Lung Cancer Mortality Is Related to Age in Addition to Duration and Intensity of Cigarette Smoking: An Analysis of CPS-I Data

James D. Knoke,[1] Thomas G. Shanks,[1] Jerry W. Vaughn,[1] Michael J. Thun,[2] and David M. Burns[1]

[1]Department of Family and Preventive Medicine, University of California at San Diego, San Diego, California and [2]Department of Epidemiology and Surveillance, American Cancer Society, Atlanta, Georgia

Abstract

Objectives: Models previously developed for predicting lung cancer mortality from cigarette smoking intensity and duration based on aggregated prospective mortality data have employed a study of British doctors and have assumed a uniform age of initiation of smoking. We reexamined these models using the American Cancer Society's Cancer Prevention Study I data that include a range of ages of initiation to assess the importance of an additional term for age. Methods: Model parameters were estimated by maximum likelihood, and model fit was assessed by residual analysis, likelihood ratio tests, and χ^2 goodness-of-fit tests. Results: Examination of the residuals of a model proposed by Doll and Peto with the Cancer Prevention Study I data suggested that a better fitting model might be obtained by including an additional term specifying the ages when smoking exposure occurred. An extended model with terms for cigarettes smoked per day, duration of smoking, and attained age was found to fit statistically significantly better than the Doll and Peto model ($P < 0.001$) and to fit well in an absolute sense (goodness-of-fit; $P = 0.34$). Finally, a model proposed by Moolgavkar was examined and found not to fit as well as the extended model, although it included similar terms (goodness-of-fit; $P = 0.007$). Conclusions: The addition of age, or another measure of the timing of the exposure to smoking, improves the prediction of lung cancer mortality with Doll and Peto's multiplicative power model. (Cancer Epidemiol Biomarkers Prev 2004; 13(6):949–57)

Introduction

Eighty to ninety percent of lung cancer mortality is attributable to cigarette smoking, and the relative risks for lung cancer increase with both the number of cigarettes smoked per day (CPD) and the duration of smoking (1). Doll and Peto (2) proposed a model based on a multistage theory of carcinogenesis (3) that used CPD and duration of smoking (in years) to estimate lung cancer risk. This model was fit to aggregated data on lung cancer mortality from the 20-year follow-up of the British Doctors Study (4). Doll and Peto included only subjects who smoked no more than 40 CPD and who began smoking between ages 16 and 25; they assumed a fixed age of initiation of 19 for all subjects. A different biological, ''two-stage'' model was proposed by Moolgavkar et al. (5) and fit using these same data. The Moolgavkar model assumed that the consequences of a fixed intensity of smoking are a nonlinear function of age but less for individuals under age 20 than for those over 20. The model included terms for CPD, age of initiation of smoking, and attained age. However, because the model was fit using the same data as Doll and Peto, a fixed age of initiation of 19 was assumed for all subjects.

In the present report, we use data from the American Cancer Society's Cancer Prevention Study I (CPS-I; ref. 6) to develop new parameters for these models. The CPS-I has considerably more deaths from lung cancer than the British Doctors Study and contains age of initiation of smoking as well as attained age. Using the CPS-I data, we then perform a statistical comparison of the Doll and Peto model, the Moolgavkar model, and an extended model with an additional term for attained age.

Materials and Methods

The CPS-I. In 1959, 1,078,894 subjects were recruited by the American Cancer Society to participate in the 12-year CPS-I prospective mortality study (7). Follow-up questionnaires, which assessed mortality and continuing smoking status, were administered after ~2, 4, and 6 years and at the conclusion of follow-up. Mortality was additionally assessed at 1, 3, 5, and 11 years. To parallel the British Doctors Study, the primary study group for this report was the subgroup of white, male, current cigarette smokers ages 40 to 79, who initiated smoking by age 35 and did not also smoke cigars or pipes. Age, age at initiation of cigarette smoking (initiation), and dose

Received 7/16/03; revised 1/9/04; accepted 2/2/04.

Grant support: American Legacy Foundation.

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

Requests for reprints: James D. Knoke, Tobacco Control Policies Project, University of California at San Diego, Suite 310, 1545 Hotel Circle South, San Diego, CA 92108. Phone: 619-294-3708; Fax: 619-220-0228. E-mail: jknoke@ucsd.edu

(CPD) were assessed by questionnaire at the baseline visit, and it was these values that were used in our analyses. Initiation and dose were categorized on the questionnaire, as detailed in Table 1. Descriptive analyses of the mortality data collected by this study have been presented previously (7-9).

Eighty-eight percent of the primary study group (n = 174,993) either died (26%) or were successfully followed for the entire 12 years. Death certificates were requested from the appropriate state heath departments and were obtained for ~95% of deaths. Lung cancer listed as the primary, secondary (contributing), or tertiary (any mention) cause of death on the death certificate was the outcome for our analyses.

There were 163,643 white males who continued to smoke throughout their period of observation and had complete data on age, initiation, dose, and outcome. Of these, 3,405 subjects died of lung cancer during follow-up. Those continuing smokers ages 40 to 79 during some portion of the follow-up period, who initiated smoking by age 35, were tabulated into 1,200 cells (40 × 6 × 5, age × initiation × dose). This tabulation advanced age and duration with year of follow-up and censored subjects when they died of causes other than known lung cancer or were lost to follow-up. Intensity of smoking (dose) as reported at the baseline evaluation was assumed to continue throughout follow-up.

We also studied a secondary group of 92,307 subjects who were white, male, and had never smoked any tobacco product to assess the rate of lung cancer in the absence of tobacco smoking. Of these, 215 died of lung cancer during follow-up. The nonsmoker subgroup has been studied before, with some variations, by Whittemore (9), Garfinkel and Silverberg (10), Burns et al. (8), Thun et al. (11), and Leenhouts (12). Whittemore reported fewer deaths, so may have used only lung cancer as the primary cause of death in her analysis. Leenhouts used only the first 6 years of follow-up of the CPS-I.

