Published in final edited form as:

Risk Anal. 2012 July ; 32(0 1): S151–S165. doi:10.1111/j.1539-6924.2011.01754.x.

Yale Lung Cancer Model

Theodore R. Holford¹, Keita Ebisu², Lisa McKay¹, Cheongeun Oh³, and Tongzhang Zheng¹

¹Yale School of Public Health, 60 College Street, New Haven, Connecticut 06520, USA

²Yale School of Forestry and Environmental Studies, 195 Prospect Street, New Haven, Connecticut 06511, USA

³New York University Langone Medical Center, 650 First Avenue, New York, New York 10016, USA

Abstract

The age-period-cohort model is known to provide an excellent description of the temporal trends in lung cancer incidence and mortality. This analytic approach is extended to include the contribution of carcinogenesis models for smoking. Usefulness of this strategy is that it offers a way to temporally calibrate a model that is fitted to population data and it can be readily adopted for the consideration of many different models. In addition, it provides diagnostics that can suggest temporal limitations of a particular carcinogenesis model in describing population rates. Alternative carcinogenesis models can be embedded within this framework. The two stage clonal expansion model is implemented here. The model was used to estimate the impact of tobacco control following dissemination of knowledge of the harmful effects of cigarette smoking by comparing the observed number of lung cancer deaths to those expected if there had been no control compared to an ideal of complete control in 1965. Results indicate that 35.2% and 26.5% of lung cancer deaths that could have been avoided actually were for males and females, respectively.

Keywords

Age-period-cohort calibration; lung cancer; cigarette smoking; population risk; two-stage clonal expansion model

1. INTRODUCTION

Trends in lung cancer incidence rates have been well described by age-period-cohort (APC) models ⁽¹⁻³⁾, which take into account three temporal factors: age (*a*) at diagnosis, period (*p*) or date of diagnosis, and cohort (*c*) which represents generational effects. As a method of analysis for rates in descriptive epidemiology, the APC model can provide valuable clues that are useful to explore in analytical studies. However, it does not quantify the effect of population exposure to risk factors on the vital rates, and the interpretation is somewhat heuristic.

For lung cancer, cigarette smoking is thought to be the predominant cause of disease⁽⁴⁻⁶⁾ and APC models indicate that cohort effects predominate over period in describing these trends. These are consistent with the concept of smoking initiation generally taking place among individuals in their late teens or early twenties, resulting in generational trends or cohort effects that would result from effective promotion by tobacco manufacturers. In the US, large changes in cigarette smoking resulted from free distribution of cigarettes to military recruits in World War II and advertising campaigns directed at women seeking equality in gender rights in the 1960s and 1970s. While period effects were generally found to be much

smaller, they were not entirely absent, and they could have resulted in part from antismoking campaigns or changes in the manufacturing of cigarettes. These inferences provide a qualitative rationale for observed trends, but they do not make use of the vast literature of analytical studies that have quantified the association between cigarette smoking and lung cancer mortality risk.

The rationale for the Yale Lung Cancer Model is to provide a framework for analyzing the extent to which results obtained from analytical studies, and population data on exposure to cigarette smoking can account for effects of age, period and cohort on lung cancer mortality. An accurate measurement of exposure to cigarettes and an accurate model for the effect of exposure should account for observed temporal trends in rates. Limitations in our ability to account for these temporal effects can arise from an inaccurate model, inadequate smoking exposure data, or changes in exposure to another cause of this disease. The model seeks to characterize unexplained temporal trends and to use the unexplained model parameters to calibrate the results in order to improve agreement with observed data for a population.

This macro scale model is fitted to population rates by considering summary exposure information for subgroups, which is distinct from the micro simulation models that simulate the experience of individuals that are then combined to represent a simulated population. The statistical approach involves the fitting of a model to observed data in the population, and then using the fitted model parameters to obtain estimated or predicted values for alternative distributions of exposure. Analytical epidemiology studies provide estimates of carcinogenesis model parameters that characterize the effects of exposure to cigarette smoking, which may be broken down by age of initiation, length of exposure, and duration of cessation. By combining mortality rate estimates for subgroups using estimates of the relative frequency for the subgroup we obtain predicted mortality rates for larger subgroups (e.g., combining over length of exposure to obtain an overall rate for smokers) or the overall population (e.g., combining rates for never, current or former smokers). These overall rates are then calibrated so that they yield values that (a) correspond to the overall population rates, and (b) correct for temporal effects of age, period and cohort that are not well described by the carcinogenesis model.

This analysis can provide estimates of lung cancer mortality rates and number of lung cancer deaths under alternative smoking histories using maximum likelihood estimates of a multiplicative calibration factor. The actual tobacco control (ATC) that occurred in the US will be compared to hypothetical alternatives that might have occurred. The alternative scenarios considered are no tobacco control (NTC) which could have resulted from ignoring scientific evidence on health effects of cigarette smoking and continuing behavior that existed earlier. In addition, we consider an idealized scenario in which complete tobacco control (CTC) resulted in cessation of smoking following publication of the Surgeon's General's Report⁽⁴⁾ in 1964. The two stage clonal expansion (TSCE) model for cancer is the primary focus of this work, but the model can be easily modified to consider alternative carcinogenesis models for the effects of smoking on lung cancer risk, or models that use parameters derived from different study populations.

