Variations in Lung Cancer Risk Among Smokers

Peter B. Bach, Michael W. Kattan, Mark D. Thornquist, Mark G. Kris, Ramsey C. Tate, Matt J. Barnett, Lillian J. Hsieh, Colin B. Begg

Background: **Although there is no proven benefit associated with screening for lung cancer, screening programs are attracting many individuals who perceive themselves to be at high risk due to smoking. We sought to determine whether the risk of lung cancer varies predictably among smokers.** ***Methods:*** **We used data on 18 172 subjects enrolled in the Carotene and Retinol Efficacy Trial (CARET)—a large, randomized trial of lung cancer prevention—to derive a lung cancer risk prediction model. Model inputs included the subject's age, sex, asbestos exposure history, and smoking history. We assessed the model's calibration by comparing predicted and observed rates of lung cancer across risk deciles and validated it by assessing the extent to which a model estimated on data from five CARET study sites could predict events in the sixth study site. We then applied the model to evaluate the risk of lung cancer among smokers enrolled in a study of lung cancer screening with computed tomography (CT).** ***Results:*** **The model was internally valid and well calibrated. Ten-year lung cancer risk varied greatly among participants in the CT study, from 15% for a 68-year-old man who has smoked two packs per day for 50 years and continues to smoke, to 0.8% for a 51-year-old woman who smoked one pack per day for 28 years before quitting 9 years earlier. Even among the subset of CT study participants who would be eligible for a clinical trial of cancer prevention, risk varied greatly.** ***Conclusions:*** **The risk of lung cancer varies widely among smokers. Accurate risk prediction may help individuals who are contemplating voluntary screening to balance the potential benefits and risks. Risk prediction may also be useful for researchers designing clinical trials of lung cancer prevention. [J Natl Cancer Inst 2003;95:470–8]**

There is no analog in lung cancer to the Gail Model in breast cancer, in which an individual's absolute risk of disease can be estimated from his or her exposure history *(1)*. It might be argued that such a model is unnecessary, in that most people who have smoked are at high risk of lung cancer, whereas those who have never smoked are not. However, if there were to be large, predictable variations in lung cancer risk among people who have smoked, the ability to determine individual risk from the person's history of exposures would be a useful adjunct to both patient care and clinical research.

Individuals who are considering lung cancer screening, for example, could balance their likelihood of developing lung cancer against the harms that can result from false-positive findings, including complications, costs, and anxiety associated with subsequent diagnostic tests *(2)*. Individual lung cancer risk prediction could also be incorporated into the design, recruitment, and analysis of studies of lung cancer prevention, potentially reducing the sample size required to achieve the desired statistical power. Conceivably, risk prediction could even be used to help determine whether highly sensitive screening technologies such as computed tomography (CT) lead to overdiagnosis of lung cancer by comparing the rates of predicted and detected incident disease within a screened population *(3,4)*.

In this article, we describe the development and validation of a model of individual lung cancer risk that can be applied in both

Affiliations of authors: P. B. Bach (The Health Outcomes Research Group, Department of Epidemiology and Biostatistics and Department of Medicine), M. W. Kattan (The Health Outcomes Research Group, Department of Epidemiology and Biostatistics and Department of Urology), Memorial Sloan-Kettering Cancer Center, New York, NY; M. D. Thornquist, M. J. Barnett, Fred Hutchinson Cancer Research Center, Seattle, WA; M. G. Kris, Department of Medicine, Memorial Sloan-Kettering Cancer Center; R. C. Tate, L. J. Hsieh, C. B. Begg, The Health Outcomes Research Group, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center.

Correspondence to: Peter B. Bach, M.D., Memorial Sloan-Kettering Cancer Center, 307 E. 63rd St., New York, NY 10021 (e-mail: bachp@mskcc.org).

See "Notes" following "References."

Journal of the National Cancer Institute, Vol. 95, No. 6,

clinical and research settings. The model is based on data from participants in the Carotene and Retinol Efficacy Trial (CARET), a large, randomized trial of lung cancer prevention. To determine whether the risk of lung cancer varies, we examined the predicted 10-year lung cancer risk among subjects enrolled in an ongoing CT screening program. To ascertain the usefulness of the model as an adjunct to clinical research, we assessed the extent of variation in risk among a cohort of individuals who meet typical eligibility criteria for cancer prevention studies.

Subjects and Methods

Study Subjects

Our model was derived from data collected during CARET, a multicenter, randomized, controlled study that evaluated the impact of beta-carotene and vitamin A supplementation on lung cancer incidence and mortality *(5,6).* CARET enrolled two populations. One consisted of 14 254 heavy smokers (men and women, aged 50–69 years), who had at least 20 pack-years of smoking exposure and were either current smokers or had quit within 6 years of enrollment. The other consisted of 4060 asbestos-exposed men (aged 45–69 years, either current smokers or former smokers who had quit within 15 years of enrollment) who had either radiologic evidence of asbestos exposure or a history of employment in a trade that put them at high risk for asbestos exposure (primarily shipyard or construction workers).

A total of 18 314 individuals were randomly assigned to receive either placebo or the study drug (30 mg/day beta-carotene and 25 000 IU/day retinyl palmitate). Randomization for the pilot study began in June 1985, followed by randomization for the full-scale study in June 1989; study accrual ended in September 1994. The intervention itself was stopped in January 1996 after preliminary results revealed definitive evidence of no benefit and substantial evidence of possible harm *(6),* but study subjects continue to be followed annually by mail, with additional data collection on reported endpoints. The subjects included in our analyses were the 18 172 individuals who had a documented history of current or former smoking. Our analyses of the CARET data were approved by the Institutional Review Board at the Fred Hutchinson Cancer Research Center, Seattle, WA.

Risk Prediction Model

Our model is configured to estimate the absolute risk that an individual will be diagnosed with lung cancer within 10 years. We chose the 10-year time horizon because it is probably in excess of the time it takes for lung cancer to progress from an undetectable size to an untreatable stage; consequently, it is a useful perspective from which to counsel patients about screening. In addition, as Woloshin et al. *(7)* have suggested, the 10-year time frame is one that most patients can imagine.

