etzel**[1], **Qiong Dong**[1], **Christopher I. Amos**[1], **Qingyi Wei**[1], **Xifeng Wu**[1], and **Waun Ki Hong**[2]

[1]Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas

[2]Division of Cancer Medicine, The University of Texas M. D. Anderson Cancer Center, Houston, Texas

Abstract

Risk prediction models are useful in clinical decision making. We have published an internally validated prediction tool for lung cancer based on easily obtainable epidemiologic and clinical data. Because the precision of the model was modest, we now estimate the improvement obtained by adding two markers of DNA repair capacity.

Assay data (host-cell reactivation and mutagen sensitivity) were available for 725 White lung cancer cases and 615 controls, all former or current smokers, a subset of cases and controls from the previous analysis. Multivariable models were constructed from the original variables with addition of the biomarkers separately and together. Pairwise comparisons of the area under the receiver operating characteristic curves (AUC) and 3-fold cross-validations were done.

For former smokers, the AUC and 95% confidence intervals were 0.67 (0.63–0.71) for the baseline model and 0.70 (0.66–0.74) for the expanded model. For current smokers, the comparable AUC values were 0.68 (0.64–0.72) and 0.73 (0.69–0.77). For both groups, the expanded models were statistically significantly better than the baseline models ($P = 0.006$ and $P = 0.0048$, respectively), although the increases in the concordance statistics were modest. We also recomputed 1-year absolute risks of lung cancer as described previously for two different risk profiles and showed that individuals who exhibited poor repair capacity or heightened mutagen sensitivity had increased absolute risks of lung cancer.

Addition of biomarker assays improved the sensitivity of the expanded models.

We recently constructed and internally validated clinical tools for lung cancer risk prediction for current, former, and never smokers based on epidemiologic and clinical data derived from 1,851 non-Hispanic White lung cancer patients and 2,001 matched controls (1). Variables included were environmental tobacco smoke (for never smokers only), family cancer history, dust exposure, prior respiratory disease and hay fever, and smoking history variables (pack-years for current smokers and age stopped smoking for former smokers). All have strong biologically plausible etiologic roles and are relatively easy to ascertain through patient interview. The validated concordance statistics for the former and current smoker models were modest (0.63 and 0.58, respectively), although consistent with those from other risk prediction models. The purpose of this new analysis was to improve the precision of the models by incorporating select markers of host DNA repair capacity into the prediction tool.

Requests for reprints: Margaret R. Spitz, Department of Epidemiology, Unit 1340, The University of Texas M. D. Anderson Cancer Center, P.O. Box 301439, Houston, TX 77230-1439. Phone: 713-792-3020; Fax: 713-745-1165; E-mail: mspitz@mdanderson.org.

Disclosure of Potential Conflicts of Interest No potential conflicts of interest were disclosed.

Deficiency in DNA damage recognition and repair leads to destabilization of the genome and increased risk of mutagenesis. Previously, we have shown that suboptimal DNA repair capacity (defined as below the median percent repair in the control population, as measured by an *in vitro* lymphocyte culture–based host-cell reactivation assay) is associated with up to 2-fold statistically significant increased lung cancer risks (2,3). Likewise, the *in vitro* mutagen sensitivity assay quantifies chromatid breaks induced by bleomycin as an indirect reflection of repair ability. Higher bleomycin sensitivity (median breaks per cell) is associated with 1.6- to 1.9-fold lung cancer risks with evidence of a dose-response relationship ($P_{\text{trend}} < 0.001$; refs. 3,4). The risk estimate for heightened sensitivity for both markers combined is almost 3-fold (3).