2. METHODS

The Yale Lung Cancer Model describes the impact of a distribution of exposure history for cigarette smoking in a population of individuals in a particular age-group at a period of time. It makes use of (a) a quantitative description of the relationship between smoking history and the lung cancer mortality rate, (b) the distribution of smoking history summaries, and (c) a calibration that aligns observed population rates with those from equations in (a). Let $\mathbf{Z}$ represent a summary of smoking history, and $\lambda(\mathbf{Z})$ the mortality rate in a population

resulting from the specified distribution of exposure. Calibration of the rate is accomplished by introducing a multiplicative factor that may either be a constant to be estimated, or a function of parameters to be estimated that can depend on times from critical reference points, $\mathbf{t}$, giving rise to an estimated rate for the population,

$$\lambda^* = \theta(\mathbf{t}) \lambda(\mathbf{Z})$$

where $\theta(\mathbf{t})$ represents the calibration factor. In this section we describe the smoking and the calibration models that were used to obtain these estimates.

2.1. Approach/Model

The TSCE model was used in this work with parameters estimated using data from the Health Professionals Follow-up Study (HPFS) for males and the Nurses' Health Study (NHS) for females. Moolgavkar *et al*⁽⁷⁻¹¹⁾ proposed the TSCE model in which the carcinogenesis process is initiated in a cell that then multiplies to form a clone and further detail on the TSCE model is provided in Chapter 8 of this monograph.⁽¹²⁾ A second hit on one of these initiated cells transform it into a cancer cell that subsequently multiplies further until it forms a tissue mass that can be clinically identified as cancer. The functional form for the TSCE model is complex, but it has been found to provide an excellent description of the effect of age on lung cancer incidence and mortality.

To model the effect of smoking on lung cancer mortality rates, we regard the population as a mixture of never, current and former smokers, each with prevalence p_0 , p_1 and p_2 respectively, giving the overall rate

$$\lambda = p_0 \lambda_0(a) + p_1 \lambda_1(a, \bar{d}_1, \bar{a}_{i,1}) + p_2 \lambda_2(a, \bar{d}_2, \bar{a}_{i,2}, \bar{a}_Q) \quad (1)$$

where $\lambda_0(\cdot)$, $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are the rates for the corresponding smoking categories. Other parameters in the model are age (a), mean number of cigarettes smoked per day ($\bar{d}_i$) for current ($i = 1$) and former smokers ($i = 2$), mean age of smoking initiation ($\bar{a}_i$), and mean age quit ($\bar{a}_Q$).

Among those who never smoked, the mortality rate $\lambda_0(a)$, is a function of age alone which reflects the underlying effect of the aging process on lung cancer risk.

The mortality rate among current smokers, depends not only on age, but on dose or the number of cigarettes smoked per day, d , and the age of initiation, a_{i1} . To apply the TSCE model to relatively homogeneous subgroups of the population, we used average values for dose and age of initiation, but greater accuracy can be achieved by further subdividing the broad class of current smokers by age at initiation.

For former smokers, mortality is not only dependent on current and initiation ages and dose, but also on age of quitting, a_Q , yielding $\lambda_2(a, d, a_{i2}, a_Q)$. Implicit in these covariates is time quit, $a - a_Q$. Because time quit is highly variable as a cohort gets older, risk estimated from a carcinogenesis model is likely to also vary greatly within the population of former smokers. In order to improve accuracy, the former smoker category was broken down by years quit as follows:

1. 1-2 years
2. 3-5 years

3. 6-10 years
4. 11-15 years
5. 16 years or more.

In duration category j , mean dose ($\bar{d}_{2j}$), mean age of initiation ($\bar{a}_{2l,j}$), mean age quit ($\bar{a}_{q,j}$) and proportion of former smokers in the category (q_{2j}) were determined, yielding the mortality rate among former smokers

$$\bar{\lambda}_2 = \sum_j q_{2j} \lambda_{2j} \left(a, \bar{d}_{2j}, \bar{a}_{2l,j}, \bar{a}_{q,j} \right)$$

2.2. Data

Data used to calibrate and validate the model were the number of lung cancer deaths, and the US population reported for single year categories. SEER*STAT provides these data and documentation of methodology used to obtain the estimates are provided on the SEER website (<http://seer.cancer.gov/>).

Estimates of rates of smoking initiation and cessation, as well as dose were obtained for five year birth cohort categories using data from National Health Interview surveys, and further detail on how these were derived are given in Chapter 2.⁽¹³⁾ These values provided estimates of the actual experience in the US for birth cohorts born from 1900 onward, and they are used to generate smoking exposure for the actual tobacco control (ATC) scenario. For the no tobacco control (NTC) scenario, the rates and dose were assumed to remain the same after 1955 when knowledge about the harmful effects of smoking cigarettes began to be disseminated, as they were just before that date. Before 1955 the same rates as those used for the ATC case were used. Finally, the complete tobacco control (CTC) case assumed a hypothetical ideal in which smoking initiation ceased with publication of the Surgeon General's report in 1964 and all current smokers quit. To determine inputs for the Yale Lung Cancer Model, summary estimates of the distribution of smoking histories for the US population were calculated by running the smoking history generator many times using the smoking initiation and cessation rates and doses under the alternative scenarios, reporting summary statistics for the parameters of interest in relevant subgroups. Further details on the smoking history generator (SHG) and the manner in which population smoking histories were generated are described in Chapter 5,⁽¹⁴⁾ and Chapter 4⁽¹⁵⁾ provides a discussion of the approach used to develop the counterfactual tobacco control scenarios.