To determine the absolute risk of lung cancer for an individual within 10 years, we created two 1-year models. One predicts the probability of being diagnosed with lung cancer (the focus of our study), and the other predicts the probability that an individual will die without having been diagnosed with lung cancer (the competing risk). We then recursively estimated 10-year lung cancer risk by cycling these two 1-year models 10 times. In each year, the risk of lung cancer diagnosis and the risk of death in the absence of lung cancer were estimated (both were absorbing states in the model). Then, for each subsequent cycle, we incrementally changed the values of the predictors and reduced the at-risk pool to simulate one of two scenarios: continued smoking (at the same level) and continued abstinence from smoking. For former smokers, we modeled only the risk if abstinence were continued. Because any-cause death rates in clinical trials are often not consistent with those seen in the general population, we ran additional recursive models in which the risk of death in the absence of a lung cancer diagnosis was based on other sets of assumptions. We found that our estimates of 10-year lung cancer risk did not change substantially, whether we used estimates of any-cause death rates from the age- and sex-specific sections of the National Center for Health Statistics decennial life tables *(8),* which are somewhat lower than the any-cause death rates in CARET, or whether we used any-cause death rates that were three times larger than those listed in the life tables, thus exceeding both what was observed in CARET and the relative excess mortality that is typically seen in a population of heavy smokers *(9).*

Outcomes and Predictors

Both 1-year models were developed using similar methods and predictors. We chose predictors—age, sex, prior history of asbestos exposure, duration of smoking, average amount smoked per day while smoking, and duration of abstinence from smoking for former smokers—that met two criteria. First, they are identifiable from a clinical history; second, they are established or strongly suspected risk factors for lung cancer. All of these predictors are also risk factors for all-cause mortality *(10,11).* Other potential predictors that may be germane to risk prediction, including history of obstructive lung disease, brand of cigarette smoked, type of asbestos exposed to, findings on chest x-ray, and exposure to radon or secondhand smoke, were not considered because they were either not easily assessable through subject interview or were not recorded as part of the CARET study. An individual's age at the time he or she started smoking was not included as a predictor in our analyses because it is a function of predictors that were included (i.e., age, duration of smoking, and duration of abstinence).

Derivation of 1-Year Models

Cox proportional hazards regression was used to estimate the multivariable relations between the risk factors and the outcomes (i.e., lung cancer diagnosis and death in the absence of lung cancer diagnosis) *(12).* For the regression analyses, data gathered from each individual was divided into individual person–time periods. The beginning of each time period was defined by date of an encounter with a study coordinator (either initial or follow-up). The end of the time period was defined by either the date of the outcome or the date of a censored event, which included a subsequent follow-up encounter, the achievement of the alternative outcome, or the end of the follow-up period. The data were analyzed in this manner to take advantage of the fact that the values of the smoking exposure predictors in the CARET study were updated at each encounter.

Continuous predictors (age, duration of abstinence, duration of smoking, and number of cigarettes smoked per day) were fit with restricted cubic splines to allow for nonlinear and nonmonotonic effects; the knots separated quartiles of the data *(13).* Study arm, sex, and asbestos exposure were treated as categorical variables. All decisions with respect to the coding of variables were made prior to modeling.

The values of the predictors for each subject were determined from the responses of the CARET participants. To adhere to the principle that risk factors be identifiable through subject interviews, we characterized asbestos exposure as present if the study subject reported an exposure history that included a first occupational exposure occurring 15 or more years previously and a minimum duration of 5 years in a trade that put him or her at high risk of asbestos exposure (i.e., asbestos worker, insulator, lagger, plasterboard worker, drywaller, plasterer, ship scaler, ship fitter, rigger, shipyard boilermaker, shipyard welder, shipyard machinist, shipyard coppersmith, shipyard electrician, plumber/pipefitter, steamfitter, sheet metal worker). Smoking exposure history was determined from responses to the following questions or to similar questions: What is the total length of time, in years, that you have smoked cigarettes? On the average of the entire time you smoked, how many cigarettes did you smoke per day? and How old were you when you quit smoking cigarettes?

Validation of 1-Year Models

The proportional hazards assumption that the hazard ratio was constant over time was verified by tests of correlations with time and examination of residual plots *(12).* Discrimination was assessed by the concordance index, after the optimistic bias was reduced through 10-fold cross-validation *(13–15).* For each of the continuous predictors in the model, we compared the estimated relation with lung cancer risk with previously reported relations between that same predictor and lung cancer risk. We chose results from prior studies in which the comparators were presented in the context of controlling for or stratifying on the other continuous factors in our model. To present the results graphically, some results were rescaled and some results were adapted from published equations.

We then assessed internal calibration of the model with a 10-fold cross-validated calibration plot, and we assessed internal validity by determining the extent to which a model estimated on data from five of the six study sites in CARET could predict events in the sixth study site. Of the six study sites (Seattle, WA; Baltimore, MD; New Haven, CT; Portland, OR; San Francisco, CA; and Irvine, CA), we performed this validation three times, holding out in turn each of the three largest study sites (i.e., Seattle, Portland, and Irvine). We evaluated each of these analyses by comparing the rates of accumulated predicted and observed events across deciles of predicted risk.

Forecasted Risk

To assess the extent of variation in lung cancer risk among smokers, we analyzed data on 300 randomly selected subjects enrolled in an ongoing volunteer study of low-dose computed tomography (CT) at the Mayo Clinic in Rochester, MN. To determine whether risk prediction has the potential to enhance clinical research in lung cancer prevention and detection, we drew a Lorenz curve of the distribution of lung cancer events (based on 1-year risk) that are likely to occur among the subset of 201 subjects in the Mayo Clinic study who meet the entry criteria for the National Lung Screening Trial (NLST) *(16).* The NLST, which has begun patient accrual, is a large, randomized, controlled trial of lung cancer screening whose entry criteria are typical of those in lung cancer prevention and detection studies: subjects must be aged 55–74 years, have smoked a minimum of 30 pack-years, and be current smokers or former smokers who quit within the last 15 years *(17,18).* The analyses of the Mayo Clinic data were approved by the Institutional Review Board at the Mayo Clinic.