Patients and Methods

This expanded analysis is based on data from 725 White non-Hispanic cases and 615 controls, a subset of those used to construct the original risk model (1), and for whom complete epidemiologic and assay data were available. Due to the small number of never smokers overall and especially with assay data, this analysis is restricted to current and former smokers. As previously described (1,2), the case patients are all newly diagnosed histologically confirmed lung cancer patients enrolled before initiation of chemotherapy or radiation therapy, with no restriction on age, stage, or histology. Healthy control subjects were recruited from a multispecialty physician group and frequency matched to the case patients by age (±5 y), sex, ethnicity, and smoking status (never, former, or current). An individual who had smoked at least 100 cigarettes in his or her lifetime but quit >12 mo before lung cancer diagnosis (for case patients) or before the interview (for control subjects) is considered a former smoker. Current smokers include those currently smoking and "recent quitters" [i.e., those who quit <12 mo 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. The study was approved by institutional review boards at M. D. Anderson and Kelsey-Seybold Clinics. All subjects provided written informed consent for participation.

The assays are done blinded to case-control status and in batches. They are run consecutively as cases and controls are enrolled into the study and there should not be selection bias in availability of assay data. Because control selection lags case selection, there are fewer controls than cases with completed assay data. The host-cell reactivation *in vitro* assay measures the activity of a reactivated reporter gene in cells transfected with benzo(*a*)diol epoxide–treated plasmids (5). Because a single unrepaired DNA adduct can effectively block transcription (6), any activity reflects the ability of the host cells to remove benzo(*a*)diol epoxide–induced adducts from the plasmids. Mutagen sensitivity is measured by quantifying the chromatid breaks induced by a challenge mutagen (in this instance, bleomycin) in short-term lymphocyte cultures, averaged to the number of breaks per cell (7).

Statistical analyses were done using SAS software version 9.1 (SAS Institute, Inc.). Pearson's χ^2 test was used to assess the differences in categorical variables, and Student's *t* test for continuous variables between cases and controls. All tests were two sided. For the risk models for former and current smokers, we retained the original epidemiologic variables and added the biomarker variables into the models separately and together. We performed pairwise comparisons (8) of the area under the receiver operator characteristic curve (AUC) for (*a*) the baseline multiple logistic model, (*b*) the multiple logistic model with the DNA repair capacity variable added, (*c*) the multiple logistic model with bleomycin sensitivity added, and (*d*) the multiple logistic model with both markers incorporated. We also conducted 3-fold cross-validation by randomly dividing data into three equally sized groups and building four multiple logistic risk models (as described above) using two groups. For each risk model, we then used the remaining data group to calculate the AUC. We repeated this process thrice and calculated the mean AUC, $\bar{C}$, and its associated SD, $S_{\bar{C}}$, across the cross-validations. We then calculated 95% confidence intervals (95% CI) as for each model $\bar{C}$, ± 1.96$S_{\bar{C}}$.

Results

Table 1 summarizes the demographic details for the subjects by smoking status. Current smokers (350 cases, 244 controls) were well matched on age and gender. Among former smokers (375 cases, 371 controls), controls were, on average, 2 years younger than the case patients but well within the 5-year age matching criterion mandated by the study design. Case patients predictably were significantly heavier smokers that their respective control subjects. In former smokers, the mean DNA repair capacity (%) for the case patients was 8.2 (±2.6), compared with 9.0 (±3.5) for the control subjects ($P < 0.001$). The mean bleomycin sensitivity breaks per cell data were 0.8 (±0.4) and 0.7 (±0.4), respectively, for the case patients and control subjects ($P < 0.001$). The comparable data for current smokers were 8.3 (±2.9) and 9.2 (±3.6) for DNA repair capacity ($P = 0.002$) and 0.7 (±0.4) versus 0.6 (±0.3) for bleomycin sensitivity ($P < 0.001$).