The Doll and Peto Model. The Armitage and Doll (3) multistage model of carcinogenesis implies that a cell progresses to malignancy through a sequence of states of a Markov chain. As applied to the effect of cigarette smoking on lung cancer risk, it assumes that the number of events occurring in each cell is Poisson distributed with incidence (mean value) a specialized multiplicative power function of duration and dose. Specifically, the incidence is a function of three parameters a, b, and c:

$$\mu_1(a,b,c,) = a \cdot (\text{dose} + 6)^b \cdot (\text{duration} - 3.5)^c \quad \text{(A)}$$

Doll and Peto (2) found that parameter values of b = 2.0 and c = 4.5 fit the British doctors' data well. Peto (13) noted that these values imply that duration of smoking has a more profound effect on lung cancer risk than does dose. The constants 6 and 3.5 differentiate the Doll and Peto model from a simple Poisson model for which the incidence would be:

$$\mu_0(a,b,c,) = a \cdot (\text{dose})^b \cdot (\text{duration})^c$$

The constant 6 accounts for the background risk of lung cancer in the absence of smoking, and the constant 3.5 accounts for the time lag between a single cell first becoming a cancer and death from subsequent growth of that cancer. Doll and Peto considered models with constants other than 6 and 3.5 and reported that the fit to these alternative models was similar that of model A, although the estimated parameters took on different values. Consequently, these constants are somewhat arbitrary. In particular, a value greater than 3.5 may be more appropriate for the second constant (9). In this article, however, model A is always used with the constants 6 and 3.5 to allow comparison with the original Doll and Peto analysis. Notice that, by definition, this model is intended to be applied to smokers; nonsmokers have duration of zero; thus, the third term would be negative 3.5 raised to a power.

The Moolgavkar Model. Moolgavkar et al. proposed an alternate biological model for carcinogenesis (14, 15) and fitted this ''two-stage'' model to observed lung cancer death rates using the British doctors' data (5). Like the Doll and Peto model, Moolgavkar's model assumes that events in each cell follow the Poisson distribution, although the incidence is a complicated nonlinear function of age, initiation, and dose. One feature of Moolgavkar's model is the assumption that the adverse health consequences of a fixed intensity of smoking for a fixed period are less for smokers under age 20 than for those over age 20. This is based on the presumption that the number of susceptible lung tissue cells continues to increase until ~age 20 and then remains constant. The incidence function is:

$$\mu_2(t,d) = (c_0 + c_2 d)\{c_0 \exp[bd(t - t_0)] \cdot \int_0^{t_0} X(s)\exp[a(t-s)]ds + (c_0 + c_1 d)\int_{t_0}^{t} X(s)\exp[(a+bd)(t-s)]ds\} \quad \text{(B)}$$

where a, b, c_0, c_1, and c_2 are model parameters, d is dose, t_0 is age of initiation, t is attained age, and $X(s)$ is the mean number of susceptible cells at age s (see ref. 15

Table 1. Categorizations of age of initiation of smoking and CPD

Range	Coded as
Age of initiation (y)	
≤9	7
10-14	12
15-19	17
20-24	22
25-29	27
30-34	32
CPD	
1-9	5
10-19	15
20	20
21-39	30
≥40*	45

*The categories of 40 and >40 in the original data have been combined for our analyses due to small frequencies.

for a detailed description of this function). It is plausible to include nonsmokers in an analysis of this model. Dose can be set to zero and attained age to t_0; the second integral then evaluates to zero.

Moolgavkar described the parameters in terms of the first-mutation and second-mutation rates ($c_0 + c_1 d$ and $c_0 + c_2 d$) per lung tissue cell per year of smoking exposure and the net proliferation rate of intermediate cells ($a + bd$). Due to the nonlinear incidence, none of the five parameters are uniquely associated with an individual explanatory variable (age, initiation, or dose). However, Moolgavkar refers to b, c_1, and c_2 as dose-related parameters because of their multiplicative relation to dose. When fit to the British doctors' data (with nonsmokers included in the analysis), Moolgavkar's initial estimate of the parameter b was a small negative value; consequently, he set b to zero. Further analysis by Moolgavkar showed that a similar fit to the data could be obtained by setting any one of the three dose-related parameters, b, c_1, or c_2, to zero or by setting $c_1 = c_2$. These results suggest that the original, five-parameter model may be overparameterized.

Extended Models. The effects of a given intensity and duration of smoking may vary depending on the age at which smoking occurs. It has been suggested that the lung may be either more (16) or less (5) susceptible to a given carcinogenic exposure early in life. An increased susceptibility to carcinogenic exposures with advancing age has also been postulated (17, 18). We tested whether the addition of a term for age of initiation or attained age to the incidence function of the Doll and Peto model leads to a better fitting model. While age of initiation, attained age, and duration of exposure can all be postulated to have independent biological effects, it is not possible to include all three terms in a mathematical model because including any two fixes the value for the third. We examined all three combinations of these terms in the models presented below. When any two terms are used in the model, they specify the duration of smoking and where that duration occurred in the life span of the smoker:

$$\mu_3(a,b,c,d) = a \cdot (\text{dose} + 6)^b \cdot (\text{duration} - 3.5)^c \cdot \text{initiation}^d \quad \text{(C)}$$

$$\mu_4(a,b,c,d) = a \cdot (\text{dose} + 6)^b \cdot (\text{duration} - 3.5)^c \cdot \text{age}^d \quad \text{(D)}$$

$$\mu_5(a,b,c,d) = a \cdot (\text{dose} + 6)^b \cdot (\text{age} - 3.5)^c \cdot \text{initiation}^d \quad \text{(E)}$$

Models C and D are of interest because model A is a *reduction* of each. This nesting allows likelihood ratio testing of whether they statistically significantly better fit the data than model A. Models C to E all use the same *statistical information* and have the same degrees of freedom, but one model may fit better than another due to the form of the model and its assumptions. Comparisons between these models cannot be made by likelihood ratio testing, however, because there is no nesting among them. Similarly to the Doll and Peto model, it is preferable to include only smokers in analyses of these models.

