Evidence for Age-specific Genetic Relative Risks in Lung Cancer

W. James Gauderman and John L. Morrison

Recent studies of familial aggregation suggest that family history of lung cancer among first-degree relatives is associated with increased risk for early-onset, but not late-onset, lung cancer. To assess whether this could be explained by variability in genetic relative risk across age, segregation analysis was performed on the Louisiana Lung Cancer Dataset. This data set consisted of 337 probands who died of lung cancer between 1976 and 1979 and their first-degree relatives. A variation of the Cox proportional hazards model was used that allowed estimation of age- and genotype-specific incidence rates, from which the authors obtained estimates of age-specific genetic relative risks. The best-fitting model included an autosomal dominant locus (allele frequency, 0.043), with carrier-to-noncarrier relative risks that exceeded 100 for ages less than 60 years and declined monotonically to 1.6 by age 80. The hypothesis of proportional genetic relative risk across age was rejected ($p = 0.009$). The estimated proportion of persons with lung cancer who carry the high-risk allele exceeds 90% for cases with onset at age 60 years or less and decreases to approximately 10% for cases with onset at age 80 years or older. These findings support previous evidence of a major susceptibility locus for lung cancer and suggest that linkage studies should preferentially recruit young lung cancer cases and their relatives. Am J Epidemiol 1999;151:41-9.

genetics; lung neoplasms; smoking

Lung cancer is the most common cause of cancer death in the United States for both men and women, accounting for 33 percent of cancer deaths in men and 22 percent in women. Between 1974 and 1991, the overall age-adjusted lung cancer incidence rate increased from 37.8 to 68.2 per 100,000 persons. There is some evidence that the rates for males reached a plateau in the 1980s and have since declined, but incidence rates for females appear to still be increasing (1).

Case-control (2, 3) and cohort (4-6) studies have established tobacco smoking as a risk factor for lung cancer. Several case-control studies have indicated that the relative risk due to smoking is higher in females than in males (7-12). However, a recent cohort study has not found such a difference between the sexes (13). Additional environmental agents that have been implicated in lung cancer risk include asbestos (14), environmental tobacco smoke (15), and a variety of other chemical exposures (16).

Several analyses suggest that there are also genetic determinants of lung cancer. Studies of familial aggregation have found familial clustering of the disease (17) and have demonstrated that family history of lung cancer was a significant risk factor after smoking and other risk factors were controlled for (18, 19). In a case-control study of 337 lung cancer probands and their unaffected spouses, relatives of probands were 2.4 times more likely to have lung cancer compared with relatives of their spouses, after age, sex, pack-years of smoking, and occupational exposures were controlled for (20). Analysis of the families of these 337 probands, hereafter referred to as the Louisiana Lung Cancer Dataset, showed evidence of a putative major gene, with an allele frequency of 0.05 and a codominant inheritance pattern (21). In a subsequent analysis of the Louisiana Lung Cancer Dataset, the data were stratified into two groups on the basis of the age of the proband as a surrogate for smoking behaviors pre- and post-World War I (22). Segregation analysis showed different inheritance patterns in the two groups, suggesting the possibility of a gene-smoking interaction. However, direct modeling of an interaction utilizing all families in a single analysis showed no evidence of an interaction between a major gene and pack-years of tobacco exposure (23).

In our previous segregation analysis, the relative increase in lung cancer risk due to a major gene was
assumed to be constant over all ages (23). Results from two case-control association studies indicated that a first-degree family history of lung cancer was associated with increased risk for early-onset, but not late-onset, lung cancer (24, 25). Although these studies cannot distinguish genetic from shared-environment effects, they suggest that the relative effect of a predisposing gene may vary with age. For breast cancer, segregation analysis demonstrated that the effect of genotype on risk varies by age, with an estimated genetic relative risk of 98 for women aged 20–29 years, which declines to a genetic relative risk of only two in women age 80 years or older (26). As with breast cancer, understanding the age-specific dependence of lung cancer risk on genotype is important in developing useful estimates of penetrance for counseling at-risk individuals and in determining which subjects are likely to be the most informative in genetic linkage studies.

In this paper, we utilize the Louisiana Lung Cancer Dataset to investigate whether the dependence of lung cancer on genotype varies by age. Segregation analyses are presented to compare various models of inheritance and to test for evidence of age modification of genetic relative risk. All analyses include adjustment for smoking, and we investigate the sensitivity of the findings to alternative metrics of smoking, including ever/never smoked, packs per day, and pack-years of tobacco exposure. Analysis of gene-smoking interaction is also presented. The best-fitting model is used to compute age-specific cumulative risks of lung cancer and age-specific case-carrier probabilities, stratified by sex and smoking status.

