Use of a multistage model to predict time trends in smoking induced lung cancer

Joel B Swartz

Abstract

Study objective—The aims were to use a mathematical model to predict the time course of smoking induced lung cancer, and to investigate to what extent the most recent increases in lung cancer mortality are due to cigarette smoking.

Design—A mathematical model was developed and solved by simulation to construct detailed smoking histories of the US white male population given available prevalence data by age and cohort. A multistage carcinogenesis model was used to predict the time course of smoking induced lung cancer given the detailed smoking histories.

Source of data and model parameters—The smoking prevalence figures were taken from work by Harris who calculated them using data collected in the Health Interview Survey. The parameters of the multistage model were taken from Whittemore who fitted the model to several sets of smoking and lung cancer data.

Main results—The smoking model was used to construct detailed smoking histories of the US white male population from 1900 to 1985. In turn the multistage model was used to predict age and cohort specific smoking induced lung cancer mortality rates over this period. These results were compiled to predict the overall age adjusted trend in smoking induced lung cancer from 1970 to 1985. The model predicts a 12% decline in smoking induced lung cancer for this group over the 15 year period.

Conclusions—The model calculations predict a 12% decline in smoking induced lung cancer for this group, during a period when the actual total rate of lung cancer increased by 26%. Taken together with the decline in average tar content in cigarettes over this period, and the relatively constant dose rate among smokers, these results strongly suggest that the recent increase in lung cancer among white males in the USA is due entirely or in large part to factors other than cigarette smoking.

The aim of this study was to use a mathematical model to predict the trend in the smoking induced cancer rate for white males in the USA over a recent 15 year period. This is a matter of substantial public health importance, as there has been a major increase in the lung cancer rate for white males, as well as for all other sectors of the US population during this period. One proposed explanation is that this is the delayed effect of males taking up smoking during the period starting at the beginning of the 20th century and ending in about 1960. A second explanation is that the lung cancer increase is due to other environmental factors, such as exposure to synthetic organic carcinogens, the production and use of which rose exponentially during the postwar period. Proponents of the latter explanation point out that the relative risk of lung cancer declines sharply after a person stops smoking. Risks in smoking prevalence 25 or more years ago, which were followed by steady declines in prevalence, could not therefore be responsible for the current increase in lung cancer rate.

This initial study was confined to white males in the USA because of the relatively large amount and accuracy of appropriate data, and because of the long period of decline in smoking prevalence for this group. The model can easily be extended to other groups in the population. In this paper, lung cancer rate will always mean mortality rate, although the difference between the mortality and incidence rate for lung cancer is small.

BACKGROUND ON CARCINOGENESIS MODELS

The multistage model is the basis of the model used to calculate the age incidence distribution for lung cancer. This model is biologically based, and has been successfully fitted to a range of experimental and epidemiological data. It assumes that a cell makes discrete transitions in the process of changing from normal to cancerous. One or more of these transitions is affected by carcinogens. It predicts that, given constant exposure, the cancer rate increases with age to the \((k-1)\)st power, where \(k\) is the number of transitions. Whittemore has devised expressions for the age-mortality function for intermittent exposure, and has fitted data on lung cancer mortality and smoking to a semi-empirical function derived from the multistage model.

EFFECT OF EXPOSURE CESSION ON CANCER MORTALITY PATTERNS

The multistage model predicts, and many experimental and human studies have demonstrated, that cessation of exposure to a carcinogen can dramatically reduce the excess cancer risk due to that exposure. This phenomenon can occur even late in the life of the exposed subject and after several decades of exposure.

According to the multistage theory of cancer, the later the stage affected by the carcinogen, the greater will be the reduction in relative risk from cessation of exposure late in life. This phenomenon has been shown for such carcinogens as 2-AAF, cigarette smoke condensate, ionizing radiation, arsenic, and DDT.
For our purposes, the crucial point is that cessation of smoking brings about a rapid decline in the relative risk for lung cancer. This has been shown by numerous independent epidemiological studies. Typically, five years after smoking cessation, lung cancer mortality rates decline to about 50% of those for continuing smokers; 15 years after smoking has stopped, the relative risk versus non-smokers drops from the range of 15 to 20, to 2 or 1. These results are summarised in table I.