A Poisson Model for Nonsmokers. This model assumes that the number of events occurring in each 1-year age interval is Poisson distributed with incidence a simple power function of age:

$$\mu_6(a,b) = a \cdot (\text{age} - 3.5)^b \quad \text{(F)}$$

The constant 3.5 was used for consistency with the Doll and Peto and extended models. This model differs slightly from the Poisson model for nonsmokers reported by Burns et al. (8), which did not subtract 3.5 years from age.

Statistical Methods. The Doll and Peto and extended models were fit using PROC GENMOD of the SAS System (Cary, North Carolina), which evaluated maximum likelihood estimates of the model parameters and likelihood ratio tests of nested models. The likelihood ratio tests evaluated whether a reduced model fits the data as well as a specific alternative model with additional parameters (19). The Moolgavkar model was fit with PROC NLIN of the SAS following the approach described by Jennrich and Ralston (20) for obtaining maximum likelihood estimates. Confidence intervals for the parameters were estimated by the Wald approximation (19).

χ^2 goodness-of-fit tests of models were performed on a reduced number of combined cells (353 rather than 1,200) to meet the minimum expected cell frequencies suggested by Cochran (21) for goodness-of-fit testing. The algorithm used for cell combination is described in Appendix 1. The goodness-of-fit test assessed the overall fit of the data to the model; small P values for the test indicated that the data do not fit the model well. These overall tests are absolute and not relative to another model as are the likelihood ratio tests.

Graphical residual analysis employing standardized residuals, the signed square roots of the contributions to χ^2 for the cells, was also performed using the combined cells. The residual value for a combined cell was associated with the weighed averages of initiation and dose, with weights equal to the expected frequency for the original cell divided by the sum of the expected frequencies for all cells combined. For each residual plot, the linear regression coefficient was tested for nonzero slope.

Results

Replication of Prior Results Using the British Doctors Study. We fit the data from the British Doctors Study with our computer routines to ensure that the routines were programmed correctly, to compare estimation techniques for the Doll and Peto model, and to explore three-parameter and four-parameter Moolgavkar models. The results are displayed in Table 2. Our maximum likelihood estimates for the parameters of the Doll and Peto model, using only the data on smokers, are similar to those obtained by graphical interpolation of indirectly age-standardized relative risk estimates on the same data (2). The confidence intervals for our maximum likelihood estimates included the values obtained by the more heuristic method of Doll and Peto.

Table 2. Models fitted to the British doctors' data: Parameter estimates (95% confidence intervals)

Doll and Peto model (formula A), with smokers only

	a (Constant)	b (Dose)	c (Duration)
Doll and Peto (1978)*	2.73×10^{-12}	2	4.5
Present report	7.17×10^{-13} [(0.43-119) $\times 10^{-13}$]	1.88 (1.44-2.33)	4.37 (3.75-4.98)

Moolgavkar model (formula B)

	a	c_0	c_1	c_2
Moolgavkar et al. (1989)	0.114 (0.098-0.130)	6.51×10^{-8} [(3.02-10.0) $\times 10^{-8}$]	8.34×10^{-8} [(−10.6 to 27.3) $\times 10^{-8}$]	5.49×10^{-9} [(3.93-7.05) $\times 10^{-9}$]
Present report (with nonsmokers)	0.114 (0.099-0.128)	6.59×10^{-8} [(3.19-9.98) $\times 10^{-8}$]	7.73×10^{-8} [(−12.0 to 27.5) $\times 10^{-8}$]	6.12×10^{-9} [(−12.7 to 25.0) $\times 10^{-9}$]
Present report (smokers only)	0.113 (0.099-0.128)	7.54×10^{-8} [(1.79-13.3) $\times 10^{-8}$]	4.97×10^{-8} [(−11.1 to 21.2) $\times 10^{-8}$]	9.44×10^{-9} [(−21.9 to 40.8) $\times 10^{-9}$]
Present report (with nonsmokers)	0.115 (0.101-0.130)	6.23×10^{-8} [(3.10-9.36) $\times 10^{-8}$]	2.08×10^{-8} [(1.48-2.68) $\times 10^{-8}$]	†
Present report (smokers only)	0.114 (0.099-0.128)	6.77×10^{-8} [(1.23-12.3) $\times 10^{-8}$]	2.11×10^{-8} [(1.40-2.82) $\times 10^{-8}$]	†

*Doll and Peto did not obtain estimates by maximum likelihood and did not report confidence intervals.
†Models with c_1 set equal to c_2.

Our results for the Moolgavkar model differed somewhat from the four-parameter maximum likelihood estimates obtained by Moolgavkar et al. (5), which included nonsmokers. Our parameter estimates are within the confidence intervals of Moolgavkar's estimates; however, our SE for c_2 is 10 times larger than Moolgavkar's. This difference may be due to overparameterization, which could result in an unstable iterative estimation process. Our analysis disclosed a substantial negative correlation between c_1 and c_2 (−0.98) and confidence intervals for both c_1 and c_2, which included zero, implying redundancy between c_1 and c_2. With nonsmokers excluded, the results similarly suggested overparameterization. A three-parameter Moolgavkar model with c_1 set equal to C_2 conversely showed no evidence of overparameterization with either nonsmokers included or excluded (Table 2).

Lung Cancer among CPS-I Nonsmokers. The lung cancer mortality rate among nonsmokers in the CPS-I is presented in Fig. 1 as a function of 5-year intervals of age. Model F closely fit the mortality rate, with parameter estimates $\hat{a} = 5.29 \times 10^{-13}$ and $\hat{b} = 4.83$ and goodness-of-fit $\chi^2 = 15.71$ (df = 25, $P = 0.92$). The estimate of the parameter b is greater than that reported by Whittemore (9) for a similar model. However, Whittemore subtracted 5 years instead of 3.5 from age and included fewer deaths in her analysis.