Table 2 summarizes the results of the multivariable logistic regression analyses for former and current smokers separately. Although the data were derived from only a subset of subjects included in the original analysis, the same panel of risk factors identified from the univariate analysis were statistically significant in these multivariable logistic models. Specifically, for former smokers, lung cancer was statistically significantly associated with personal history of emphysema [odds ratio (OR), 2.08], exposure to dust (OR, 1.64), family history of any cancer (OR, 1.71), age at smoking cessation (OR, 2.61 for those who quit after age 54), and no prior history of hay fever (OR, 1.69). For current smokers, the ORs were 2.48 for emphysema, 1.68 for the heaviest smoking category, 2.02 for smoking-related family history, and 1.74 for asbestos exposure. The risk estimates for dust exposure (OR, 1.39) and no prior hay fever (OR, 1.51) were only of borderline significance ($P = 0.09$ and $P = 0.08$, respectively).

For former smokers, the OR (95% CI) for DNA repair capacity (%) as a continuous variable was 1.11 (1.06-1.17). For mutagen sensitivity (breaks per cell), the OR was 1.87 (1.24-2.82). The comparable risk estimates for current smokers were 1.07 (1.02-1.14) and 4.33 (2.45-7.66), respectively. There was only a weak inverse correlation between DNA repair capacity and bleomycin sensitivity that achieved statistical significance only in currently smoking control subjects (data not shown), but this correlation did not lead to a significant colinearity or affect the overall model fit.

For former smokers, the AUC (95% CI) was 0.67 (0.63-0.71) for the model with the epidemiologic variables, 0.69 (0.65-0.73) with the addition of DNA repair capacity, 0.68 (0.64-0.72) for bleomycin, and 0.70 (0.66-0.74) when both assay data were added to the epidemiologic model (Table 3). For current smokers, the AUC for the baseline model was 0.68 (0.64-0.72) compared with 0.69 (0.64-0.73) with DNA repair capacity added, 0.72 (0.68-0.77) with only bleomycin sensitivity included, and 0.73 (0.69-0.77) with incorporation of both assays. Pairwise comparisons of the receiver operator characteristic curves showed that the expanded models incorporating both markers were statistically significantly better than the baseline models for both current and former smokers ($P = 0.006$ and $P = 0.0048$, respectively; data not shown).

The mean AUC for the overall expanded model derived from the 3-fold cross-validation showed good calibration as indicated by nonstatistically significant Hosmer-Lemeshow goodness of fit test statistics (Table 3). The cross-validation AUC values were 0.70 (0.63-0.77) and 0.68 (0.62-0.75) for current and former smokers, respectively.

In our previous article, we computed 1-year absolute risks for lung cancer using national incidence and mortality data for hypothetical scenarios with extreme risk profiles (1). For example, a currently smoking 75-year-old White man, with a 58-pack-year smoking history, report of physician-diagnosed emphysema, one or more first-degree relatives diagnosed with a smoking-related cancer, and prior asbestos exposure, had a 1-year absolute risk (compared with a man of similar age, but without these risk factors) of 8.7% (Table 4). If, in addition, he exhibited suboptimal DNA repair capacity, his 1-year absolute risk was increased to 8.9%. When we added the mutagen-sensitive phenotype into the equation, his risk almost doubled to 16.3%. Likewise, for a female former smoker, age 66 y, who quit smoking at age 54 y, reported regular exposure to dusts, but denied any family cancer history, her baseline 1-year risk was 1.2% (Table 4). Adding poor repair capacity yielded an estimated 1-year risk of 2.3%. Together with mutagen sensitivity, this risk doubled from baseline to 3.1%.

Discussion

Our expanded models that incorporate DNA repair capacity and mutagen sensitivity data performed better than the baseline data-driven models. The concordance statistics (i.e., the ability to discriminate between cases and controls) were 70% for current smokers and 68% for former smokers. These statistics, although modest, compare favorably with the Gail model for breast cancer (*C*-statistic of 0.67; ref. 9).