2.3 Calibration and validation

The purpose of the Yale Lung Cancer Model is to provide a quantitative description of lung cancer mortality trends in the US as a function of available data on cigarette smoking. Estimating parameters for a carcinogenesis model requires the use of data from a cohort study in which subjects are followed over time, e.g., HPFS and NHS used here. Not only do these groups differ socio-economically from the overall population, but their knowledge about factors affecting health risk is likely to be comparatively high. Using a carcinogenesis model derived from these populations may not yield results that agree well with population rates because of bias in model parameters with respect to the US population, an inconsistency that would be expected when one population is not well represented by another. Another potential source of bias is the exposure estimates, which are derived from surveys conducted in the 1970s and later. Smoking behaviors vary widely and these have changed considerably over time. However, surveys necessarily simplify what can be a complex smoking history that often relies on a subject's memory. All of these limitations

can result in biased estimates of lung cancer mortality rates when applied to the entire population. Calibration provides a correction for these discrepancies that result from direct use of a model that may be imperfect for any or all of these reasons. In addition, these limitations may not only result in systematic differences in scale, but in temporal differences, as well.

An APC model was employed to calibrate the carcinogenesis model in order to bring rates into conformity with rates for the overall population. Let $\mathbf{t} = (a, p, c)$ represent a vector of temporal elements: age, period and cohort, respectively. Details on exposure to cigarette smoking in the population at a particular time $(a, p$ and $c=p-a)$ is given by the vector $\mathbf{Z}(\mathbf{t})$. A carcinogenesis model provides an estimate of the mortality rate as a function of the population smoking exposure data, $g=I\{\mathbf{Z}(\mathbf{t})\}$. We calibrate estimates from a carcinogenesis model using a multiplicative factor that depends on the temporal vector,

$$\lambda^* \{\mathbf{Z}(\mathbf{t}); \mathbf{t}\} = \theta(\mathbf{t}) \lambda \{\mathbf{Z}(\mathbf{t})\}, \quad (2)$$

which is a log-linear function of the temporal elements, similar to the approach employed by Luebeck *et al*⁽¹⁰⁾,

$$\theta(a, p, c) = \exp \left\{ \mu + \alpha_a + \pi_p + \gamma_c \right\} \quad (3)$$

The intercept, μ , scales the rates so that the estimates from the model correspond overall with those observed in the US population. Temporal elements for age (α_a , $a=1, \dots, A$), period (π_p , $p=1, \dots, P$) and cohort (γ_c , $c=1, \dots, C$) provide corresponding calibration for temporal elements in the carcinogenesis model that do not correspond well to the effects observed in the population as a whole. If temporal effects are all 0, then the model is in good temporal agreement with the population, and the extent to which these effects become parallel to the abscissa indicates the adequacy of the carcinogenesis model's characterization of the corresponding temporal trend in the population rates. Poor agreement could result from either a limitation in the carcinogenesis model or limitations in the population estimates of exposure.

The well recognized identifiability problem in APC models affects the calibration function parameters, and the phenomenon has been discussed in considerable detail⁽¹⁶⁻²⁰⁾. In this form, log θ resembles an analysis of variance model, and the usual constraints imply that

$$\sum_a \alpha_a = \sum_p \pi_p = \sum_c \gamma_c = 0$$

but the linear dependence among age, period, and cohort extends to indices for the three time effects, in that $c = p - a + A$. Hence, the design matrix for a linear model that includes all three factors is not of full rank, and a unique set of parameters for a corresponding generalized linear model does not exist.^(16, 17) While not offering a solution to the identifiability problem, it is possible to develop ways of understanding the source of the difficulty so that one can express estimable components that are easily interpreted. This can be accomplished by partitioning each temporal effect into overall slope or direction of the trend and curvature or deviation from linear trend.^(21, 17) For example, we can represent the age effects by

$$\alpha_a = \left(a - \frac{A+1}{2} \right) \beta_a + \alpha_{ca} \quad (4)$$

where β_a is the underlying slope for age, and α_{Cj} the curvature or departure from linear trend. It has been shown, using similar partitions for period and cohort, that curvature terms (α_{Cj} , π_{Cj} and γ_{Ck}) are identifiable, but slopes (β_a , β_π and β_γ) are not.^(21, 17) In effect, the slopes are aliased by an indeterminate constant, ν that is hopelessly entangled with all three effects, so that any particular set of slope estimates (indicated by asterisks) is associated with a true slope by

$$\begin{aligned}\beta_a^* &= \beta_a + \nu \\ \beta_\pi^* &= \beta_\pi - \nu \\ \beta_\gamma^* &= \beta_\gamma + \nu\end{aligned}\quad (5)$$

From the data alone, there is no way to estimate ν , but some linear combination of the slopes can be estimated, e.g., drift which is defined by $(\beta_\pi + \beta_\gamma)$.^(22, 23) It is also well known that fitted values are an estimable function of the parameters; hence, the identifiability problem only affects individual temporal parameters and *not* the calibration factor.