Statistical Analyses

Statistical analyses were performed using S-Plus 2000 Professional software (Insightful Corp., Redmond, WA) with additional functions (the Design library) *(19)* and SAS software (version 6.12; SAS Institute Inc., Cary, NC). Our recursive model was run in Microsoft Excel 2000 (Microsoft Corp., Redmond, WA). All *P* values are two-sided.

RESULTS

CARET Cohort

From 1985 through 1994, 18 314 individuals were enrolled in the CARET study; we analyzed data on the 18 172 who had documented current or former smoking history (14 254 from the

Table 1. Characteristics of the cohort of subjects in the Carotene and Retinol Efficacy Trial (CARET) who were used to generate a model of lung cancer risk

Variable	No. of subjects (column %)	No. of lung cancer cases (row %)	No. of deaths (row %)
All	18 172	1070 (5.9)	3175 (17.5)
Heavy smoker cohort	14 254 (78.4)	847 (5.9)	2331 (16.4)
Asbestos cohort	3918 (21.6)	223 (5.7)	844 (21.5)
Age at study entry, y			
44–54	6317 (34.8)	190 (3.0)	511 (8.1)
55–60	4638 (25.5)	247 (5.3)	654 (14.1)
61–75	7217 (39.7)	633 (8.8)	2010 (27.9)
Sex			
Male	11 883 (65.4)	732 (6.2)	2371 (20.0)
Female	6289 (34.6)	338 (5.4)	804 (12.8)
Smoking status*			
Former	7168 (39.4)	303 (4.2)	1103 (15.4)
Current	11 004 (60.6)	767 (7.0)	2072 (18.8)
Race			
White	16 939 (93.2)	996 (5.9)	2938 (17.3)
Black	524 (2.9)	40 (7.6)	142 (27.1)
Hispanic	273 (1.5)	7 (2.6)	28 (10.3)
Asian/Pacific Islander	205 (1.1)	11 (5.4)	25 (12.2)
American Indian/Alaskan	148 (0.81)	8 (5.4)	27 (18.2)
Other/unknown	83 (0.45)	8 (9.6)	15 (18.1)
Arm			
Placebo	8825 (48.6)	458 (5.2)	1411 (16.0)
Intervention	9347 (51.4)	612 (6.5)	1764 (18.9)
Enrollment site			
Seattle, WA (including pilot study†)	6655 (36.6)	449 (6.7)	1325 (19.9)
Baltimore, MD	822 (4.5)	58 (7.1)	190 (23.1)
New Haven, CT	1041 (5.7)	46 (4.4)	186 (17.9)
Portland, OR	4564 (25.1)	283 (6.2)	772 (16.9)
San Francisco, CA	866 (4.8)	45 (5.2)	177 (20.4)
Irvine, CA	4224 (23.2)	189 (4.5)	525 (12.4)
Education			
No high school diploma	2044 (11.3)	180 (8.8)	471 (23.0)
High school	4320 (23.8)	239 (5.5)	736 (17.0)
Some college	6226 (34.3)	322 (5.2)	895 (14.4)
Completed college	3737 (20.6)	156 (4.2)	489 (13.1)
Unknown	1845 (10.2)	173 (9.4)	584 (31.7)

*Ranges were as follows: No. of cigarettes smoked per day = 1–90; duration of smoking in years = 1–63; duration of abstinence in years = 0–51.

†Pilot study contributed 1705 subjects to our analyses.

heavy smoking cohort and 3918 from the asbestos cohort). The characteristics of these subjects are listed in Table 1. Subjects contributed an average of 13.6 observational intervals (median = 13), with a mean duration of 265 days per interval (median = 200 days, interquartile range = 120–365 days). As of February 25, 2002, the subjects had been followed to an outcome of lung cancer for 168 343 person-years and an outcome of death for 169 785 person-years, during which time 1070 of the subjects were diagnosed with lung cancer (incidence rate = 636 per 100 000 person-years) and 3175 of the subjects died (mortality rate = 1870 per 100 000 person-years). Among the observed lung cancer cases, both the distribution of histologic subtypes