MATERIALS AND METHODS

Louisiana Lung Cancer Dataset

Probands were Caucasians who died of lung cancer in one of 10 Louisiana parishes in the period 1976–1979 (n = 440). First-degree relatives were identified for 337 (77 percent) of the 440 probands, for a total sample size of 4,356 subjects. Pedigrees ranged in size from four to 28 members, and there were 108 familial lung cancer cases in addition to the 337 ever/never smoked, packs per day, and pack-years. The latter is computed as the product of packs per day and number of years of consumption. No information was obtained for smokers on when (in calendar time) they began (or quit) smoking. Smoking status (ever/never) was missing for 2.4 percent of the probands, 14.8 percent of the affected nonprobands, and 21.3 percent of the unaffected nonprobands. Among the nonprobands, there was a larger proportion of missing smoking data in deceased subjects (26.8 percent) than in those living at the time of study (17.6 percent).

Models

Let d denote the disease status for an individual, and let t represent their corresponding age of onset if they are diseased (d = 1) or their last known age unaffected if they are nondiseased (d = 0). Let z denote a vector of measured covariates, for example, smoking status and sex, and let g denote the genotype at an unobserved major locus. We assumed that this major locus was diallelic with high-risk allele A and normal allele a and population frequency $q_A$ of the high-risk allele.

We furthermore assumed Hardy-Weinberg equilibrium, which implies that the frequencies of genotypes AA, Aa, and aa are given by $q_A^2$, $2q_A(1-q_A)$, and $(1-q_A)^2$, respectively.

To account for variability in age of onset, we used a proportional hazards model (27), with modification to allow for nonproportional hazard (incidence) rates with respect to genotype. The age-specific incidence rate, conditional on genotype and covariates, is given by

$$\lambda(t|g, z) = \lambda_0(t, G) \exp(\beta' z).$$

(1)

In this model, the form of the age-specific incidence rate curve is allowed to vary by genetic susceptibility G. For example, under dominant inheritance $G = 1$ if $g = AA$ or Aa and $G = 0$ if $g = aa$, and the genetic relative risk at age t for a carrier compared with a noncarrier is computed as $\lambda_0(t, 1)/\lambda_0(t, 0)$. Each parameter $\beta$ represents the log-relative risk of disease per unit increase in the corresponding covariate. The model can be augmented to include gene-environment interaction terms (28).

Since a particular parametric form for the baseline dependence of lung cancer risk on age is not known,
we modeled it as a step function, i.e.,
\[ \lambda_0(t, G) = \lambda_{Gk} \text{ for } t_{k-1} < t \leq t_k, \]
where \( t_k, k = 1, \ldots, K \) were predefined cutpoints of age. We utilized \( K = 5 \) age intervals, with \( t_1 = 50, t_2 = 60, t_3 = 70, t_4 = 80, \) and \( t_5 = 100. \) As an example, in the dominant model, the baseline parameters are \( \lambda_{11}, \ldots, \lambda_{15} \) for carriers of the high-risk allele (\( G = 1 \)) and \( \lambda_{01}, \ldots, \lambda_{05} \) for noncarriers (\( G = 0 \)). Since there were no lung cancer cases aged 30 years or younger, \( t_0 \) was set to 30, and \( \lambda_0(t) \) was set to zero for \( t \leq 30. \)

**Choice of covariates**

In previous analyses of these data, cumulative pack-years at the time of study was used as the metric of tobacco exposure (21–23). It is well known that treating a time-dependent covariate as a fixed cumulative exposure in analysis of a variable-age-of-onset trait will lead to a biased effect estimates (29) and, in some cases, can make a detrimental exposure appear protective (30). The model in equation 1 could be augmented to allow for a time-dependent covariate, but to properly use pack-years in this way would require knowledge of when a smoker started and stopped smoking. These data are not available in the Louisiana Lung Cancer Dataset.