Doll and Peto and Extended Models for the CPS-I Smokers. Parameter estimates for the Doll and Peto model on the 1,200 cells of the CPS-I male current smokers data are reported in Table 3. Both exponential parameters are highly statistically significantly different from zero but of lower magnitude for the CPS-I than for the British Doctors Study. The value of the χ^2 statistic for duration ($\chi^2 = 1{,}952.3$) is more than five times the value of that for dose ($\chi^2 = 373.3$), which is consistent with Peto's (13) observation that duration has the more profound effect on lung cancer risk. Analyses of residuals, however, indicated that the fit of model A was lacking in both dimensions of age of initiation and attained age (Fig. 2).

We then fit extended models C to E to the CPS-I data and compared them with model A. Because model A is

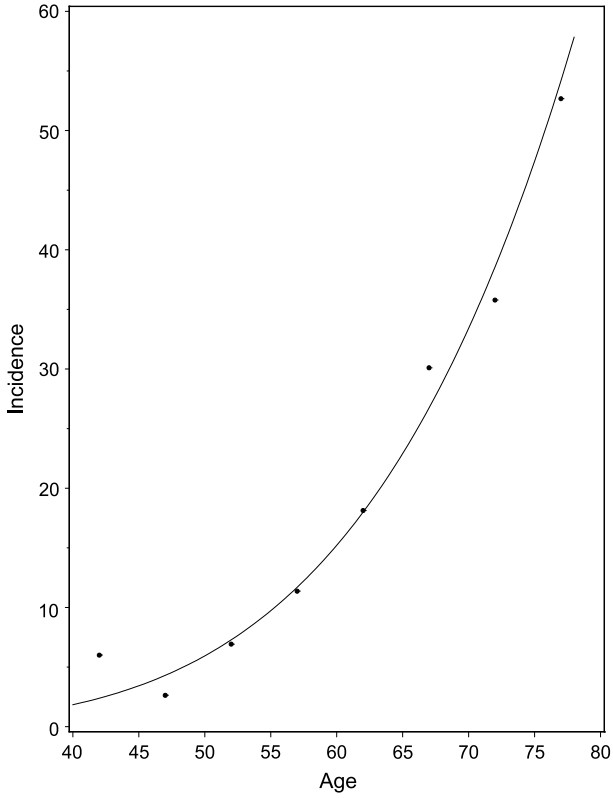

Figure 1. Incidence of fatal lung cancer, nonsmokers. By 5-year age intervals and fit by the Poisson model.

Table 3. Doll and Peto and extended models with CPS-I data, current smokers: Parameter estimates (95% confidence intervals)

Model*	a (Constant)	b (Dose)	Duration	Initiation	Age	LR†	$\chi^{2‡}$	P (χ^2)
A	1.01×10^{-10} [(0.48-2.10) $\times 10^{-10}$]	0.96 (0.86-1.05)	3.74 (3.58-3.91)	—	—	—	446.0	<0.001
C	3.41×10^{-12} [(1.18-9.8) $\times 10^{-12}$]	1.02 (0.92-1.12)	4.08 (3.90-4.27)	0.68 (0.53-0.83)	—	75.1	366.8	0.23
D	2.21×10^{-13} [(0.49-9.95) $\times 10^{-13}$]	1.02 (0.95-1.12)	2.35 (2.02-2.68)	—	2.68 (2.11-3.25)	85.6	358.3	0.34
E	3.93×10^{-14} [(1.15-13.5) $\times 10^{-14}$]	1.03 (0.93-1.13)	—	−0.87 (−1.00 to −0.75)	5.88 (5.62-6.14)	—	379.6	0.12

*Models are defined in Materials and Methods.
†Likelihood ratio χ^2 test that this model fits as good as model A; df = 1, $P < 0.001$.
‡χ^2 goodness-of fit test; df = 349 for model A and 348 for other models.

nested in models C and D, likelihood ratio tests could be performed. Both models C and D statistically significantly better fit the data than model A. Model E also appeared to fit the data well, although a formal test of model E versus model A was not possible. However, model E was compared by likelihood ratio with its two nested reduced models, those without attained age or age of initiation, and was found to fit the data statistically significantly better than either reduced model (results not shown). χ^2 goodness-of-fit tests for models A and C to E, using the 353 combined cells described in Appendix 1, suggested (Table 3) that the extended models fit the data rather well in an absolute sense, while model A did not fit well.

Although models C to E could not be compared with each other by formal statistical hypothesis testing, we did add an interaction (cross-product) term to each (results not shown). There was a significant interaction between duration and initiation in model C. The interactions were not significant between duration and age in model D or between age and initiation in model E. We selected model D for further analysis based on the likelihood ratio tests, the goodness-of-fit tests, and the logical coherence of a model with parameters for dose, duration, and attained age. This model is subsequently called the extended D-P model. The residual plots with respect to age of initiation and attained age for model D are presented in Fig. 3 and are essentially flat.

Moolgavkar's Model for the CPS-I Smokers. Moolgavkar's model was also fit to the 1,200 cells of white male current smokers. This model, at the outset, encompasses five parameters, c_0, c_1, c_2, a, and b. However, after finding it to be negative in a preliminary analysis of the British doctors' data, Moolgavkar set b to zero. We performed a similar preliminary analysis using the CPS-I cells and found a similar negative estimate of b. Consequently, we set b to zero and examined a reduced, four-parameter model.