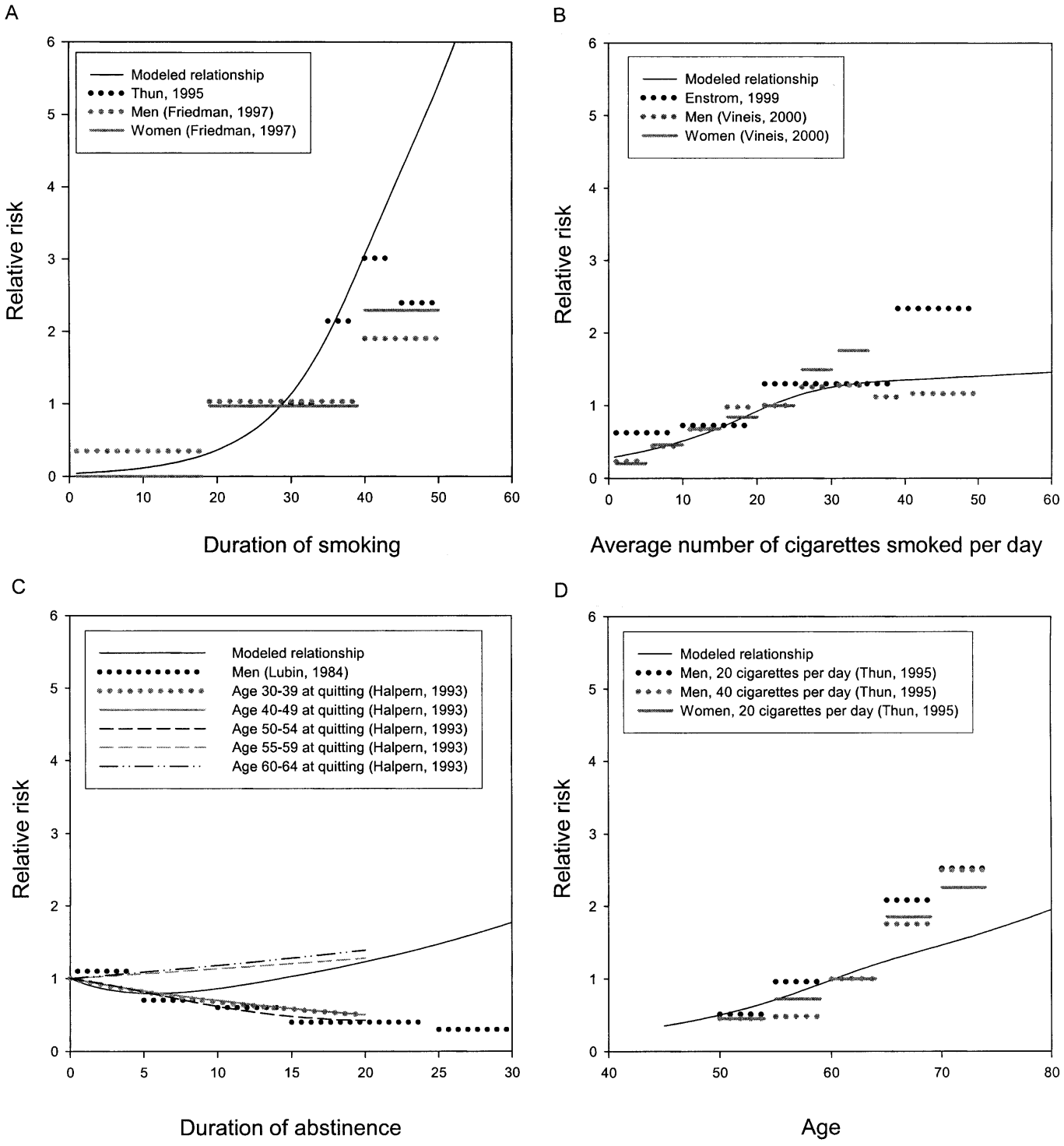

Fig. 1. Modeled multivariable relations between 1-year lung cancer risk and each of the four continuous predictors. Superimposed over each of the modeled relationships are parallel results previously reported by other authors for the same continuous predictors that had been analyzed in similar multivariable contexts in which the other continuous predictors were either controlled for or stratified on. Relative risks of lung cancer are shown for (**A**) duration of smoking *(45,46)*, (**B**) average number of cigarettes smoked per day *(47,48)*, (**C**) duration of abstinence *(33,35)*, and (**D**) age *(46)*. When the previous reports grouped the populations into categories, the expanse of the groupings is indicated by the length of the horizontal bars. When the previous reports included a regression equation, the shape of the reported response has been reproduced.

and the survival distribution were consistent with national statistics *(20)*: 77% of cases were non-small-cell cancers, 18% were small-cell cancers, and the median survival after diagnosis was 7.4 months.

Lung Cancer Risk Model

Our regression analyses yielded two multivariable 1-year risk models: one that predicted the probability of being diagnosed with lung cancer and one that predicted the probability of dying without a lung cancer diagnosis. (*See* supplemental equations on the Journal's Web site at http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue6/index.shtml.) We report here our validation efforts for the former model. The latter model was used only in the recursive estimation process and was subjected to sensitivity analyses rather than further validation. In the 1-year lung cancer risk model, the associations between risk factors and lung cancer occurrence were consistent with those in previous reports for both continuous predictors (duration of smoking, average number of cigarettes smoked per day, duration of abstinence, and age [Fig. 1]) and categorical predictors. The study drug (i.e., beta-carotene and retinyl palmitate) increased the risk of lung cancer to a degree consistent with previously published data from CARET (hazard ratio [HR] = 1.20, 95% confidence interval = 1.06 to 1.25; $P = .004$) *(6)*. A history of asbestos exposure, one of the categorical predictors, was associated with an independent increase in lung cancer risk (HR = 1.24, 95% CI = 1.04 to 1.48; $P = .02$). There was no statistical evidence that sex independently influenced lung cancer risk (HR = 0.94, 95% CI = 0.92 to 1.08; $P = .41$).

The cross-validated concordance index was 0.72, and the cross-validated calibration plot by risk decile was consistent with excellent calibration (Fig. 2). One might argue that a model that takes as inputs only age, the presence of a smoking history (all of our subjects), and study drug (a factor unique to our analyses) as predictors might be sufficient for lung cancer risk prediction. However, in our analyses, this model had a cross-validated concordance index of only 0.66 (data not shown). Furthermore, this model cannot identify variations in risk between individuals of a particular age, which can be quite large, based on individual smoking history. These variations are most clearly illustrated in the risk tables (Appendix). In each of the three analyses in which we derived a model from five sites and validated it on the sixth site, the observed rates of lung cancer across the risk deciles for the held-out site closely matched those that had been predicted by the corresponding model derived from the five included sites (Fig. 3), adding strength to the argument that our model will be generalizable to other populations of smokers.

Variation in 10-Year Lung Cancer Risk Among Subjects Undergoing Voluntary Screening

Two examples of actual individuals in the Mayo Clinic study illustrate the range of modeled 10-year lung cancer risk among smokers (Table 2). A 51-year-old female who smoked one pack per day for 28 years and quit smoking 9 years earlier is in the 5th percentile of risk. Assuming that she remains abstinent, her 10-year risk of lung cancer is less than 1% (0.80%, which could also be characterized as 1 in 120). To provide a benchmark for this estimate, the 10-year risk of lung cancer for an individual of similar age who has never smoked is roughly an order of magnitude lower, approximately 0.07% (1 in 1400) *(21)*.