There are two alternative variables in the data set that can be used as tobacco exposure measures. The first is an indicator variable of ever/never smoked. Although this is also time dependent, it has the advantage that most smokers probably began smoking prior to the age at which they would be at risk for developing lung cancer (around age 30 years). Thus, for the age periods of risk, ever smokers and never smokers are likely to be correctly classified using this variable. The drawback of using ever/never smoked is that it does not include any information about smoking intensity, which has been shown to be an important factor in lung cancer risk. The second variable that can be considered is packs per day. Although this variable provides dose information, a single measure of packs per day will not correctly account for any time-varying patterns in smoking intensity and is more prone to errors in reporting.

Since ever smoking is likely to be the most correctly classified tobacco-exposure variable, we utilized it in our primary analyses. However, we also investigated the sensitivity of our findings to the choice of tobacco-exposure variable, as described below. Preliminary Cox regression analyses using relatives of probands indicated that sex was also a significant risk factor after adjustment for smoking and, therefore, was included as a covariate in all analyses. The proportional hazards assumption for smoking and sex was tested by adding age-covariate interactions to the Cox model (31); neither covariate showed significant deviation from proportionality.

**Comparison of models**

Previous segregation analyses of these data supported either the Mendelian codominant model (21) or the Mendelian dominant model (23). In each of these analyses, a single parameter was included in the model to quantify the age-specific effect of genotype on risk, thus precluding a direct test of whether genetic relative risk varied across age. We performed segregation analysis using the age-specific genetic relative risk model to confirm that Mendelian inheritance was supported under this formulation and for formal testing for age modification of the genetic relative risk.

Segregation analysis was conducted by comparison of the general transmission model (32) with several nested alternatives by using likelihood ratio tests. Included as alternatives were the Mendelian codominant, dominant, and recessive models, as well as the sporadic (no major gene) and environmental (no parent-to-child transmission) models. In general, Mendelian inheritance is supported when both the sporadic and the environmental models can be rejected while one or more of the Mendelian models cannot. To compare nonnested models, Akaike's Information Criteria (AIC) was utilized. AIC was computed as \(-2(\text{log-likelihood}) + 2(\text{number of free parameters})\), and the model with the minimum value was chosen as the most parsimonious. To account for the sampling of pedigrees through a diseased proband, we applied a correction appropriate for single ascertainment (33), i.e., we conditioned the likelihood of each pedigree on the proband being affected at his or her age at examination.

To test whether a model that allows for age-specific genetic relative risks was necessary, the best-fitting model from the segregation analysis was compared with a model in which the effect of genotype was assumed to be proportional across age, i.e.,

\[ \lambda(t | g, z) = \lambda_0(t) \exp\{z'z G\}. \]  

For example, under dominant inheritance, this model constrains the genetic relative risk for carriers (\( G = 1 \)) to noncarriers (\( G = 0 \)) to be \( e^z \) for all ages. Since this is a nested alternative of the model shown in equation 1, a likelihood ratio test was used for comparison, with 4 df under the five-step baseline hazard model described above. To determine whether the test for age-specific genetic relative risks was sensitive to the choice of smoking covariate, we also performed
this likelihood ratio test after replacing ever/never smoked with packs per day and then again replacing it with pack-years and pack-years squared. All analyses were performed using the Genetic Analysis Package (34), modified by the authors to fit the age-specific relative risk model shown in equation 1. A two-sided significance level of 0.05 was used for all hypothesis testing.

**Estimation of cumulative risks and genotype probabilities**

Suppressing subscripts and letting $\Omega = \{\beta, \lambda_0, \lambda_2, \ldots\}$ denote the parameters of the disease model, the penetrance for a given subject is given by

$$f(d, t | g, z, \Omega) = \lambda(t | g, z, \Omega)S(t | g, z, \Omega).$$

The survival function, $S(t | g, z, \Omega) = \exp(-\int_0^t \lambda(s | g, z, \Omega)ds)$, is the probability of remaining disease free up to age $t$. The probability of being afflicted with lung cancer prior to age $t$ for a subject with genotype $g$ and covariate vector $z$ is then given by $1 - S(t | g, z, \Omega)$. The probability that a diseased subject with age of onset $t$ carries genotype $g^*$ at a major lung cancer susceptibility locus, conditional on his/her covariate vector $z$ and the maximum likelihood estimates of the parameters $(\hat{\Omega}, \hat{q}_A)$, was computed with Bayes formula as

$$Pr(g^* | d = 1, t, z, \hat{\Omega}, \hat{q}_A) = \frac{f(d = 1, t | g^*, z, \hat{\Omega})Pr(g^* | \hat{q}_A)}{\sum_f f(d = 1, t | g, z, \hat{\Omega})Pr(g | \hat{q}_A)}.$$

The prior probability $Pr(\cdot | \hat{q}_A)$ was computed under Hardy-Weinberg equilibrium, and the sum in the denominator was taken over all three genotypes.