The reduced, four-parameter model converged to a negative estimate of c_2 that, although small, was statistically significantly less than zero (Table 4). In addition, there were substantial negative correlations, −0.84 between c_0 and c_1 and −0.63 between c_1 and c_2. The negative estimates and large negative correlations suggested that the four-parameter model was still overparameterized. We then considered three-parameter models and obtained reasonable estimates with either c_1 or c_2 set to zero. The model with c_1 set equal to c_2 fitted the CPS-I data more

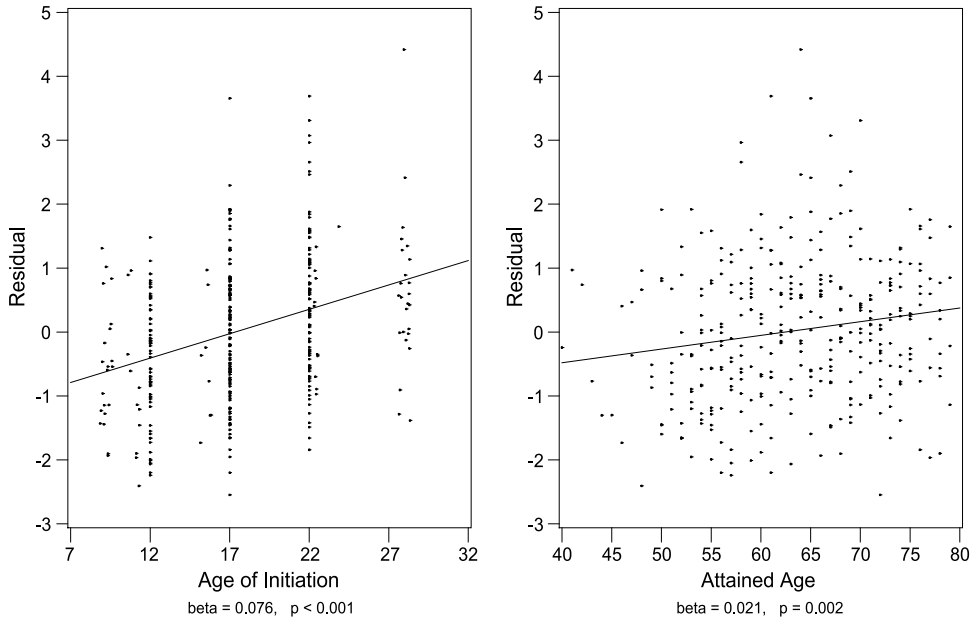

Figure 2. Residuals of the Doll and Peto model. *Left,* by age of initiation; *right,* by attained age.

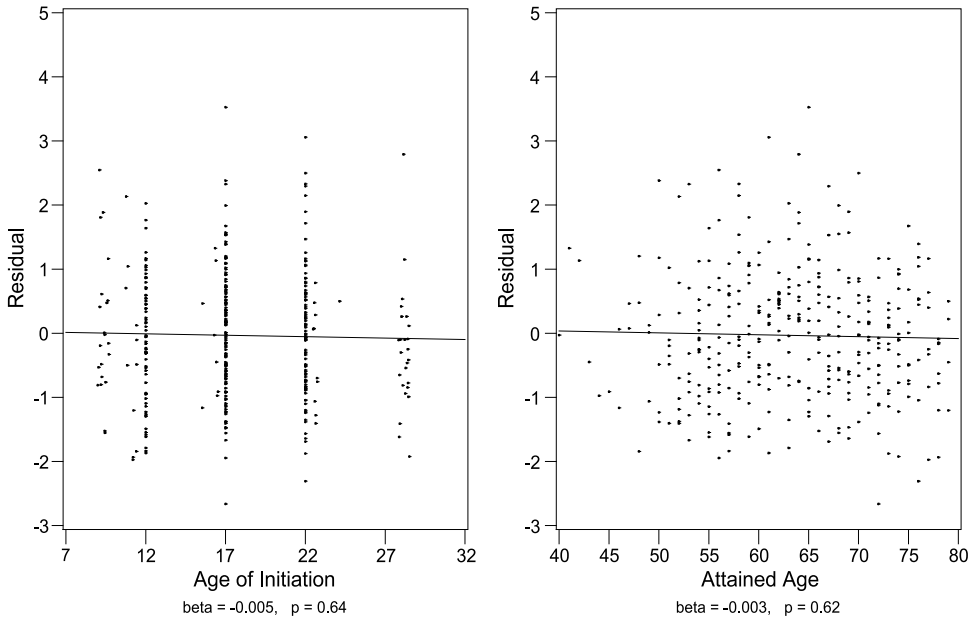

Figure 3. Residuals of the extended D-P model. *Left,* by age of initiation; *right,* by attained age.

poorly than either of the other three-parameter models. We arbitrarily reported the model with $c_2 = 0$, because this version exhibited slightly better fit than that with $c_1 = 0$.

χ^2 goodness-of-fit tests suggested that the three-parameter Moolgavkar model fitted the data as least as well as the four-parameter model, better than the Doll and Peto model, not as well as the extended D-P model, and not well in an absolute sense. Residual analysis indicated that, similarly to the Doll and Peto model, the fit of the three-parameter model was lacking primarily in the dimension of age of initiation (Fig. 4), although age of initiation was a term in the model.

The three-parameter and four-parameter versions of the Moolgavkar model were also fit with the 40 additional cells of nonsmokers included to parallel the approach taken by Moolgavkar et al. (5). The results (Table 4) were similar to the results based on only the smokers' data, although there are small but statistically significant differences in the parameter estimates.

Discussion

The Doll and Peto model has proven to be a valuable tool for estimating lung cancer rates; it has been cited by over 300 research reports (as reported in the *Science Citation Index*) and been used in U.S. Environmental Protection Agency risk assessments (22, 23). However, it has also been found to not fit data from other populations as well as it fits the British doctors' data. In particular, Mizuno et al. (24) found the maximum likelihood estimates of the parameters b and c for male Japanese smokers to be smaller than Doll and Peto found for the British doctors. In our analyses of the CPS-I data, the parameters b and c are also found to be smaller than for the British data and there is a significant improvement in fit when a term for attained age is added.