A 68-year-old male current smoker who has smoked two packs per day for 50 years is in the 95th percentile in the Mayo Clinic study and faces approximately 15–20 times the risk of the

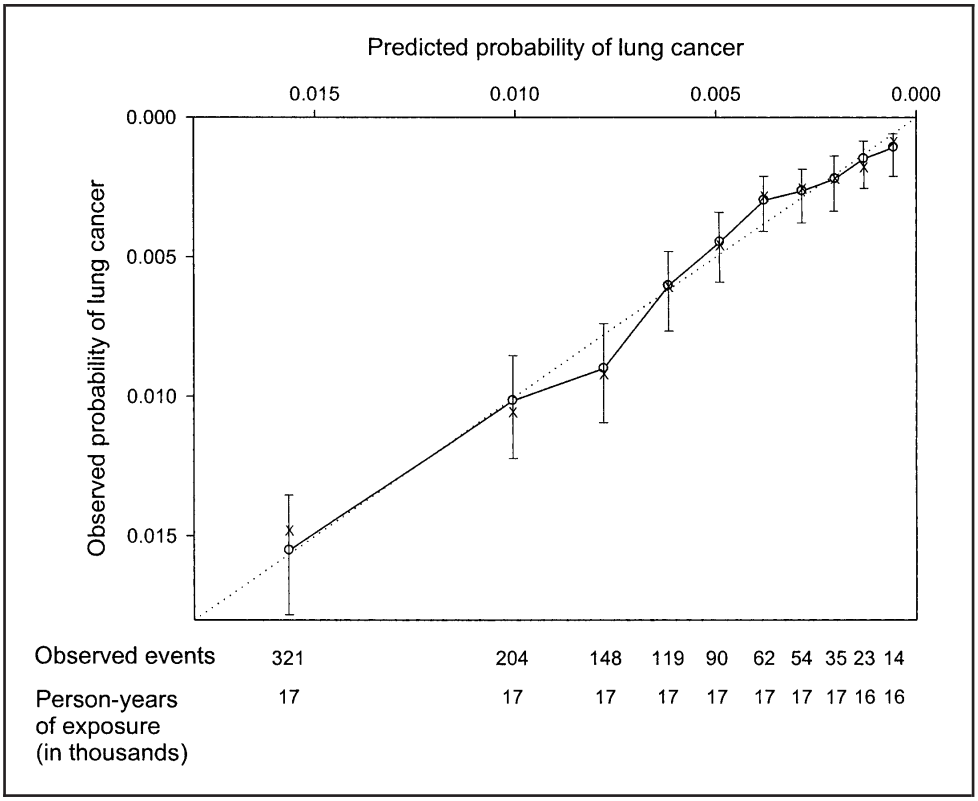

Fig. 2. Calibration of the 1-year lung cancer risk model by decile of risk. The predicted and observed probabilities of lung cancer for each group are graphed on the horizontal and vertical axes, respectively. The **dashed line** indicates the reference line, on which an ideal model would lie. **Circles** mark the apparent predictions for each decile; **X's** mark the cross-validated predictions for each decile. **Vertical bars** indicate 95% confidence intervals around the apparent prediction. The numbers of observed events and person-years of exposure in each decile are listed below the graph.

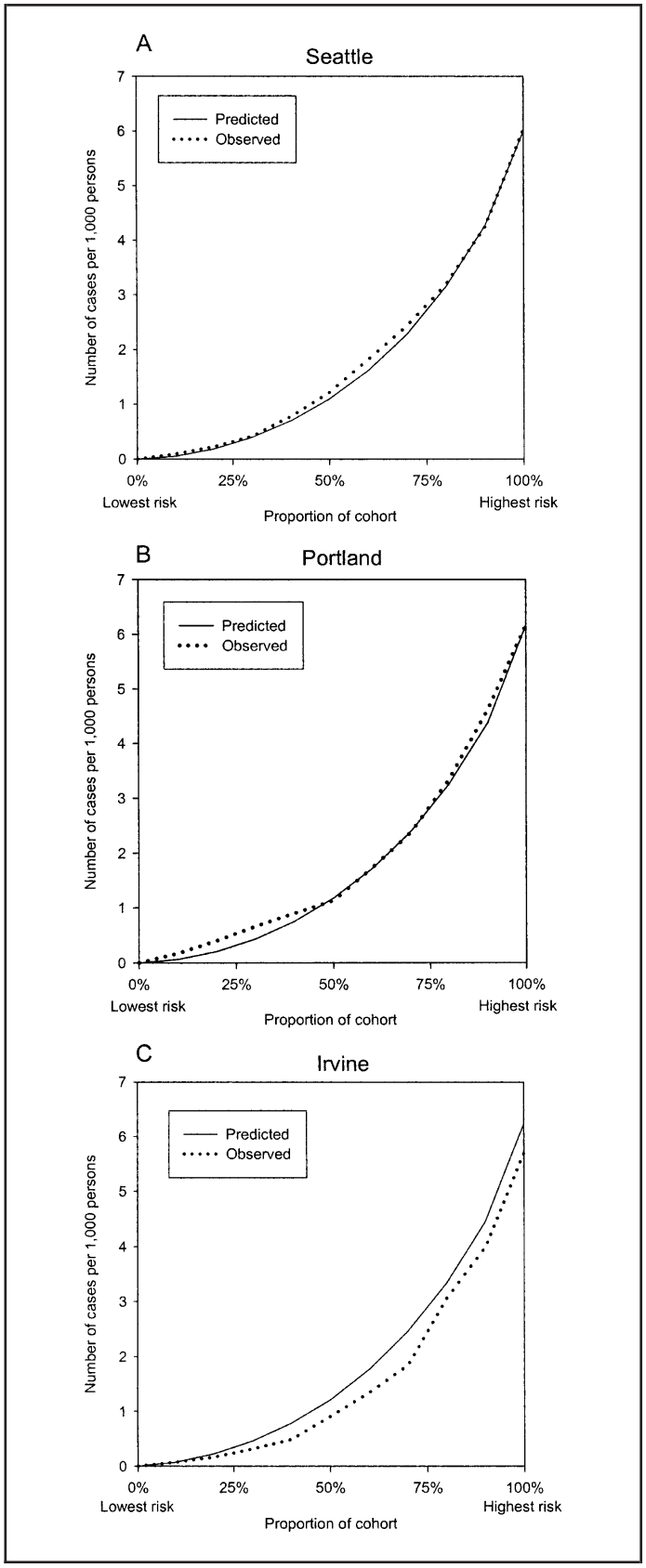

Fig. 3. Comparison of predicted and observed rates of lung cancer across risk deciles. In each of the analyses, data from five of the six Carotene and Retinol Efficacy Trial study sites were used to derive a model of 1-year lung cancer risk that was then applied to individuals in the sixth held-out study site. The graphs display the cumulative predicted and observed rates of lung cancer in each of the held-out study sites; the rates are based on deciles of predicted risk. (**A**) Seattle held out. (**B**) Portland held out. (**C**) Irvine held out.