Calibration requires fitting the APC model for $\theta(\cdot)$ to a function of the observed rates, and thus obtaining optimal estimates of the temporal parameters. We assume that the number of lung cancer deaths, Y , has a Poisson distribution, and the denominator for the rate, D , is known. The observed calibration factor, $\hat{\theta} = Y/D\lambda$, is the maximum likelihood estimate for the group, and the variance of the estimate would be $Var(\hat{\theta}) = \theta/D\lambda$. If we also assume a log-linear model for the calibration factor, then maximum likelihood estimates of the parameters can be obtained by fitting a generalized linear model in which the linear predictor, η , is related to the calibrated rate, λ^* through the link function

$$\eta = \log \theta = \log(\lambda^* / \lambda) = \mu + \alpha_a + \pi_p + \gamma_c$$

We specify a Poisson distribution for the response (i.e., the observed calibration factor) and introduce a scale weight equal to the denominator for the factor, $D\lambda$.⁽²⁴⁻²⁶⁾ Estimates of the model parameters were obtained using PROC GENMOD in SAS®. Estimates of a calibrated rate given a particular set of smoking exposure covariates, $\mathbf{Z}$, employs both the estimated rate from the carcinogenesis model and the corresponding maximum likelihood estimate of the calibration factor for the given age, period and cohort, $\hat{\theta}(a, p, c) \lambda(\mathbf{Z})$.

The likelihood ratio goodness of fit statistic provides an overall summary of fit. The APC model without any smoking contribution is known to provide a good description of temporal trends,^(27, 1-3) so it should not be surprising when calibrating for all three temporal factors that one obtains good agreement between fitted and observed rates. Dropping one temporal factor from the calibration demonstrates how well it is characterized by the model. For example, if one dropped age and only calibrated for period, cohort and a constant, comparing observed and fitted rates, then systematic departure would suggest that age is not well characterized by the model.

Estimates of age, period and cohort calibration parameters provide model validation by indicating elements of trends that are not well described by a particular carcinogenesis model. When a plot of these parameter estimates is overlaid onto a similar plot of parameters from the APC model without an embedded carcinogenesis model, one can see how much of the trend has been explained. If the carcinogenesis model completely explained temporal effects estimated in the APC model then the parameters from the calibration should be zero. Intervals in which the temporal effects are not constant, on the

other hand, point to epochs in which the carcinogenesis model is not providing a good characterization of trend.

3. RESULTS

Estimates of the number of age-specific lung cancer mortality rates and annual lung cancer deaths were determined using the TSCE models with parameters estimated from HPFS and NHS for males and females respectively. Calibration methods included temporal adjustment for age, period and cohort.

3.1. Hypotheticals

Figure 1(a) shows age-specific mortality rates derived from the TSCE models with parameters estimated from HPFS males. The hypothetical groups considered differ in their smoking histories, i.e., age started, whether they quit smoking, and the number of cigarettes smoked per day. The rates for nonsmokers are considerably lower than the smokers and they increase with age. Two ages at smoking initiation of 20 cigarettes per day were considered, 14 and 25, and the TSCE model implies a large difference in risk for individuals who begin smoking at an early age. Both age initiation groups were divided into hypothetical groups who continued to smoke and who quit at age 35. A clear advantage becomes quickly apparent indicating the benefit in reducing lung cancer mortality risk by quitting. However, risk is still substantially higher than that of nonsmokers and this difference shows no sign of abating up to age 84 (not shown) which is the age range considered in this work. The implication of the TSCE model is that one can never recover completely the harm done by cigarette smoking. Finally, doses of 10, 20 and 40 cigarettes per day were considered for those who begin smoking at 25 and quit at 35. The model implies a clear dose response relationship for lung cancer mortality risk, although the magnitude of that effect over this range is not as great as the effect of age at initiation and cessation. Figure 1(b) shows the corresponding scenarios for women using data from NHS, and the temporal patterns are quite similar to those observed for men.

3.2. Calibration and Validation

An overall summary of the calibration results determined by estimating the age, period and cohort parameters are shown in Table I. The scaled deviance test for goodness of fit of the TSCE model was 1,830.0 for males and 1,554.8 for females, which would be compared to chi-square on 1,272 df, which strongly indicate a lack of fit. A comparison of the fitted rates from the Yale Lung Cancer Model with the observed rates, some of which are shown in Figures 2(a,b), suggests that the lack of fit is random, thus the significance was regarded as extra-Poisson variation and the corresponding scale parameter was estimated to be 1.44 (i.e., the variance about the fitted rates was 44% greater than expected from a Poisson distribution) for males and 1.22 for females. Because the overall linear trends are not estimable, the summary for the individual components of trend only test for curvature, which is estimable, and to accomplish this an F-test was used where the scale estimate (Pearson chi-square divided by its df) was used as the denominator. In each of the three aspects of temporal trend, curvature is highly significant ($P < .0001$) which indicates that there are aspects of temporal trend that are not completely characterized by the TSCE model. This suggests that calibration for these temporal effects may be important for improving the estimated number of cancer cases.