A recent report by Flanders et al. (25), using data from the American Cancer Society's Cancer Prevention Study II, fitted two-parameter (duration and intensity), simple Poisson models stratified by decade of attained age. They concluded that their results confirmed Peto's (13) observation that duration of smoking is more important than intensity (CPD) in predicting lung cancer. They also found that the estimated coefficients for both duration and intensity decreased with increasing age. Thus, their two-parameter, stratified modeling approach using Cancer Prevention Study II data is generally consistent with our conclusion, using CPS-I data, that age is an important third parameter in predicting lung cancer risk.

Table 4. Moolgavkar Models with CPS-I Data: Parameter Estimates (95% Confidence Intervals)

	a	c_0	c_1	c_2	χ^{2*}	P (χ^2)
Smokers only	0.094 (0.089-0.098)	1.62×10^{-7} [(1.24-2.00) $\times 10^{-7}$]	1.41×10^{-7} [(0.82-2.00) $\times 10^{-7}$]	-7.11×10^{-10} [(-12.1 to -2.12) $\times 10^{-10}$]	423.0	0.004
Both smokers and nonsmokers	0.091 (0.088-0.095)	1.17×10^{-7} [(1.03-1.30) $\times 10^{-7}$]	2.43×10^{-7} [(2.11-2.75) $\times 10^{-7}$]	-7.64×10^{-10} [(-10.3 to -4.95) $\times 10^{-10}$]	—	—
Smokers only	0.095 (0.091-0.099)	1.82×10^{-7} [(1.55-2.10) $\times 10^{-7}$]	9.69×10^{-8} [(7.71-11.68) $\times 10^{-8}$]	†	417.5	0.007
Both smokers and nonsmokers	0.093 (0.089-0.097)	1.16×10^{-7} [(1.03-1.30) $\times 10^{-7}$]	1.89×10^{-7} [(1.71-2.07) $\times 10^{-7}$]	†	—	—

*χ^2 goodness-of fit test; df = 348 for the first model and 349 for the second model.
†c_2 set to zero.

That the estimated coefficient for age in the extended D-P model is (significantly) positive suggests that a given duration of smoking might be more hazardous when experienced later in life. There are several possible explanations for this observation. There may be an accumulation of exposure to other lung carcinogens as is apparent in nonsmokers, and there may be an increased susceptibility to carcinogenic exposure with advancing age. Another possibility, suggested by Moolgavkar et al. (5), is that the young might be at lower risk because they have fewer lung tissue cells at risk. Conversely, there is evidence that a given exposure to carcinogens may be more damaging when received at a younger age as demonstrated by chromosome loss at 3p21 (16). These explanations cannot be differentiated by an analysis of the CPS-I data because the inclusion of any two of the three age/duration terms fixes the third.

A single value for dose is unlikely to adequately characterize the lifetime intensity of smoking. Cross-sectional surveys have shown that CPD is not constant; it increases from initiation to ~age 30 and increases more slowly to ~age 50 and then declines (26, 27). This phenomenon would result in CPD reported at older ages underestimating the lifetime smoking exposure compared with CPD reported at midlife. The lower than predicted risk of smoking for earlier ages of initiation thus may be due to the pattern of smoking intensity with age in addition to the fewer lung cells to be exposed as Moolgavkar has suggested. It might also account for the apparent contradiction present for younger ages of initiation between the increased susceptibility to molecular change from carcinogenic effects and the observed lower lung cancer incidence.

Examination of the residuals and χ^2 goodness-of-fit tests suggested that the Moolgavkar model also did not fit the CPS-I data as well as did the extended D-P model. The lack of fit of the three-parameter Moolgavkar model appeared largely due to how the model incorporates age of initiation. The Moolgavkar model's assumption of constant risk after age 20 may be too young of an age for the transition if CPD actually rises to age 30. Our results confirmed the observation of Moolgavkar et al. (5) that the term for net proliferation rate of intermediate cells is not affected by dose. However, our additional observations that only one of the dose-related parameters (c_1, c_2, and b) is estimable with the CPS-I data and that the overall fit is not especially good suggest that the two-stage biological model underlying Moolgavkar's mathematical model may not well describe these data.

There are apparent differences between the British and the American populations similar to the previously observed differences between British and Japanese populations (24). For the Doll and Peto model, estimates of both parameters b and c were smaller for the American population than for the British (the confidence intervals for b did not overlap). The estimates for the three-parameter Moolgavkar model also differed between the two populations.

The differences found in how smoking affects lung cancer risk between the British and the American populations are not unexpected. Neither the British Doctors Study nor the CPS-I were probability samples. The British doctors were enrolled about 8 years before the CPS-I participants were enrolled. There were demographic, environmental, and dietary differences between the two populations. In addition, there were differences in the composition of cigarettes consumed in the two countries (28). Subsequently, there have been changes in the demographics of smokers (29) and additional changes to the composition of cigarettes (28), suggesting that if either study were repeated today, there may well be differences in resulting estimates of model parameters.

We included continuing smokers who reported consuming an excess of 40 CPD; such heavy smokers have been excluded from previous analyses of the British doctors' data. While there were relatively few such heavy smokers in our data, we included them because this did not have a deleterious effect on model fit as assessed by