**Changes in smoking patterns**

In general, smoking prevalence increases from age 20 years to the mid-30s, and then begins to decline, slowly at first but rapidly in the mid-60s. The changes in calendar years are well documented; prevalence increased from about 10% in 1920 to over 55% in the 1950s, followed by a subsequent decline to the current levels of about 35%. An important observation is that calendar year changes occur throughout all age groups, i.e., a general decline in prevalence consists of a decline across all age groups, appearing as a slowed increase in smoking for younger cohorts and a more rapid drop for older cohorts.

Prevalence data from Harris were used. Working from the results of health interview survey questionnaires, Harris reconstructed prevalence data for eight 10 year cohorts, starting in 1880. These data show that the decline in smoking prevalence began about 1960, with some variation among cohorts. Selected prevalences are presented in table II.

**Smoking model and lung cancer mortality function**

The smoking model used is a multistage model in which the first and penultimate stages are affected by smoking. Whittemore has provided a closed form solution for the cancer mortality rate as a function of current age, age starting smoking, and age stopping smoking for this model. The solution contains both a linear and a quadratic term in dose. This closed form solution applies only to the case where the subject smokes for one continuous interval. Whittemore fitted the multistage model to data from the British Physician's study, the Dorn US Veteran's study, and a case-control study of white males in New Mexico. The model provided excellent fits to the data from the first and the third studies, and provided a better fit than traditional models using logistic regression on the New Mexico data. The cancer mortality function for this model is given in equation (1):

\[
M(t) = 2\cdot01 \times 10^{-12} [(t-t_0) + \frac{p}{(1+c/2pc(t-t_1-t_0) + 2pc(t-t_1-t_0))}] \quad (1)
\]

where \( t \) is current age, \( t_0 \) is age started smoking, \( t_1 \) is age stopped smoking, and \( M(t) \) is the mortality rate at age \( t \); \( p \) is a constant = 0.207; \( c \) is the smoking rate in packs per day.

**Other work in this area**

A number of other investigators have attempted to predict population trends in lung cancer given smoking data. We summarise two important ones below.

Stevens and Moolgavkar developed a statistical-phenomenological model to predict lung cancer rates in Great Britain from smoking data. This models lung cancer incidence as a function of age, cohort, and a smoking measure. In analogy with the logistic model, it assumes that the lung cancer risk is proportional to the cumulative cigarette consumption for the given cohort for the given year.

The model makes accurate predictions for time trends in lung cancer, and the values found for the relative risk from smoking are consistent with those found in epidemiological studies. The success of its predictions is diminished by the large number of parameters required.

Brown and Kessler developed an age-period-cohort model to predict lung cancer rates in the USA for 1985 to 2025. In one version the lung cancer incidence is assumed to be proportional to the product of an age, a cohort, and a year parameter. In a second version the period factor is replaced by a factor that depends on total cigarette consumption and on tar content in the cigarettes. Estimates of future cigarette consumption have been used to make projections of lung cancer rates into the 21st century. The model is an improvement over other models in that it takes account of the changing tar content of cigarettes.

**Methods**

**Constructing smoking submodel**

The smoking submodel is needed to calculate smoking history by age and calendar year. This is done by simulation of the changes in smoking practices, and of mortality due to smoking induced lung cancer and other causes. Specifically, we need to know for each cohort at each age the number of “persons” who started smoking and stopped smoking for every possible combination of ages. We assume that the smoking prevalence for a given cohort at a given age is given by the Harris study. Linear interpolation is used to estimate values for years between those for which data are available.

**Categorisation**

All “persons” must be in one of three categories: never smoked, current smoker, former smoker.
"Persons" in the current smoker category are divided into subcategories according to the age they started to smoke. "Persons" in the former smoker category are divided into subcategories according to the age they started smoking and the age they stopped smoking. So $cs(i)$ is the number of "persons" who started smoking at age $i$; $fs(i,j)$ is the number of "persons" who started smoking at age $i$, and stopped smoking at age $j$.