To address the question of how much the TSCE model with the available smoking data for the population can explain temporal trends, we compared the temporal effects with the carcinogenesis model included with those without a carcinogenesis model. A summary of the impact of temporal effects with and without the carcinogenesis model are shown in

Table I. In each case, the F-tests for temporal effects are considerably smaller using the TSCE model, and column four shows the percent curvature explained by the carcinogenesis model for each temporal effect. The model accounts for 90% of the age curvature, and 51% and 68% of period and cohort curvature, respectively, for men. Similarly for females, the amount of curvature explained by the TSCE model is 74% for age, 68% for period and 75% for cohort. Figure 3 (a) shows the estimated age effects for men and women using the TSCE model and the model with no carcinogenesis contribution included, using the constraint of zero slope for period. If the model offered a perfect description of temporal trend, effects would be zero, and trends parallel to the x-axis indicate no temporal effect. For ages over 50, the model provides a fairly good summary of age trends for females, although the declining trend shows the need for a correction that decreases for the older age groups, i.e., the model tends to overestimate the rates compared to younger ages. The decline is greater for males. For the youngest ages, the relative correction is greater than for the oldest ages, but the effect on the overall rates is much less because lung cancer rates are low in this group. Period effects, shown in Figure 3 (b) employ the same scale as the other temporal effects to allow comparison of magnitude, and these are constrained to have zero slopes to achieve a unique set of estimates. A clear pattern is apparent, and the effects without the carcinogenesis model have greater curvature than the effects with the model. However, period required much small calibration than either age or cohort. Finally, the estimated cohort effects using the constraint for period are shown in Figure 3(c). It is important to recognize that the estimates for the most recent cohorts are determined from as few as a single rate in the youngest age groups, resulting in considerably less precision. Thus, the large fluctuations to the right of the curves in this graph are likely to be random. It is also apparent that the TSCE model that includes smoking history data has explained much of the existing cohort trend but not all of it, especially for early cohorts.

This calibration function was applied to the specific hypothetical smoking groups. The effect was to modify not only the overall level for the rates, but it corrected aspects of trend that was not appropriately accounted for in the carcinogenesis model. Figures 4(a) and (b) show APC calibrated trends for hypothetical smoking histories in the 1921 birth cohort for males and females respectively. While the overall patterns are similar to those seen in Figure 1, the proportionate increase in the calibrated trends tends to be somewhat greater. While the patterns for females are similar to those seen for males, the overall levels are somewhat lower.

3.3. Tobacco Control Scenarios

Trend in the number of lung cancer deaths for actual tobacco control (ATC) using the TSCE model is shown in Figures 5 (a) for males and (b) for females. Table II gives the estimated number of lung cancer deaths by gender, which is identical to the observed. Yearly trends in the estimates that include a constant calibration parameter are similar to but slightly different from those that are APC calibrated. For both males and females the overall increase is less pronounced when temporal trends for age, period and cohort are not calibrated. The total number of cases, on the other hand, is essentially the same which is induced as a result of the normal equations solved when finding maximum likelihood estimates in Poisson regression.

Yearly trend estimates assuming no tobacco control (NTC) using the TSCE model are displayed in Figures 5(a,b) for males and females respectively. The temporal trends in later years are somewhat steeper when temporal calibration is not invoked. For both genders, the temporal calibrated estimates show fewer lung cancer deaths in the earlier years compared to those obtained using constant calibration and the reverse for more recent years. The estimated number of lung cancer deaths that would have occurred had there been no tobacco

control was 2.67M for males and 1.27M for females. Thus, the estimated number of lung cancer deaths avoided by tobacco control was 0.82M (0.60M males and 0.22M females).

Estimates of annual number of lung cancer deaths if complete tobacco control (CTC) had been achieved following publication of the Surgeon General's Report are shown in Figure 5. The impacts of calibration are similar to those noted for the no tobacco control case. Estimated numbers of lung cancer deaths that would have occurred under this ideal scenario are 0.96M and 0.44M for males and females respectively. Overall, an estimated 2.54M lung cancer deaths (1.71M males and 0.83M females) could have been avoided under this ideal circumstance. This suggests that the controls that were implemented avoided about 35.2% of the potential for males and 26.5% for females.

4. DISCUSSION

The Yale Lung Cancer Model with TSCE as the embedded carcinogenesis model found that tobacco control that was implemented in the US reduced lung cancer deaths from an estimated 2.67M to 2.07M (21%) in males and 1.27M to 1.05M (17%) in females. This is about 0.82M lives saved. Under idealized circumstances in which complete tobacco control was implemented, lung cancer deaths could have been reduced still further to 0.96M and 0.44M in males and females respectively, for a total of 2.54M lives saved. An alternative approach for evaluating the effectiveness of the existing control program is to consider the proportion of the ideal difference that was actually achieved, which we found to be 35.2% for males and 26.5% for females. In Chapter 14, Holford and Levy⁽²⁸⁾ consider four alternative models that have been proposed for describing the effect of cigarette smoking on male lung cancer mortality using a similar approach to that employed here to estimate the effect of tobacco control. They found that the number of lives saved to be quite different among the various models, but the percent of the ideal that was actually achieved is fairly consistent with what is found here for males, i.e., 35-40% of the ideal. Major reasons for differences among models lie in estimates of the impact of smoking cessation. For example, the Armitage-Doll multistage carcinogenesis model fitted to the CPS-I data by Knoke *et al* ⁽²⁹⁾ posits that former smokers have a much quicker return to background rates seen among never smokers than the TSCE model. Still, estimates of the proportion of the ideal achieved are similar. Some support for a greater benefit from smoking cessation than is suggested by TSCE is also provided in an analysis by Doll using data from the British Doctors Study which suggested a flattening of the rates when a smoker quits until the effect of age intervenes as risk approaches that of nonsmokers.⁽³⁰⁾