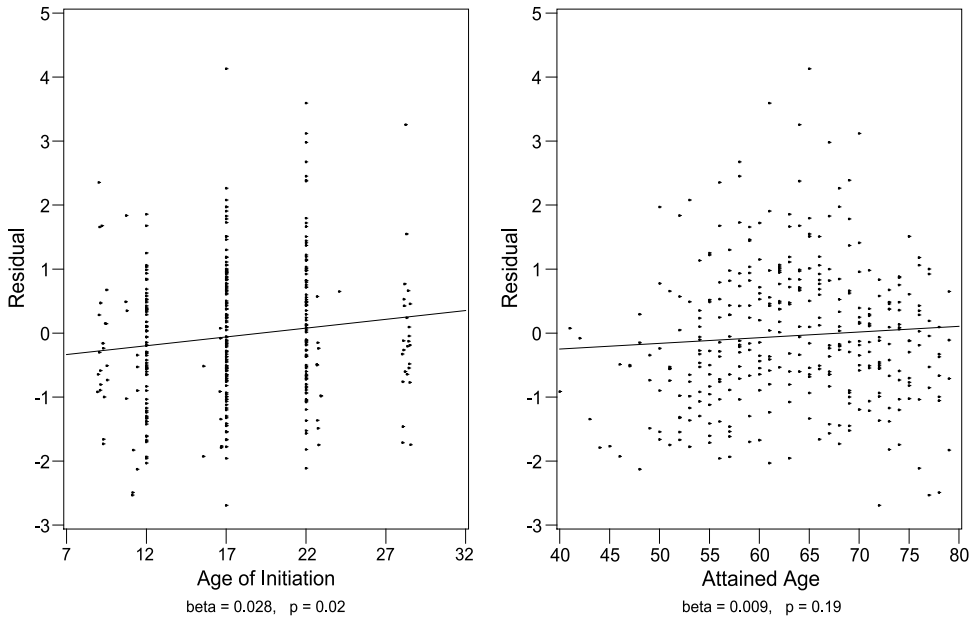

Figure 4. Residuals of the three-parameter Moolgavkar model. *Left,* by age of initiation; *right,* by attained age.

either the residual or the goodness-of-fit analyses. When the 40+ CPD smokers were excluded from Doll and Peto model estimation, the parameter for duration was little changed (3.79 instead of 3.74). The parameter for dose increased slightly (1.20 instead of 0.96); however, its confidence interval (1.05-1.34) still did not overlap the confidence interval for dose parameter with the British data (1.44-2.33).

Previous reports have not agreed on whether nonsmokers should be included when modeling the effect of smoking on the risk of lung cancer. The Doll and Peto and extended D-P models, by definition, do not extend to nonsmokers, while the Moolgavkar model does. Our analyses show that the inclusion of nonsmokers in the Moolgavkar model has only a modest, although statistically significant, effect on the parameter estimates compared with the estimates when only smokers are included. This report, like the cited previous reports, does not attempt to model risk among former smokers. Likely, more complicated models than studied here will be required to effectively model risk among this important and growing population subgroup (30).

In conclusion, our analysis of the CPS-I data shows that adding a third parameter, a measure of where in the life span exposure to smoking occurs, to the Doll and Peto model, in addition to terms for dose and duration of smoking, improves the accuracy of model prediction. This confirms the importance of Moolgavkar's inclusion of two age-related terms but not how age is incorporated in his nonlinear incidence function. Although the underlying biological phenomenology for the importance of two age-related terms is conjectural at this time, the extended D-P model appears useful for future projections of the health consequences of cigarette smoking.

Appendix 1

Cochran (21) suggested minimum expected cell frequencies for goodness-of-fit testing to maximize power and minimize the loss of precision from the asymptotic approximation to the χ^2 statistic. These criteria say that the minimum expected cell frequency should be one, and most cells (80%) should have expected cell frequencies of at least 5. The algorithm for combining cells was chosen to the meet the Cochran criteria for the extended D-P model D and to not bias the results.

The combined cells are detailed in Appendix Table 1. The youngest ages had the sparsest data and all subjects ages 40 to 45 were combined into one cell for each year of age. Those ages 46 to 47 were combined into two cells for each year of age and so forth through those ages 52 to 53 being combined into eight cells. The intermediate ages, 54 to 74, had the most data and were combined into 13 cells for each year of age. Those ages 75 to 79 again had sparser data and were combined into a decreasing number of cells as age increased. In general, those who initiated smoking (init) between ages 15 and 19 and smoked one or more packs of cigarettes per day had the most data. Those who initiated between 10 and 14 or between 20 and 24 had the next most data. For the intermediate ages and ages of initiation, there were several cells for which no combination was necessary for those who smoked one or more packs of cigarettes per day.

The minimum expected cell frequency for all models was at least 1. For model D, 285 cells (80.7%) had expected cell frequencies greater than 5. The Cochran criteria were not quite met for the other models. Only 259 cells (73.4%) of the Doll and Peto model A had expected cell frequencies greater than 5. For the four-parameter Moolgavkar model, 76.3% of the cells had expected cell frequencies of at least 5; for the three-parameter Moolgavkar model, 79.6% of the cells had expected cell frequencies of at least 5.

A.1 Appendix Table 1: Definition of the reduced number of cells for goodness-of-fit testing

Cells	Age	Init	CPD
1-6	40-45	All	All
7,9	46,47	≤19	All
8,10	46,47	≥20	All
11,14	48,49	≤14	All
12,15	48,49	17	All
13,16	48,49	≥20	All
17,23	50,51	≤14	All
18,24	50,51	17	≤19
19,25	50,51	17	20
20,26	50,51	17	30
21,27	50,51	17	45
22,28	50,51	≥20	All
29,37	52,53	7 + 12	7 (All) + 12 (≤20)
30,38	52,53	12	≥21
31,39	52,53	17	≤19
32,40	52,53	17	20
33,41	52,53	17	30
34,42	52,53	17	45
35,43	52,53	≥20	≤20
36,44	52,53	≥20	≥21
45,. . .,305	54,. . .,74	7 + 12	7 (All) + 12 (≤19)
46,. . .,306	54,. . .,74	12	20
47,. . .,307	54,. . .,74	12	30
48,. . .,308	54,. . .,74	12	45
49,. . .,309	54,. . .,74	17	≤19
50,. . .,310	54,. . .,74	17	20
51,. . .,311	54,. . .,74	17	30
52,. . .,312	54,. . .,74	17	45
53,. . .,313	54,. . .,74	22	≤19
54,. . .,314	54,. . .,74	22	20
55,. . .,315	54,. . .,74	22	30
56,. . .,316	54,. . .,74	22	45
57,. . .,317	54,. . .,74	≥25	All
318,327	75,76	7 + 12	7 (All) + 12 (≤20)
319,328	75,76	12	≥21
320,329	75,76	17	≤19
321,330	75,76	17	20
322,331	75,76	17	30
323,332	75,76	17	45
324,333	75,76	22	≤19
325,334	75,76	22	≥20
326,335	75,76	≥25	All
336,343	77,78	≤14	All
337,344	77,78	17	≤19
338,345	77,78	17	20
339,356	77,78	17	≥21
340,347	77,78	22	≤19
341,348	77,78	22	≥20
342,349	77,78	≥25	All
350	79	≤14	All
351	79	17	≤19
352	79	17	≥20
353	79	≥20	All