**RESULTS**

The distributions of age, sex, and smoking, stratified by proband and disease status, are shown in table 1. The mean age of onset for both probands and nonproband cases was approximately 64 years. The age of onset was age 60 years or younger for 35 percent of the probands and 39 percent of the nonprobands. A history of smoking was reported for 92 percent of the probands, with a mean of 1.8 packs per day, and for 82 percent of the nonproband cases, with a mean of 1.9 packs per day. Segregation analysis results are shown in table 2. Both the sporadic ($p = 0.045$) and the environmental ($p < 0.0001$) transmission models were rejected, but none of the Mendelian alternatives could be rejected. Based on the AIC, the dominant model provides the most parsimonious fit to the data. We found no evidence for deviation from Hardy-Weinberg equilibrium.
Weinberg equilibrium in the dominant model \( (p = 0.31) \).

Table 3 shows the parameter estimates and log-likelihoods from the best-fitting dominant model and from the corresponding dominant model in which the genetic relative risk is constrained to be proportional across age. The estimated relative risk for ever smokers was 4.9 \( (p < 0.0001) \) in the proportional model.
and 4.3 ($p < 0.0001$) in the age-specific model, and males showed a significant increase in risk relative to females ($p < 0.05$ in each model). In the proportional model, the single genetic relative risk estimate was 13.6, and the estimated frequency of the high-risk allele was 2.5 percent. In the age-specific model, the allele frequency estimate was higher (4.3 percent), and the estimated genetic relative risks varied substantially with age, from 138.1 and 123.0 for the age groups 31–50 and 51–60 years, respectively, decreasing to 1.6 in the age group 80 years or more. Comparison of the two models indicates that the age-specific model fit the data better did than the proportional model ($\chi^2 = 13.64, p = 0.009$), i.e., that genetic relative risk for lung cancer varies significantly with age.

Table 4 shows results of additional analyses to determine whether the comparison of the age-specific and proportional models was sensitive to the choice of adjustment covariates. For the model with no adjustment (model 1) and that with adjustment for sex only (model 2), no subjects were excluded because of missing data on smoking. In both cases, the proportional genetic relative risk model can be rejected ($p = 0.003$ and $p = 0.007$, respectively). Model 3 is a summary of the result shown in table 3, while models 4 and 5 show that the proportional model can also be rejected when either packs per day ($p = 0.006$) or pack-years ($p = 0.035$) are utilized as the tobacco exposure covariates.

As described above, previous analysis of these data showed no evidence for a gene $\times$ pack-years interaction under the proportional genetic risk model (23). Since the current analysis supported the use of both a different type of model (age-specific genetic relative risks) and a different smoking variable (ever smoking), we revisited the question of whether there was evidence for a gene-smoking interaction. To do this, we added a single gene-smoking interaction term to the age-specific model shown in table 3. The estimated interaction relative risk was 4.6, suggesting that gene carriers are more susceptible to the effects of tobacco exposure than are noncarriers. However, the 1 df likelihood ratio test of the interaction coefficient yielded $\chi^2 = 2.64 (p = 0.10)$, indicating a nonsignificant interaction at the 5 percent significance level.

Figure 1 shows plots of age-specific cumulative lung cancer risk by carrier and smoking status for males and females based on estimates for the age-specific model in table 3. The estimated lifetime risk (at age 80 years) of lung cancer for the highest risk group, male carrier smokers, is 42.1 percent, while it is only 1.2 percent in the lowest risk group, female non-carrier nonsmokers. For both males and females, the estimated cumulative risk for gene carriers increases steadily between ages 30 and 80 years, while for non-carriers the estimated cumulative risk is nearly zero until age 60, but increases steadily thereafter. For two groups, carrier nonsmokers (upper solid line) and non-carrier smokers (lower dashed line), the cumulative risks at age 85 years are nearly equivalent, although their trajectories follow quite different patterns.