Movement through categories

"Persons" can move from the category non-smokers to the category current smokers, and from current smokers to former smokers. No backward movement is permitted. This rule was used because the solution of the multistage model is much simpler for the case where there is only one continuous smoking period, and because more detailed data on changes in smoking habits are not available. At the end of each year, "persons" are moved forward one category according to the above rules. This rule is relaxed later in a perturbation calculation. In the case of a decline in prevalence, "persons" are moved proportionately from all current smoking subcategories, i.e., if $cs(30)$ is twice as big as $cs(32)$ then twice as many "persons" are taken from $cs(30)$ as from $cs(32)$. This is equivalent to the assumption that the probability of stopping smoking is independent of age starting smoking. The validity of this assumption is confirmed in the work of Novotny et al. 16

For ease of calculation it was assumed that everyone smokes at the same rate, in this case one pack per day. This assumption is relaxed in a perturbation calculation and discussed later. If probability of stopping smoking is roughly independent of dose, then this assumption has no impact on the model results. Data on trends in the proportion of the population smoking at various ages strongly suggest that deviations from this assumption are small. 19

![Figure 1 Carcinogenesis submodel](http://jech.bmj.com/)

Figure 1 Carcinogenesis submodel

### Table III Comparison of actual lung cancer rates with smoking induced rates from model: age adjusted rates for white males

<table>
<thead>
<tr>
<th>Year</th>
<th>Actual rate</th>
<th>Rate from model due to smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>1971</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td>1972</td>
<td>106</td>
<td>100</td>
</tr>
<tr>
<td>1973</td>
<td>107</td>
<td>99</td>
</tr>
<tr>
<td>1974</td>
<td>110</td>
<td>98</td>
</tr>
<tr>
<td>1975</td>
<td>112</td>
<td>97</td>
</tr>
<tr>
<td>1976</td>
<td>114</td>
<td>96</td>
</tr>
<tr>
<td>1977</td>
<td>116</td>
<td>95</td>
</tr>
<tr>
<td>1978</td>
<td>119</td>
<td>94</td>
</tr>
<tr>
<td>1979</td>
<td>120</td>
<td>93</td>
</tr>
<tr>
<td>1980</td>
<td>122</td>
<td>92</td>
</tr>
<tr>
<td>1981</td>
<td>122</td>
<td>92</td>
</tr>
<tr>
<td>1982</td>
<td>124</td>
<td>90</td>
</tr>
<tr>
<td>1983</td>
<td>124</td>
<td>89</td>
</tr>
<tr>
<td>1984</td>
<td>125</td>
<td>89</td>
</tr>
<tr>
<td>1985</td>
<td>126</td>
<td>88</td>
</tr>
</tbody>
</table>

Note: Source for actual rates is reference. 30 Rate is normalised so that the 1970 rate equals 100.

### Prevalence by age and cohort

The smoking prevalence for a particular cohort at a particular age is always fixed to be equal to the prevalence determined from the Harris study. 14 An adjustment is made at the end of each calendar year. If the Harris prevalence exceeds the prevalence at the end of the preceding year, then an appropriate number of "persons" is moved from the non-smoking category to the smoking category. If the prevalence needs to be adjusted down, then "persons" are moved from the current smoking category to the former smoking category.

### CARCINOGENESIS SUBMODEL

The carcinogenesis submodel is expressed in equation (1). It is used to calculate the lung cancer mortality rate at a particular age given the age for starting smoking and the age for stopping smoking. This is the exact solution for the multistage model with the first and penultimate stage affected by smoking. The function is semi-empirical because of the exponent of 4.5. Strictly speaking, under the multistage model the exponent should be an integer. However, the exponent of 4.5 gives the best fit. This submodel is shown diagrammatically in fig 1.

### MODEL FLOW DESCRIPTION

1. A computation is performed for each cohort, starting with those born in 1900 and finishing with those born in 1945. Each cohort contains all "persons" born in one calendar year. (2) For each cohort the computations begin at age 21 years. The ratio of $ns$ to $cs(21)$ is set to be the same as that of $(1-prevalence)$ to prevalence. (3) At the end of each year, $j$, mortality is calculated for all subcategories of non-smokers, current smokers, and former smokers using equation (1). (4) The $ns$, $cs$, and $fs$ are adjusted so that the sum of $cs(i)$ divided by the total population equals the prevalence. If the prevalence declines then the appropriate number of "persons" is moved from $ns$ to $cs(i)$. If the prevalence increases then the appropriate number of "persons" is moved from each $cs(i)$ to the appropriate $fs(i,j)$. The number taken from each subcategory of $cs$ is proportional to the size of the subcategory, i.e., the same proportion is taken from each subcategory.