It is not uncommon in epidemiology studies to summarize smoking history for an individual by a single measure, e.g., pack-years of smoking. A person who started smoking 20 cigarettes per day at 14 and quit at 35 would have 410 pack-years for the rest of their lives. A person who started smoking at the same rate at 25 would need to continue smoking until age 46 to obtain the comparable number of pack-years. However, we see a sizable difference in risk for these two scenarios in Figures 1 and 4 for both genders. A fundamental implication of carcinogenesis models that have been fitted to data from large cohort studies, including the TSCE models employed here, is that the impact of carcinogens like tobacco smoke is not easily reduced to a simple measure of exposure. There are huge differences in smoking behavior in a population and we have tried to capture the nuances that would result from these differences in the models described here.

The calibration approach used in the development of the Yale Lung Cancer Model provided excellent agreement between observed and estimated mortality rates (see Figure 2). The total number of cases each year for the APC calibration was identical to the observed for the ATC scenario, but this is a result of solving the normal equations used to obtain maximum

likelihood estimates. One should not be overly confident that a model is correct based on agreement between observed and fitted values, once it has been calibrated for age, period and cohort. Alternative models can also produce excellent fit even if they result from very different estimates of exposure effects, as can be seen in Chapter 14.⁽²⁸⁾ An evaluation of the extent to which the carcinogenesis model accounts for the age, period and cohort effects can be more useful in validating the model. In addition, the effects can help to identify which aspects of temporal trend are not well characterized by the carcinogenesis model. However, even here it may be impossible to determine whether a particular aspect of temporal trend is missed due to the model itself or to inadequacies in the exposure data. Models that appear to have equally good fit to observed data can produce quite different estimates of rates by smoking history, thus affecting estimates of the impact of a control program. The TSCE appears to be based on a rationale that more closely corresponds to the biology of cancer,⁽⁷⁻¹¹⁾ although parameters derived from alternative study cohorts give rise to somewhat different estimates of the effect of tobacco control, especially for estimates of lives saved.

The age contribution to the APC calibration shown in Figure 3(a) has a substantial bend before age 50 and the line is fairly straight thereafter. This suggests that the model provides a better description of the age effect for those older than 50 and the downward correction at younger ages points to an over estimate of risk. A model with three instead of two stages that would be cloned may provide a better description of the age effect. However, this would have little effect on our overall estimates of the number of lung cancer deaths because (a) the calibration corrects for the disparity, and (b) the rates are very low in the younger ages so they contribute few deaths to the total.

In order to obtain estimates of calibration parameter that are not identifiable we adopted the constraint of zero slopes for period, but we emphasize this cannot be verified within the available data. We see in Figure 3(b) that there is very little nonlinear correction required for period, so the TSCE model does well in determining this aspect of trend. However, there are cohort related factors that the model has not fully captured, especially among women. The calibration essentially lowers the estimated rates, especially for cohorts born before 1930. One can only speculate as to the reasons for this, but among the generational aspects that our model is not able to capture are (a) manufacturing changes or brand choice that could affect lethality of cigarettes, (b) behavior changes in the manner in which cigarettes are smoked, (c) differences in efficacy of anti-smoking campaigns, and (d) changes in exposure to other risk factors for lung cancer including secondhand smoke, asbestos, radiation and air pollution.

A limitation of the model described in this manuscript is in the approximation that entailed using a single level of exposure for a particular smoking category. For some components of smoking history like age at initiation and dose, this approximation may not seriously affect the calculations because the variance of the distribution for these values is relatively small in comparison with the average level. However, for duration of smoking or time since cessation, the variance can be great and this can result in large differences in risk, as can be inferred from smoking history scenarios shown in Figure 1. Micro simulation models do not have this difficulty because they simulate the entire smoking history for each individual, which can be considered to be a group of one. Accuracy of the calculations can be improved greatly by refining the categories, especially the smoking duration categories for former smokers. In comparing our results for the TSCE model with those of the Fred Hutchinson Cancer Center we see good agreement. However, the agreement could be improved further by taking more detailed tabulated summaries of exposure for both former and current smokers.

Population based smoking history data are also a limitation in this effort to describe the US lung cancer mortality experience. Cross-sectional data were used to create age-specific initiation and cessation rates, and distributions of dose. These data necessarily rely on recall, and a single value for each individual is generated, which does not account for those who may gradually start or cease smoking or may vary their dose over time. Carcinogenesis models use dates that result from the rates, which were derived through simulation and not directly determined by a survey. In addition, the distribution of the exposure categories can be affected somewhat by changes in the population that result from causes of death other than lung cancer, and these competing causes were not controlled in this version of the model. However, these effects are usually small and not thought to have a sizable impact on the estimates.