References

1. U.S. Department of Health and Human Services. Reducing the health consequences of smoking: 25 years of progress. A report of the Surgeon General. Atlanta, Georgia: U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. Department of Health and Human Services Publication No. (CDC) 89-8411; 1989.
2. Doll R, Peto R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. J Epidemiol Commun Health 1978;32:303-13.
3. Artimage P, Doll R. Stochastic models for carcinogenesis. Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 4. Berkeley: University of California Press; 1961. p. 19-38.
4. Doll R, Hill AB. Mortality in relation to smoking: ten years observations of British doctors. BMJ 1964;1:1399-410, 1460-7.
5. Moolgavkar SH, Dewanji A, Luebeck G. Cigarette smoking and lung cancer: reanalysis of the British doctors' data. J Natl Cancer Inst 1989; 81:415-20.
6. Garfinkel L. Selection, follow-up, and analysis in the American Cancer Society prospective studies. Natl Cancer Inst Monogr 1985;67: 49-52.
7. Hammond EC. Smoking in relation to the death rates of one million men and women. Natl Cancer Inst Monogr 1966;19:127-204.
8. Burns DM, Shanks TG, Choi W, Thun MJ, Heath CW, Garfinkel L. The American Cancer Society Cancer Prevention Study I: 12-year follow-up of 1 million men and women. Changes in cigarette-related disease risks and their implication for prevention and control: smoking and tobacco control. Monograph No. 8, U.S. Department of Health and Human Services, NIH, National Cancer Institute. NIH Publication No. 97-4213; 1997.
9. Whittemore AS. Effect of cigarette smoking in epidemiological studies of lung cancer. Stat Med 1988;7:223-38.
10. Garfinkel L, Silverberg E. Lung cancer and smoking trends in the United States over the past 25 years. CA Cancer J Clin 1991;41:137-45.
11. Thun MJ, Day-Lally C, Meyers DG, et al. Trends in tobacco smoking and mortality from cigarette use in Cancer Prevention Studies I (1959 through 1965) and II (1982 through 1988). Changes in cigarette-related disease risks and their implication for prevention and control: smoking and tobacco control. Monograph No. 8, U.S. Department of Health and Human Services, NIH, National Cancer Institute. NIH Publication No. 97-4213; 1997.
12. Leenhouts HP. Radon-induced lung cancer in smokers and non-smokers: risk implications using a two-mutation carcinogenesis model. Radiat Environ Biophys 1999;38:57-71.
13. Peto R. Influence of dose and duration of smoking on lung cancer rates. IARC Sci Publ 1986;23-33.
14. Moolgavkar SH. The multistage theory of carcinogenesis and the age distribution of cancer in man. J Natl Cancer Inst 1978;61:49-52.
15. Moolgavkar SH, Knudson AG Jr. Mutation and cancer: a model for human carcinogenesis. J Natl Cancer Inst 1981;66:1037-52.
16. Hirao T, Nelson HH, Ashok TD, et al. Tobacco smoke-induced DNA damage and an early age of smoking initiation induce chromosome loss at 3p21 in lung cancer. Cancer Res 2001;61:612-5.
17. Anisimov VN. Carcinogenesis and aging. Adv Cancer Res 1983;40: 365-424.
18. Cohen HJ. Biology of aging as related to cancer. Cancer 1994;74: 2092-100.
19. Rao CR. Linear statistical inference and its applications. Wiley series in probability and mathematical statistics. New York: Wiley; 1973.
20. Jennrich RI, Ralston ML. Fitting nonlinear models to data. Annu Rev Biophys Bioeng 1979;8:195-238.
21. Cochran WG. Some methods for strengthening the common chi-squared test. Biometrics 1954;10:417-51.
22. U.S. Environmental Protection Agency. Guidelines for carcinogenic risk assessment. Fed Regist 1986;51:33992-4003.
23. Andersen EL, and The Carcinogenic Assessment Group. Quantitative approaches in use to assess cancer risk. Risk Anal 1983;3:277-95.
24. Mizuno S, Akiba S, Hirayama T. Lung cancer risk comparison among male smokers between the ''six-prefecture cohort'' in Japan and the British physicians' cohort. Jpn J Cancer Res 1989;80:1165-70.
25. Flanders WD, Lally CA, Zhu BP, Henley SJ, Thun MJ. Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: results from Cancer Prevention Study II. Cancer Res 2003;63:6556-62.
26. Shanks TG, Burns DM. The pattern of smoking uptake from survey data. Presented at the Society for Research on Nicotine and Tobacco; 2003 Feb; New Orleans, Louisiana.
27. Centers for Disease Control and Prevention. Cigarette smoking among American Indians and Alaska Natives—behavioral risk factor surveillance system, 1987-1991. MMWR Morb Mort Wkly Rep 1992; 41:861-3.
28. Hoffmann D, Hoffmann I. The changing cigarette, 1950-1995. J Toxicol Environ Health 1997;50:307-64.
29. Burns DM, Lee L, Shen LZ, et al. Cigarette smoking behavior in the United States. Changes in cigarette-related disease risks and their implication for prevention and control: smoking and tobacco control. Monograph No. 8, U.S. Department of Health and Human Services, NIH, National Cancer Institute. NIH Publication No. 97-4213; 1997.
30. Freedman DA, Navidi WC. Ex-smokers and the multistage model for lung cancer. Epidemiology 1990;1:21-9.