Table 2. Predicted lung cancer risk for a sample of participants in an ongoing study of low-dose computed tomography screening at the Mayo Clinic*

	Percentile of risk in Mayo Clinic study				
Characteristic	5th	25th	50th	75th	95th
Age, y	51	52	58	56	68
Sex	F	F	M	F	M
Average No. of cigarettes smoked per day	20	20	25	40	40
Duration of smoking, y	28	35	40	44	50
Duration of abstinence, y	9	0	3	0	0
Asbestos exposure	No	No	No	No	No
10-year risk if no further smoking, %	0.80	1.50	4.10	5.60	10.80
10-year risk if continued smoking, %	NA	2.80	NA	8.40	14.90

*M = male; F = female; NA = not applicable.

person at the 5th percentile. His 10-year risk of lung cancer is 11% (1 in 9) if he quits smoking immediately and 15% (1 in 7) if he continues to smoke at the current level. Other examples of actual Mayo Clinic study participants in Table 2 illustrate both the broad range of risk among individuals undergoing voluntary screening and the potential reduction in lung cancer risk for the current smokers if they quit smoking.

Variation in Risk Within Cohorts Enrolled in Clinical Trials of Cancer Prevention and Detection

Even among the 201 participants in the Mayo Clinic study who fit the eligibility for randomized, controlled screening studies such as the NLST, there was a broad distribution of risk (Fig. 4). The effect of this broad distribution is that most of the incident cancers that will be observed in the study will cluster in

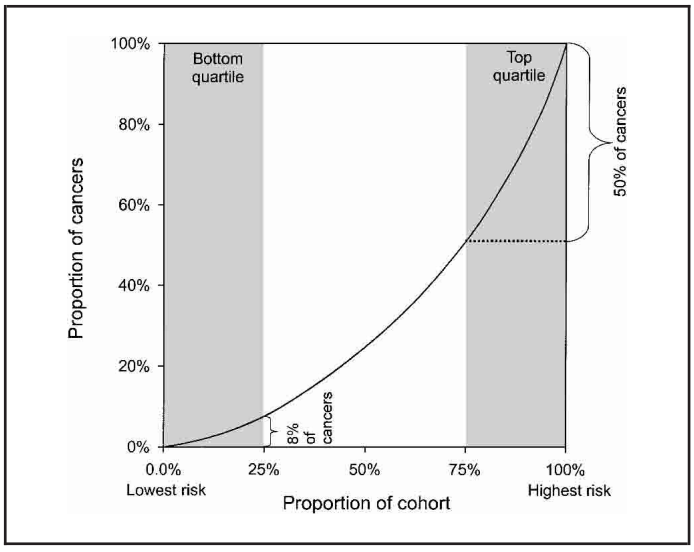

Fig. 4. Cumulative distribution of expected lung cancer events for 201 enrollees in a study of low-dose computed tomography for lung cancer detection at the Mayo Clinic, Rochester, MN, who met the entry criteria for the National Lung Screening Trial (i.e., aged 55–74 years; current or former smoker who has quit within the last 15 years, 30 pack-years of smoking or more). The figure demonstrates that 8% of lung cancer events will occur in the lowest quartile risk group and that 50% of lung cancer events will occur in the highest quartile risk group.

a small segment of the population. For example, the one-quarter of individuals who are at the lowest risk will account for approximately 8% of the incident lung cancer cases, whereas the one-quarter of individuals who are at the highest risk will account for roughly 50% of the lung cancer cases (Fig. 4). One implication of this finding is that the number of observed events in a clinical study of lung cancer detection could be enriched through risk prediction.

DISCUSSION

We derived a lung cancer risk prediction model using data from a large, multicenter, randomized, controlled trial of lung cancer prevention in which there was regular ascertainment of lung cancer risk factors, meticulous follow-up, and validation of outcomes. This model can be applied to individual patients with the use of either a risk table or one of several software applications, all detailed in the Appendix. We then applied the risk prediction model to a population of smokers enrolled in a voluntary study of CT and observed a broad range of lung cancer risk. This finding has important implications with regard to lung cancer screening recommendations and the design of lung cancer prevention trials.

Currently, individuals who are considering lung cancer screening have little information on which to base their decisions *(22).* They may be aware that low-dose CT screening identifies noncancerous abnormalities in the lung parenchyma 20%–50% of the time, most of which require some sort of follow-up evaluation *(23).* They may also have heard that CT screening has not been shown to reduce mortality from lung cancer *(17,24).* Yet the perception that early lung cancer detection should lead to improved outcome seems to be widespread, whereas the potential downstream negative consequences of screen-detected false positives are, in general, under-appreciated by both patients and health care providers.

Because our risk prediction model can help patients to locate themselves along the spectrum of lung cancer risk, it may provide them with a context for their decision about whether to be screened *(7).* For some individuals, such as those of advanced age and heavy smoking exposure, the risk of lung cancer may exceed 10% within 10 years, even if they stop smoking. Given this high risk of disease, participation in a screening program may be compelling. However, screening programs are also proving to be enticing to individuals at substantially lower risk. We found that some individuals who are participating in lung cancer screening trials have a 10-year risk of lung cancer of less than 1%. If such individuals are educated about their low level of personal risk, we anticipate that many will find screening to be unappealing.

Our analyses also suggest that risk prediction may be a useful adjunct for planning and conducting clinical trials of cancer prevention. For example, to achieve its desired statistical power, the NLST plans to enroll 50 000 individuals *(18,25).* When we forecast the distribution of lung cancer cases in a population that met the entry criteria for the study, we found that risk prediction may allow the investigators to identify those subjects in whom lung cancer is most likely to occur. If so, an enrollment strategy focused on these high-risk individuals might enable the investigators to reduce either sample size or study duration without sacrificing statistical power. However, there is a trade-off in that focusing exclusively on high-risk individuals might diminish the generalizability of a study's findings.

We examined the associations between lung cancer risk and each of our predictors as a way to validate the multivariable model, and caution should be exercised in interpreting our findings outside this context. Nevertheless, some findings are intriguing. For example, several studies *(26–28)* have suggested that, at the same level of smoking exposure, a woman's risk of lung cancer exceeds that of a man. However, the methods used in these studies have been challenged, and a number of other investigations have found no such association *(21,29–32).* Our model revealed no convincing association between sex and lung cancer risk, although systematic differences by sex in the ascertainment of other exposures could have masked a small effect.