The improvement in AUC we reported when incorporating the additional markers into the model (Table 3) is modest in absolute terms. However, it is instructive to compare the improvement in the Gail model when mammographic density, a known risk factor for breast cancer, was added to the model. Improvement in the *C*-statistic with the addition of either the BI-RADS density (9,10) or percentage density (11) was modest for every model and ranged from 0.01 to 0.06. These results are very similar to our data comparing the baseline and expanded models (increase in *C*-statistic from 0.03 to 0.05).

More than 85% of all lung cancers occur in ever smokers. The challenge is to try to predict which of the estimated 45 million current smokers and 46 million former smokers in the United States are at highest risk for developing lung cancer? (12). From a public health perspective, we need to segregate out from the totality of ever smokers in the population the smallest possible segment in which the largest number of lung cancers is predicted to arise. Cronin et al. (13) have pointed out the challenges that exist in predicting cancer risk even for those sites (like lung cancer) where the risk factors are so clearly identifiable and measurable.

There are a few lung cancer risk prediction models in the literature. Bach et al. (14) used smoking history data from a large randomized trial of retinol and carotene in heavy smokers and asbestos-exposed individuals. Their model incorporates smoking intensity (years smoked, number of cigarettes per day, duration of quitting), gender, and asbestos exposure. However, this model, although a major contribution to the risk prediction literature, is largely applicable to heavy smokers between 50 and 75 years of age, who were heavy smokers, and who had quit for <20 years. Colditz et al. (15) developed a risk index based on group consensus from literature review that is considered a general guide, rather than providing precise estimation of individual risk. Parameters included in their model are smoking phenotype, family history, air pollution, and fruit and vegetable intake. de Torres et al. (16) showed that emphysema, as detected on low-dose chest computed tomography, was a significant risk factor for lung cancer after adjustment for potential confounders. In their multivariate model that included age, gender, and pack-years, the risk for emphysema was 2.51 (1.10-6.23).

By combining case-control data (579 lung cancer cases and 1,157 age- and sex-matched population-based controls) with regional incidence rates, Cassidy et al. (17) constructed the Liverpool Lung Project model to project individual 5-year absolute risks of developing lung cancer. The variables included were similar to ours, except prior diagnosis of pneumonia is included but prior hay fever is not.

Because our cases and controls were matched on smoking status, our models cannot fully account for the overwhelming contribution of smoking phenotype to lung cancer risk. For example, in former smokers only age at cessation is included, whereas in current smokers only pack-years is considered. The risk estimates, therefore, are all relative to a person of the same age, gender, and smoking status.

For this analysis, we did not have an independent validation population. We acknowledge that the 3-fold cross-validations done on the same data set will not address the issue of overestimation of model accuracy that could occur without such validation. Therefore, the most pressing next step is validation of these models in independent data sets in other screening populations in which demographics and blood have prospectively been collected.

The benefits of reliable risk prediction models to identify high-risk subsets of current and former smokers are manifold. Whereas the uniform advice for any smoker is immediate cessation, prediction models could be helpful in the context of both screening and prevention trials. A large European screening trial of multislice computed tomography screening used risk-based selection for recruitment of current and former smokers with specific smoking intensity criteria and showed that that this approach was feasible and helped to minimize sample size requirements and, therefore, costs (18).

Our baseline model is comprehensive in terms of epidemiologic and clinical variables, but its relatively low sensitivity and specificity limit its widespread use. Addition of the biomarker assays does improve the sensitivity of the models over epidemiologic and clinical data alone. These assays, however, are time-consuming and require some level of technical expertise. Therefore, whereas feasible in a controlled academic setting, they are not applicable for widespread population-based implementation. The next step in model refinement is to include genotyping data that are inexpensive, accurate, and amenable to high-throughput analysis.

Acknowledgments

Grant support: Flight Attendant Medical Research Institute and Public Health Service grants CA55769, CA093592, and CA016672 from the National Cancer Institute, and Department of Defense TARGET grant DAMD17-02-1-0706 (W.K. Hong).