After all the simulations and calculations had been performed, the age adjusted lung cancer mortality rates were calculated for the years 1970 to 1985, with appropriate contributions from each cohort. For example, the rate for 1970 was computed using the rate for 69 year olds born in 1901, 68 year olds born in 1902, etc. The contributions from the different age groups were weighted so that the final lung cancer mortality rate would be age adjusted to the 1970 US population. The mortality rate is based only on the age range 42 to 70 years which includes the bulk of the lung cancer mortality. This range is restricted so that the same number and spectra of ages contribute to the mortality rate for each of the years 1970 to 1985.

### Results

#### MAJOR RESULTS

The major results for the study are shown in table III and in fig 2. A comparison is given between
the actual age adjusted lung cancer mortality rate for white men over the period 1970 to 1985 and the rates for smoking induced lung cancer predicted by the model. All rates are given relative to the rate in 1970. While the actual lung cancer rates climbed by 26%, the smoking induced rate declined by 12%.

<table>
<thead>
<tr>
<th>Birth year</th>
<th>Probability of dying from lung cancer at age:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1905</td>
<td>1.8 x 10^{-4} 6.4 x 10^{-4} 1.8 x 10^{-3}</td>
</tr>
<tr>
<td>1910</td>
<td>1.8             6.5             1.8</td>
</tr>
<tr>
<td>1920</td>
<td>1.9             6.7             1.8</td>
</tr>
<tr>
<td>1925</td>
<td>1.8             6.2</td>
</tr>
<tr>
<td>1930</td>
<td>1.7             6.0</td>
</tr>
<tr>
<td>1935</td>
<td>1.6</td>
</tr>
<tr>
<td>1940</td>
<td>-</td>
</tr>
</tbody>
</table>

Table IV gives more detail about the process. It shows how the probability of dying from lung cancer at specific ages changes with calendar year. One can see a peak in these probabilities and then a gradual decline.

**Sensitivity Analysis**

Some model assumptions were relaxed or varied to investigate the robustness of the overall predictions of the model.

**Drift**

In the smoking submodel, only one movement through smoking categories is permitted, i.e., either “persons” are moved from never smoked to current smoker, or from current smoker to former smoker, depending on the direction of change in the prevalence. Since it is possible for some people to start smoking in a year when prevalence decreases, and to stop smoking in a year when smoking decreases, a small drift term was added to the smoking submodel. In each year we moved an additional number of “persons” equal to 0.5% of the current smokers from never smoked to current smoker, and from current smoker to former smoker. The drop off between 1970 and 1985 remained at 12%, although the absolute mortality rate increased by 9% uniformly.

**Smoking dose**

In the main model a uniform smoking rate for all “persons”, one pack per day, was used. The cancer mortality function contains both linear and quadratic dose dependence. A run was performed with the dose doubled to two packs per day. The proportional decline was the same as in the main model, 12% over the 15 years. The absolute rate increased, by 13%.

**Age smoking started**

In the main model it was assumed that the earliest possible age for starting smoking was 21 years. A run in which the earliest age for starting smoking was 18 years was performed. This is the starting age used by Whittemore. The trends were the same as in the main model. There was a 12% decline in lung cancer rate over the 15 year period. The absolute rate was 12% higher than when smoking started at age 21 years.

**Other carcinogenesis-lung cancer functions**

 Runs were performed with two other functions that describe lung cancer mortality as a function of smoking history. The first was Whittemore’s $g$, which depends on smoking only by the cumulative number of packs smoked. The function is given in equation (2):

$$ M(t) = c(t - 5)^{4 + 5} (1 + a \text{ pk}) $$

where $t$ is age; pk is cumulative packs smoked; $M$ is mortality; and $c$, $a$ are constants.

Whittemore showed that this function adequately fitted smoking studies, although not as well as the multistate function. The run with this function produced a decline of 5% over the 15 year period, about half that found using the basic model with the multistate function.

Finally a multistage model was used, in which it was assumed that smoking affected only the fourth of five stages. We chose this model because it had been used in an earlier version of this study. The equations for current and former smokers are presented below as equations (3) and (4).

The run with the one stage affected model produced a decline of 8% over the 15 year period, about two thirds of that with the two stages affected model.

$$ M(t) = M_0(t) + [1 + b(t - 10)^{4 + 5} t < t_1] $$

$$ M(t) = M_0(t) + [1 + b(t - 10)^{4 + 5} t > t_1] $$

where $t$ is current age; $t_0$ is age start smoking; $t_1$ is age stop smoking; $M_0$ is background mortality; $M$ is mortality; and $b$ is a constant.