The relationship between lung cancer risk and duration of smoking cessation has also been uncertain. Peto *(33)* and Halpern et al. *(34)* have argued that, relative to nonsmokers, individuals who quit do not undergo further elevations in lung cancer risk but also do not show a decrease in excess risk over time relative to never smokers. By contrast, Samet *(35)* and Lubin et al. *(36),* as well as the surgeon general's report on smoking *(11),* have argued that there is an independent risk-reducing effect of quitting smoking, such that longer durations of abstinence are associated with greater reductions in risk. Our analyses were more consistent with the first set of conclusions: we did not observe an additional independent benefit associated with more prolonged quitting. Instead, the difference in risk between continuing smokers and quitters appears to be explained almost entirely by differences in duration of smoking between the two groups (Fig. 1). This finding is consistent with recent evidence of the persistence of cancer-associated genetic alterations in former smokers *(37,38).*

As a practical clinical tool, our risk prediction model has some limitations. It does not distinguish among the risks of different histologic types of lung cancer, and it is relevant only to one subset (albeit a large subset) of at-risk individuals—those aged 50 years or older who have a smoking history. Our model also would benefit from further validation, because prediction models are typically less accurate in new groups of subjects than they appear to be when they are originally described and because our subjects were participants in a clinical trial of lung cancer prevention and are therefore not perfectly representative of members of the population at large *(39–41).*

Lung cancer kills the vast majority of individuals that it afflicts. This bleak fact has inspired researchers to develop cancer prevention strategies and motivated individuals to seek out screening evaluations. However, both our risk prediction model and parallel studies of other cancers support the hypothesis that the risk of cancer is highly variable among individuals in the general population *(42).* To the extent that our model can differentiate individuals of different risks, it may serve as a useful adjunct to both investigators and patients, perhaps playing a role similar to risk prediction models for colon cancer and breast cancer *(43–45).*

APPENDIX

There are several ways in which readers can apply the lung cancer risk prediction model. One is to use the 10-year risk table (Appendix Table 1) that was generated from the model.

Appendix Table 1. Approximate 10-year risk of developing lung cancer*

	Duration of smoking					
	25 years		40 years		50 years	
Age, y	Quit, %	Still smoking, %	Quit, %	Still smoking, %	Quit, %	Still smoking, %
			1 pack per day smokers			
55	<1	1	3	5	NA	NA
65	<1	2	4	7	7	10
75	1	2	5	8	8	11
			2 packs per day smokers			
55	<1	2	4	7	NA	NA
65	1	3	6	9	10	14
75	2	3	7	10	11	15

*These tables assume that people who have quit smoking will continue to abstain for the next 10 years and those who are still smoking will keep smoking the same amount for the next 10 years. For individuals with occupational asbestos exposure, the risks should be multiplied by 1.24. There was a relative paucity of events observed among individuals in CARET outside the given age ranges (i.e., younger than age 55, older than age 85), making prediction outside the given age range potentially unreliable. NA = not available.

The lung cancer risk model can also be applied in a more individualized fashion through computer-assisted risk prediction. Downloadable and Web-enabled software is available as supplemental data at the Journal's Web site (http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue6/index.shtml).

For those readers interested in using the model in programming, the equations for 1-year probability of remaining free of lung cancer and 1-year probability of survival in the absence of lung cancer are available as supplemental data at the Journal's Web site (http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue6/index.shtml). Neither equation includes the coefficient for the presence of the study drug (i.e., beta-carotene and retinyl palmitate) because it is not used in routine practice.

REFERENCES

(1) 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.

(2) Black WC. Advances in radiology and the real versus apparent effects of early diagnosis. Eur J Radiol 1998;27:116–22.

(3) Black WC. Overdiagnosis: an underrecognized cause of confusion and harm in cancer screening. J Natl Cancer Inst 2000;92:1280–2.

(4) Patz EF Jr, Goodman PC, Bepler G. Screening for lung cancer. New Engl J Med 2000;343:1627–33.

(5) Omenn GS, Goodman G, Thornquist M, Grizzle J, Rosenstock L, Barnhart S, et al.The beta-carotene and retinol efficacy trial (CARET) for chemoprevention of lung cancer in high risk populations: smokers and asbestos-exposed workers. Cancer Res 1994;54:2038s–43s.

(6) Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. New Engl J Med 1996;334:1150–5.

(7) Woloshin S, Schwartz LM, Welch HG. Risk charts: putting cancer in context. J Natl Cancer Inst 2002;94:799–804.

(8) National Center for Health Statistics. U.S. decennial life tables for 1989–91. Vol. 1, No. 1. United States Life Tables. (PHS) 97-1150-1. Hyattsville (MD): Public Health Service. p. 44. [Accessed 2/4/2003.] Available at: http://www.cdc.gov/nchs/products/pubs/pubd/lftbls/decenn/1991-89.htm.

(9) 1989 Surgeon General Report: reducing the health consequences of smoking. [Accessed 1/31/2003.] Available at: http://www.cdc.gov/tobacco/sgr/sgr_1989/1989SGRChapter3.pdf. p. 148.

(10) Bepler G. Lung cancer epidemiology and genetics. J Thorac Imaging 1999; 14:228–34.

(11) The health consequences of smoking: cancer. A report of the Surgeon General. DHHS Publ No. 82-50179. Rockville (MD): U.S. Department of Health and Human Services; 1982.

(12) Collett D. Modelling survival data in medical research. 1st ed. New York (NY): Chapman and Hall; 1994.

(13) Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.

(14) Efron B, Gong G. A leisurely look at the bootstrap, the jacknife, and cross-validation. Am Stat 1983;37:36–48.

(15) Schumacher M, Hollander N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building? Stat Med 1997;16:2813–27.

(16) Lorenz MO. Methods for measuring the concentration of wealth. Am Stat Assoc 1905;9:209–19.