The calibrated results presented here apply the same multiplicative correction factor to all smoking categories, including never smokers. We also considered calibration models that only applied to smokers, leaving the temporal effect to be the same for nonsmokers, but this did not have a large effect on the results, and the parameters were occasionally inadmissible resulting divergence for the fitting algorithm. We have greater confidence in the model that also provided correct for nonsmokers. Data on extensive follow-up of large non-smoking populations are not available so one can only speculate as to whether or not the correction should apply to nonsmokers. However, it is conceivable that changes in exposure to various pollutants, including secondhand smoke would induce a temporal trend. In addition, demographic trends in the US could change the mix of individuals who are more susceptible to developing lung cancer.

Further work is needed to explore the sensitivity of results that may arise from alternative carcinogenesis models. Detailed results on the model parameters are derived from large cohorts that are needed to estimate the effects of the diverse smoking histories that are represented in the general population. There are differences in the parameters generated by studies currently available which may be due to the manner they were selected. It will be useful to see the extent to which not only the mathematical model, but the data used to derive estimates of model parameters affect estimates of what is observed in the population.

Future work will extend the model to estimate numbers of incidence cases and incidence rates. This extends the model by introducing an approach for extrapolating mortality rates, and using a back-calculation approach that makes use of the survival experience for lung cancer derived from SEER registries. In addition, it will explore spatial variation across among states which can arise from differences in public policy toward smoking, as well as the effectiveness of cessation efforts in different parts of the country.

Acknowledgments

This work was conducted in collaboration with the Cancer Intervention and Surveillance Network (CISNET) and we are grateful for their insights and assistance with obtaining and analyzing data on smoking behavior over time. Funding was generously provided by a National Cancer Institute grant, CA97432.

REFERENCES

1. Roush GC, Schymura MJ, Holford TR, White C, Flannery JT. Time period compared to birth cohort in connecticut incidence rates for twenty-five malignant neoplasms. *Journal of the National Cancer Institute*. 1985; 74:779–88. [PubMed: 3857375]
2. Roush, GC.; Holford, TR.; Schymura, MJ.; White, C. Cancer risk and incidence trends: The connecticut perspective. Hemisphere Publishing Corp.; New York: 1987.

3. Zheng T, Holford TR, Boyle P, Mayne ST, Liu W, Flannery J. Time trend and the age-period-cohort effect on the incidence of histologic types of lung cancer in connecticut, 1960-1989. *Cancer*. 1994; 74:1556–67. [PubMed: 8062189]
4. United States Surgeon General's Advisory Committee on Smoking and Health. Smoking and health: Report of the advisory committee to the surgeon general of the public health service. U.S. Department of Health, Education, and Welfare, Public Health Service; U.S. Government Printing Office; Washington: 1964.
5. Doll R, Peto R. The causes of cancer. *Journal of the National Cancer Institute*. 1981; 66:1192–308.
6. US Department of Health and Human Services. The health consequences of smoking: A report of the surgeon general. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; Washington:
7. Moolgavkar SH, Venson DJ. A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. *Mathematical Biosciences*. 1979; 47:55–77.
8. Moolgavkar SH, Dewanji A, Venzon DJ. A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. *Risk Analysis*. 1988; 8:383–92.
9. Moolgavkar SH, Luebeck EG. Two-event model for carcinogenesis: Biological, mathematical and statistical considerations. *Risk Analysis*. 1990; 10:323–41. [PubMed: 2195604]
10. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences, USA*. 2002; 99:15095–100.
11. Hazelton WD, Clements MS, Moolgavkar SH. Multistage carcinogenesis and lung cancer mortality in three cohorts. *Cancer Epidemiology, Biomarkers & Prevention*. 2005; 14(5):1171–81.
12. Hazelton W, Jeon J, Meza R, Moolgavkar S. Fhcr lung cancer model.
13. Anderson CM, Burns DM, Dodd KW, Feuer EJ. Birth cohort specific estimates of smoking behaviors for the u.s. Population. *Risk Analysis*.
14. Jeon, J.; Meza, R.; Clarke, L.; Levy, D. Actual and counterfactual smoking prevalence rates in the us population via micro-simulation.
15. Holford TR, Clarke L. Development of the counterfactual smoking histories used to assess the effects of tobacco control. *Risk Analysis*.
16. Fienberg, SE.; Mason, WM. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. In: Schuessler, KF., editor. *translator and editor Sociological methodology 1979*. Jossey-Bass, Inc.; San Francisco: 1978. p. 1-67.
17. Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics*. 1983; 39:311–24. [PubMed: 6626659]
18. Kupper LL, Janis JM, Salama IA, Yoshizawa CN, Greenberg BG. Age-period-cohort analysis: An illustration of the problems in assessing interaction in one observation per cell data. *Communication in Statistics-Theory and Methods*. 1983; 12:2779–807.
19. Kupper LL, Janis JM, Karmous A, Greenberg BG. Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Diseases*. 1985; 38:811–30. [PubMed: 4044767]
20. Holford, TR. Age-period-cohort analysis. In: Armitage, P.; Colton, T., editors. *translator and editor Encyclopedia of biostatistics*. John Wiley & Sons; Chichester: 1998. p. 82-99.
21. Rogers WL. Estimable functions of age, period, and cohort effects. *American Sociological Review*. 1982; 47:774–96.
22. Clayton D, Schifflers E. Models for temporal variation in cancer rates. I: Age-period and age-cohort models. *Statistics in Medicine*. 1987; 6:449–67. [PubMed: 3629047]
23. Clayton D, Schifflers E. Models for temporal variation in cancer rates. Ii: Age-period-cohort models. *Statistics in Medicine*. 1987; 6:469–81. [PubMed: 3629048]
24. McCullagh, P.; Nelder, JA. *Generalized linear models*. Second ed.. Chapman and Hall; London: 1989.
25. Aranda-Ordaz FJ. On two families of transformations to additivity for binary response data. *Biometrika*. 1981; 68:357–63.