**Discussion**

The combined model with a carcinogenesis and a smoking component was used to calculate the trends in the smoking induced lung cancer rate among white men in the USA for the period 1970 to 1985. The model predicts a 12% decline in smoking induced cancer over that period, during which the overall lung cancer rate increased by 26%. To calculate the lung cancer mortality rate (hazard function) a form of the multistate carcinogenesis model was used which had been verified by several authors and for which the parameters were established by Whittemore.

This approach should provide a more accurate prediction of lung cancer trends than those used.

![Figure 2](http://jech.bmj.com/content/46/3/311)

*Figure 2* Comparison of lung cancer trend and trend in smoking induced lung cancer predicted by model.
Predicting time trends in smoking induced lung cancer

by other investigators for several reasons. First the multistage model was used to model the carcinogenesis process. This model has biological plausibility while the statistical models are purely phenomenological. The multistage model has been shown to predict accurately the hazard function in carcinogenesis for a wide range of animal and human studies.8

By use of the multistage model, it was possible to take into account specific features of the carcinogenesis process not included in other models. In other models generally only cumulative cigarette consumption is considered. According to the multistage model, the cumulative consumption figure may be inaccurate as a predictor for an individual because age at exposure and pattern of exposure are very important. For example, smoking at the age when the penultimate transition is likely to occur is far more important than smoking at age 35 years.

This model is also an improvement over other models because it makes use of the fine structure of smoking within the population. In other models only aggregate smoking data are used for the entire cohort. In this model the smoking history of different sectors of the population is used. This is necessary under the multistage model because incidence depends greatly on pattern of smoking. For example, under other models, four people smoking for 10 years would be equivalent to one person smoking for over 40 years. Under the multistage model, the latter would incur a much higher total risk as incidence depends on period of exposure to the fourth or fifth power.

We have investigated the change in predictions of the model under changes or relaxations in the model assumptions. In all cases these produced no change in the prediction of a decline in smoking induced lung cancer over the period.

In order to justify the prediction of trends, it is necessary either to incorporate dose trends in this model or to show that they do not affect the conclusion. There are limited data on trends in smoking dose, but those available show that there is little change, and that this change is too small to affect the results. Data from the Surgeon General's Report indicate the proportion of current smokers in each of the seven dose categories for 1974 and 1985.13 For males, while there is some shuffling between categories, there is scant overall trend. The average number of cigarettes smoked per day increased by 2%, between 1974 and 1985. It is reasonable to assume that the same situation prevailed back to 1960, when the decline in prevalence began. Thus the decline in prevalence was accompanied by approximately constant dose among those who smoked. Therefore the calculations based on prevalence should be largely unaffected by changes in dose.

The other major factor which could not be included in this model is the reduction in tar content in cigarettes over the past three decades. According to Stellman and Garfinkel,28 the lung cancer risk from smoking is approximately proportional to the tar content of the cigarette. Moreover, they report that the average tar content declined by over 50%, between 1950 and 1980. Allowing for lag factors, we estimate that the decline in tar content should have caused an additional decline in lung cancer mortality over the period of study of from 5% to 15%. It is possible that the increase in dose over that period could have increased lung cancer mortality by up to 5%. Therefore the inclusion of the factors missing from the model would have caused no change in the prediction, or produced a prediction of a steeper decline in lung cancer.

Finally, Whittemore7 observed that her model overpredicted the mortality for former smokers. Therefore any error incurred from using this carcinogenesis function is also on the side of predicting a smaller than actual decline in lung cancer.

In conclusion, taking into account the assumptions and omissions of the model, these calculations make a strong case that the recent rise in lung cancer mortality among white males in the USA is not due to cigarette smoking.

I should like to thank the following for their helpful comments and suggestions: Dr Samuel Epstein, Ms Christine Riddough, Dr Lael Gatwood, Dr Michael Altman, and Dr Edward Lichter. I should like to thank the following for their programming and statistical contributions: Mr Robert Glass, Ms Christine Riddough, Mr Badar Iqbal.

I should also like to thank Ms Christine Smith for typing the manuscript at various stages. I should like specially to thank the Micropopulation Simulation Resource at the University of Minnesota and the Preventive Medicine Department at the University of Illinois Medical Center for their hospitality while I worked on this paper.


References