(17) Swensen SJ, Jett JR, Sloan JA, Midthun DE, Hartman TE, Sykes AM, et al. Screening for lung cancer with low-dose spiral computed tomography. Am J Respir Crit Care Med 2002;165:508–13.

(18) National Lung Screening Trial. National Cancer Institute. [Accessed: 1/6/2003.] Available at: http://cancer.gov/nlst.

(19) Harrell FE. Regression modeling strategies: with applications to linear models, logisitic regression, and survival analysis. New York (NY): Springer, 2001.

(20) Ries LA, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, et al. SEER cancer statistics review, 1973–1998. Bethesda (MD): National Cancer Institute; 2001. [Accessed 1/30/2003.] Available at: http://seer.cancer.gov/csr/1973_1998/.

(21) Thun MJ, Calle EE, Rodriguez C, Wingo PA. Epidemiological research at the American Cancer Society. Cancer Epidemiol Biomarkers Prev 2000;9: 861–8.

(22) Lee TH, Brennan TA. Direct-to-consumer marketing of high-technology screening tests. New Engl J Med 2002;346:529–31.

(23) Bach PB, Kelley MJ, Tate RC, McCrory DC. Screening for lung cancer: a review of the current literature. Chest 2003;123(1 Suppl):72S–82S.

(24) Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, et al. Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 1999;354:99–105.

(25) Lung screening study ties up advisors, board sends it back to NCI for revision. Cancer Lett 2001;27: 1–4.

(26) Risch HA, Howe GR, Jain M, Burch JD, Holowaty EJ, Miller AB. Are female smokers at higher risk for lung cancer than male smokers? A case-control analysis by histologic type. Am J Epidemiol 1993;138:281–93.

(27) Harris RE, Zang EA, Anderson JI, Wynder EL. Race and sex differences in lung cancer risk associated with cigarette smoking. Int J Epidemiol 1993;22:592–9.

(28) Zang EA, Wynder EL. Differences in lung cancer risk between men and women: examination of the evidence. J Natl Cancer Inst 1996;88:183–92.

(29) Perneger TV. Sex, smoking, and cancer: a reappraisal. J Natl Cancer Inst 2001;93:1600–2.

(30) Kreuzer M, Boffetta P, Whitley E, Ahrens W, Gaborieau V, Heinrich J, et al. Gender differences in lung cancer risk by smoking: a multicentre case-control study in Germany and Italy. Br J Cancer 2000;82:227–33.

(31) Osann KE, Anton-Culver H, Kurosaki T, Taylor T. Sex differences in lung-cancer risk associated with cigarette smoking. Int J Cancer 1993;54:44–8.

(32) Prescott E, Osler M, Hein HO, Borch-Johnsen K, Lange P, Schnohr P, et al. Gender and smoking-related risk of lung cancer. The Copenhagen Center for Prospective Population Studies. Epidemiology 1998;9:79–83.

(33) Peto R. Influence of dose and duration of smoking on lung cancer rates. IARC Sci Publ 1986;(74):23–33.

(34) Halpern MT, Gillespie BW, Warner KE. Patterns of absolute risk of lung cancer mortality in former smokers. J Natl Cancer Inst 1993;85:457–64.

(35) Samet JM. The health benefits of smoking cessation. Med Clin North Am 1992;76:399–414.

(36) Lubin JH, Blot WJ, Berrino F, Flamant R, Gillis CR, Kunze M, et al. Modifying risk of developing lung cancer by changing habits of cigarette smoking. Br Med J (Clin Res Ed) 1984;288:1953–6.

(37) Mao L, Lee JS, Kurie JM, Fan YH, Lippman SM, Lee JJ, et al. Clonal genetic alterations in the lungs of current and former smokers. J Natl Cancer Inst 1997;89:857–62.

(38) Fearon ER. The smoking gun and the damage done: genetic alterations in the lungs of smokers. J Natl Cancer Inst 1997;89:834–6.
(39) Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
(40) Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. New Engl J Med 1997;336:243–50.
(41) Bach PB, Schrag D, Nierman DM, Horak D, White PJ, Young JW, et al. Identification of poor prognostic features among patients requiring mechanical ventilation after hematopoietic stem cell transplantation. Blood 2001;98:3234–40.
(42) Begg CB. The search for cancer risk factors: when can we stop looking? Am J Public Health 2001;91:360–4.
(43) Zauber AG, Winawer SJ, Bond JH, Waye JD, Schapiro M, Stewart ET, et al. Can surveillance intervals be lengthened following colonoscopic polypectomy? Gastroenterology 1997;112:A50–50.
(44) Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 1998;90:1371–88.
(45) Armstrong K, Eisen A, Weber B. Assessing the risk of breast cancer. New Engl J Med 2000;342:564–71.
(46) Friedman GD, Tekawa I, Sadler M, Sidney S. Smoking and mortality: the Kaiser Permanente experience. Changes in cigarette-related disease risks and their implications for prevention and control. Smoking and Tobacco Control Monograph 8. Bethesda (MD): National Cancer Institute; 1997. p. 477–97.
(47) Thun MJ, Day-Lally CA, Calle EE, Flanders WD, Heath CW Jr. Excess mortality among cigarette smokers: changes in a 20-year interval. Am J Public Health 1995;85:1223–30.
(48) Enstrom JE, Heath CW. Smoking cessation and mortality trends among 118,000 Californians, 1960–1997. Epidemiology 1999;10:500–12.
(49) Vineis P, Kogevinas M, Simonato L, Brennan P, Boffetta P. Levelling-off of the risk of lung and bladder cancer in heavy smokers: an analysis based on multicentric case-control studies and a metabolic interpretation. Mutat Res 2000;463:103–10.

NOTES

Supported by a grant from the "Steps for Breath Fund" at Memorial Sloan-Kettering Cancer Center, New York, NY, and by Public Health Service grant RO1CA090226 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services (both to P. B. Bach).

We are indebted to the subjects, study coordinators, and investigators of both CARET and the study of CT scanning for early lung cancer detection being conducted at the Mayo Clinic, Rochester, MN. We also thank Deborah Schrag, Gary Goodman, Gil Omenn, James Jett, and Steve Swenson for their insight and encouragement.

Manuscript received September 17, 2002; revised January 8, 2003; accepted January 21, 2003.