26. Holford, TR. Multivariate methods in epidemiology. In: Kelsey, JL.; Marmot, MG.; Stolley, PD.; Vessey, MP., editors. Monographs in epidemiology and biostatistics. Oxford University Press; New York: 2002.
27. Holford TR, Roush GC, McKay LA. Trends in female breast cancer in connecticut and the united states. *Journal of Clinical Epidemiology*. 1991; 44:29–39. [PubMed: 1986055]
28. Holford TR, Levy D. Comparing the adequacy of carcinogenesis models in estimating us population rates for lung cancer mortality. *Risk Analysis*.
29. Knoke JD, Shanks TG, Vaughn JW, T MJ, Burns DM. Lung cancer mortality is related to age in addition to duration and intensity of cigarette smoking: An analysis of cps-i data. *Cancer Epidemiology, Biomarkers and Prevention*. 2004; 13(6):949–57.
30. Doll, R. Cancer and aging: The epidemiologic evidence. In: Clark, RI; Cummley, RW.; McCoy, TE., et al., editors. *translator and editor Oncology 1970. Tenth international cancer conference*. Vol. V. 1971. p. 1-28.

Figure 1.

- (a). Age trends in male lung cancer rates in the HPFS TSCE model starting age 14 or 25, quitting at 35 or never, and smoking 10, 20 or 40 cigarettes/day.
- (b). Age trends in female lung cancer rates in the NHS TSCE model starting age 14 or 25, quitting at 35 or never and smoking 10, 20 or 40 cigarettes/day.

Figure 2.
Observed (dots) and calibrated (APC, PC, AC, and AP) rates (solid lines) for selected age groups by gender.

Figure 3.

- (a). Age effects for APC calibration of TSCE model and no carcinogenesis model by gender.
 (b). Period effects for APC calibration of TSCE model and no carcinogenesis model by gender.
 (c). Cohort effects for APC calibration of TSCE model and no carcinogenesis model by gender.

Figure 4.

(a). APC calibrated age trends in male lung cancer rates in the HPFS TSCE model starting age 14 or 25, quitting at 35 or never, and smoking 10, 20 or 40 cigarettes/day for the 1921 birth cohort.

(b). APC calibrated age trends in female lung cancer rates in the NHS TSCE model starting age 14 or 25, quitting at 35 or never, and smoking 10, 20 or 40 cigarettes/day for the 1921 birth cohort.

Figure 5.

(a). Estimated number of lung cancer deaths per year among males using APC and scale calibration.

(b). Estimated number of lung cancer deaths per year among females using APC and scale calibration.

Summary of curvature effects and fit for models giving deviance chi-square tests (G^2), F-tests ($P < .0001$ in all cases), and percent of the effects explained by the Two Stage Clonal Expansion (TSCE) models by gender.

Table 1

Source	df	Male			Female		
		G ²	F-test ¹	% explained	G ²	F-test ¹	% explained
Smoking Model							
Age	53	7,511.8	141.73	89.54	7,598.9	143.37	73.78
Period	24	577.1	24.05	51.59	494.8	20.61	67.79
Cohort	78	4,018.7	24.05	68.15	5,808.0	74.46	74.61
Goodness of fit	1272	1,830.0			1,554.8		
Scale estimate	1272	1.44			1.22		
No model							
Age	53	71,844.6	1,355.56	-	28,984.9	546.89	-
Period	24	1,192.0	49.67	-	1,536.0	64.00	-
Cohort	78	12,618.4	161.77	-	22,874.4	293.26	-
Goodness of fit	1272	1,710.8			1,555.5		
Scale estimate	1272	1.35			1.22		

¹ F-tests are used because of extra-Poisson variation with numerator df shown on the row and denominator df given for the estimate of scale.

Table II

Estimated number of lung cancer deaths under the Tobacco Control, No Tobacco Control and Complete Tobacco Control by gender.

	Tobacco Control	No Tobacco Control	Complete Tobacco Control
<i>Constant Calibration</i>			
Males	2,067,778	2,608,186	1,056,518
Females	1,051,980	1,250,552	480,375
<i>APC Calibration</i>			
Males	2,067,775	2,670,897	958,862
Females	1,051,978	1,273,151	438,857