Table 5 shows estimates of the probability that a person with lung cancer carries a high-risk genotype, conditional on his or her age of onset, sex, and smoking status. For cases aged 60 years or less, the estimated probability that they carry a susceptibility genotype exceeds 90 percent, regardless of their sex or smoking history. The probabilities decline with age of onset, such that approximately 10 percent of cases are predicted to be carriers if their age of onset is greater than 80 years. As might be expected, diseased subjects with another risk factor, in this case, smoking or male sex, are less likely to be gene carriers. However, the variation in carrier probabilities with respect to sex and smoking status is substantially less than the variation observed with differing ages of onset.

**DISCUSSION**

Our results support previous findings that a major gene plays an important role in the etiology of lung cancer. We have demonstrated that the effect of genotype on lung cancer risk appears to vary by age, with age-specific relative risks greatest at young ages and declining thereafter. Our estimate of the allele frequency (4.3 percent) is comparable with estimates found in previous analyses of these data (21, 23). These findings have implications for the design of future linkage studies aimed at localizing a major locus (or loci) for lung cancer. Linkage studies will have power to detect a linked locus when the disease of the affected subjects in the
FIGURE 1. Age-specific cumulative risk of lung cancer for females (top) and males (bottom) by genotype and smoking status, based on the parameter estimates for the age-specific genetic relative risk model shown in table 3. Results derive from analysis of the Louisiana Lung Cancer Dataset, which consists of 337 probands who died of lung cancer between 1976 and 1979 and their first-degree relatives. The upper two lines are for genetically susceptible subjects (g = AA or Aa), and the lower two lines are for genetically normal subjects (g = aa).
4.3 for smokers compared with nonsmokers may be an underestimate of the true effect of tobacco exposure on lung cancer risk. Mismeasurement of smoking history can lead to biased estimates of a major gene effect if the degree of mismeasurement both is associated with smoking history and is correlated within families. The main effects of smoking are not missing completely at random (35, 36). A Gibbs sampling approach for imputing missing covariate data in pedigree studies has been described (37), but it is not conducive to model comparison and, therefore, was not utilized in this paper.

Even for subjects with smoking data, there may be some error in reporting smoking history. Nondifferential error in a measured covariate produces parameter estimates that are biased toward zero in a proportional hazards model (38). Hence, the estimated relative risk of 4.3 for smokers compared with nonsmokers may be an underestimate of the true effect of tobacco exposure on lung cancer risk. Mismeasurement of smoking history may also lead to biased estimates of a major gene effect if the degree of mismeasurement both is associated with smoking history and is correlated within families. Biases in the major gene effect estimate can also occur if there are other unmeasured factors that are correlated within the family and are related to disease risk. Factors that may fall into this category include secondhand smoke, dietary factors, radon exposure, and other genetic risk factors, such as a second major gene. However, the strong major gene effect observed in our analysis is not likely to have been solely due to mismeasurement of smoking or failure to model additional environmental or genetic factors.

We did not find a statistically significant interaction between the major gene and ever smoking. It is well known that the power to detect interactions between measured covariates is less than the power to detect main effects (39). The power to detect a $G \times E$ interaction in segregation analysis is further reduced due to the fact that the gene is unmeasured (40). Additionally, the interaction coefficient in our model only parameterizes deviations from a multiplicative effect of the gene and smoking on disease risk, and the magnitude of this interaction is assumed to be constant across age. Since we have evidence that the main effect of the gene varies with age, it is possible that a gene-smoking interaction effect may also be age dependent. However, to address this question would require modification of the model to allow both the genetic and smoking parameters to vary with age, which was not possible given the sample sizes in this data set. For these reasons, we cannot rule out the possibility that a gene acts to modify the effect of smoking on lung cancer risk.

In summary, we found evidence for a major gene that increases lung cancer risk, with relative risk that varies with age in a manner similar to that observed for breast cancer. Although the genetic relative risk is estimated to be quite high at young ages, the actual number of young cases attributable to the gene is low, primarily due to the low baseline incidence rates at these ages. Linkage studies should attempt to focus on lung cancers at young ages (less than age 60 years) to maximize the yield of gene carriers, with recruitment from multiple centers to achieve adequate sample sizes. Once a gene (or genes) is identified, the dependence of disease risk on age and the potential interaction between the gene and tobacco exposure should be reconsidered, both for understanding disease etiology and for counseling at-risk individuals.

**ACKNOWLEDGMENTS**

Supported by National Cancer Institute grant CA 52862 and National Institute of Environmental Health Science grant SP30 ES07048–03.

The authors thank Dr. Henry Rothschild for providing the Louisiana Lung Cancer Dataset.
REFERENCES