References

1. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. J Natl Cancer Inst 2007;99:715–26. [PubMed: 17470739]
2. Wei Q, Cheng L, Amos CI, et al. Repair of tobacco carcinogen-induced DNA adducts and lung cancer risk: a molecular epidemiologic study. J Natl Cancer Inst 2000;92:1764–72. [PubMed: 11058619]
3. Spitz MR, Wei Q, Dong Q, Amos CI, Wu X. Genetic susceptibility to lung cancer: the role of DNA damage and repair. Cancer Epidemiol Biomarkers Prev 2003;12:689–98. [PubMed: 12917198]
4. Wu X, Lin J, Etzel CJ, et al. Interplay between mutagen sensitivity and epidemiological factors in modulating lung cancer risk. Int J Cancer 2007;120:2687–95. [PubMed: 17290394]
5. Athas WF, Hedayati MA, Matanoski GM, Farmer ER, Grossman L. Development and field-test validation of an assay for DNA repair in circulating human lymphocytes. Cancer Res 1991;51:5786–93. [PubMed: 1933849]

6. Koch KS, Fletcher RG, Grond MP, et al. Inactivation of plasmid reporter gene expression by one benzo (a)pyrene diol-epoxide DNA adduct in adult rat hepatocytes. Cancer Res 1993;53:2279–86. [PubMed: 8485714]
7. Hsu TC, Johnston DA, Cherry LM, et al. Sensitivity to the genotoxic effects of bleomycin in humans: possible relationship to environmental carcinogenesis. Int J Cancer 1989;43:403–9. [PubMed: 2466800]
8. Hanley JA, McNeil BJ. A method of comparing the areas under the receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43. [PubMed: 6878708]
9. 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. [PubMed: 16261410]
10. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 2006;98:1204–14. [PubMed: 16954473]
11. Chen J, Pee D, Ayyagari R, et al. Projecting absolute invasive breast cancer risk in White women with a model that includes mammographic density. J Natl Cancer Inst 2006;98:1215–26. [PubMed: 16954474]
12. National Health Interview Survey. Vital and health statistics: summary health statistics for U.S. adults: National Health Interview Survey, 2005. Oct 4;2006 Series 10(232)
13. 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. [PubMed: 16670389]
14. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003;95:470–8. [PubMed: 12644540]
15. Colditz GA, Atwood KA, Emmons E, et al. Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. Cancer Causes Control 2000;11:477–88. [PubMed: 10880030]
16. de Torres JP, Bastarrika G, Wisnivesky JP, et al. Assessing the relationship between lung cancer risk and emphysema detected on low-dose CT of the chest. Chest 2007;132:1932–8. [PubMed: 18079226]
17. Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung can. Br J Cancer 2008;98:270–6. [PubMed: 18087271]
18. van Iersel CA, de Koning HJ, Draisma G, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007;120:868–74. [PubMed: 17131307]

Table 1

Distribution of case patients and control subjects by selected variables

Variable	Former smokers			Current smokers		
	Case patients (*n* = 375)	**Control subjects (*n* = 371)**	***P***	**Case patients (*n* = 350)**	**Control subjects (*n* = 244)**	***P***
Sex						
Male	222 (59.20)	232 (62.53)		181 (51.71)	125 (51.23)	
Female	153 (40.80)	139 (37.47)	0.351	169 (48.29)	119 (48.77)	0.907
Age						
Mean (SD)	64.8 (8.9)	62.5 (9.7)	0.001	58.0 (10.1)	58.5 (9.4)	0.495
Pack-year						
Mean (SD)	48.7 (31.5)	42.3 (31.1)	0.005	57.6 (29.0)	47.5 (29.0)	<0.001
DNA repair (DRC), %						
Mean (SD)	8.2 (2.6)	9.0 (3.5)	<0.001	8.3 (2.9)	9.2 (3.6)	0.002
Median (range)	8.0 (0.35–17.63)	8.6 (2.39–23.95)		8.1 (2.78–25.55)	8.7 (2.09–27.21)	
Bleomycin sensitivity, breaks/cell						
Mean (SD)	0.8 (0.4)	0.7 (0.4)	<0.001	0.7 (0.4)	0.6 (0.3)	<0.001
Median (range)	0.7 (0.10–3.24)	0.6 (0.08–2.36)		0.7 (0.04–2.74)	0.5 (0.02–2.34)	

Abbreviation: DRC, DNA repair capacity.

Table 2

Multivariable logistic models for predicting lung cancer

Parameter	Levels	Regression coef.	*P*	OR (95% CI)
Former smokers				
Emphysema	No			
	Yes	0.7342	0.0036	2.08 (1.27-3.42)
Dust exposures	No			
	Yes	0.4933	0.0029	1.64 (1.18-2.27)
Family cancer history	0 or 1			
First-degree relatives with cancer	≥2	0.5375	0.0025	1.71 (1.21-2.43)
Age stopped smoking (y)	<42			
	42–53	0.4959	0.0148	1.64 (1.10-2.45)
	≥54	0.9603	<0.0001	2.61 (1.76-3.89)
Hay fever	Yes			
	No	0.5274	0.011	1.69 (1.13-2.55)
DNA repair capacity (% decrease)		0.1069	<0.0001	1.11 (1.06-1.17)
Bleomycin sensitivity (breaks/cell increase)		0.6243	0.0029	1.87 (1.24-2.82)
Current smokers				
Emphysema	No			
	Yes	0.9096	0.0005	2.48 (1.48-4.16)
Pack-years	<28			
	28–41.9	0.2890	0.3162	1.34 (0.76-2.35)
	42–57.4	0.3036	0.2752	1.36 (0.79-2.34)
	≥57.5	0.5160	0.0478	1.68 (1.01-2.79)
Family cancer history	0			
First-degree relatives with (smoking-related cancer)	>1	0.7036	0.0022	2.02 (1.29-3.17)
Hay fever	Yes			
	No	0.4102	0.0842	1.51 (0.95-2.40)
Dusts	No			
	Yes	0.3258	0.0882	1.39 (0.95-2.02)
Asbestos	No			
	Yes	0.5528	0.0516	1.74 (1.00-3.03)
DNA repair capacity (% decrease)		0.0709	0.0131	1.07 (1.02-1.14)
Bleomycin sensitivity (breaks/cell increase)		1.4665	<0.0001	4.33 (2.45-7.66)

Table 3

Final models and results of 3-fold cross-validation analyses

	Baseline model	**Baseline model + DRC**	**Baseline model + bleomycin**	**Baseline model + DRC + bleomycin**
Former smokers				
Overall AUC	0.67 (0.63–0.71)	0.69 (0.65–0.73)	0.68 (0.64–0.72)	0.70 (0.66–0.74)
Cross-validation mean AUC (95% CI)	—	—	—	0.68 (0.62–0.75)
P, goodness of fit	—	—	—	0.610
Current smokers				
Overall AUC	0.68 (0.64–0.72)	0.69 (0.64–0.73)	0.72 (0.68–0.77)	0.73 (0.69–0.77)
Cross-validation mean AUC (95% CI)	—	—	—	0.70 (0.63–0.77)
P, goodness of fit	—	—	—	0.433

Table 4

One-year absolute risks of lung cancer for two hypothetical risk profiles

Smoking status	Baseline (%)	+Suboptimal DRC (%)	+DRC + mutagen sensitivity (%)
Current smoker*	8.70	8.92	16.33
Former smoker†	1.20	2.33	3.06

NOTE: For details of risk calculations, see ref. [1].

*Seventy-five-year-old white man, current smoker, 58 pack-years, prior emphysema, hay fever, asbestos exposure, two first-degree relatives with smoking-related cancer.

†White female former smoker, age 66 y, quit at age 54 y, dust exposure, no family history of cancer or hay fever.