STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3055–3070 (DOI: 10.1002/sim.1246)

Mechanistic modelling in large case-control studies of lung cancer risk from smoking

W. F. Heidenreich[1,*,†], J. Wellmann[2,3], P. Jacob[1] and H. E. Wichmann[3]

[1]*GSF-Institute for Radiation Protection, 85758 Neuherberg, Germany*
[2]*Institute of Epidemiology and Social Medicine, University of Münster, 48129 Münster, Germany*
[3]*GSF-Institute of Epidemiology, 85758 Neuherberg, Germany*

SUMMARY

The two-step clonal expansion (TSCE) model is applied to large case-control studies, frequency matched for age, which allow estimation of the RR of lung tumour risk caused by smoking. For estimating background hazard rates, mortality data from the study areas are used to supplement the case-control data. Two approaches are used to analyse the data, based on the unconditional and the conditional likelihoods. They are demonstrated to give nearly identical results. Some model diagnostics are performed and demonstrate a good model fit. Our results indicate that smoking acts on the promotion and transformation parameters, but not on the initiation parameter of the TSCE model. The fitted relative risk of current smokers peaks between ages 50 and 60 years. The relative risk of male ex-smokers decreases strongly with time since end of exposure, but does not reach the risk of non-smokers, and does not decrease as much as for female ex-smokers. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: mechanistic models; two-step clonal expansion model; frequency matched case-control study; smoking; unconditional likelihood; conditional likelihood

1. INTRODUCTION

Mechanistic models of cancer development can supplement traditional epidemiological analysis. They have been applied to cohort studies like the British doctors study [1, 2], or studies in the atomic bomb survivors [3] using Poisson regression. They have also been applied to data sets where individual information on all persons is available, as in uranium miners [4]. The mechanistic models allow the estimation of the hazard as a function of age and exposure histories.

Case-control studies are widely used in epidemiology for reasons of convenience and cost. Various techniques for applying mechanistic models to them have been discussed [5], but to our knowledge no applications to real data have been published. Good candidates for such an

* Correspondence to: W. F. Heidenreich, GSF-Institute for Radiation Protection, 85758 Neuherberg, Germany
† E-mail: heidenreich@gsf.de

Contract/grant sponsor: EU commission; contract/grant number: FIGH-CT1999-00005

Received March 2000
Accepted December 2001

investigation with strong statistical power are case-control studies on lung cancer risk from indoor radon in parts of Eastern and Western Germany [6, 7]. These studies were designed to detect the relatively small lung cancer risk of indoor radon exposure, and collected detailed individual smoking histories in order to adjust for the much larger risks from smoking. They are frequency matched for region, sex and age. It is therefore not possible to estimate age-effects from these data alone, unless external information is supplied. Since the mechanistic models mentioned above include a term for the 'spontaneous lung cancer risk', which depends on age, the case-control data are supplemented with lung cancer mortality data from the study area to make it possible to estimate the spontaneous hazard rate. For the unconditional likelihood this is done by estimating the selection probabilities, for the conditional likelihood the smoking behaviour of the controls is used as a sample of the population.

To these data the two-step clonal expansion (TSCE) model is fitted which has also been used for analysis of the cohort data sets mentioned above. For comparison also two *ad hoc* heuristic models are fitted with the same statistical techniques.

2. DESCRIPTION OF THE DATA ON LUNG CANCER MORTALITY

Wichmann *et al.* [6–9] performed large case-control studies in parts of West and East Germany to investigate the risk of lung cancer due to indoor radon. The data were collected from 1990 to 1997. Cases were collected in specialized lung clinics of the study area. Controls were drawn at random from the mandatory registries of residents of the reference communities (east) or by the random digit dialing technique, adapted to the German telephone system (west). They were frequency-matched according to administrative units (combinations of neighbouring counties ('Landkreise')), sex, and age groups of up to 50, 51–55, 56–60, 61–65, 66–70 and 71–75 years. Older persons were not included in the study. For each person in the study the year of beginning and end of up to 16 smoking intervals were recorded as well as the number of cigarettes per day in each interval. Details can be found in references [6, 7, 9]. Some descriptive statistics for these data are given in Table I. Lung cancer mortality rates from the database for the German cancer atlas [10] for the time period 1986–1990, differentiated for sex, age group and region, were supplied by the German Cancer Research Center. Newer data or incidence data were not available.

Table I. Some numbers for the four study groups obtained by combinations of east, west and male, female. In each group is given the number of strata, cases, controls and the corresponding average number of pack-years (packs per day times smoking period in years).

Study	Sex	Number of strata	Cases		Controls	
			Number	Pack-years	Number	Pack-years
west	male	78	1928	38.7	1999	19.0
west	female	73	570	22.8	634	6.3
east	male	60	1552	28.1	1615	12.9
east	female	56	254	9.0	278	1.8

3. MODELS AND STATISTICAL METHODS

3.1. The two-step clonal expansion model

The two-step clonal expansion (TSCE) model is used, as it is described in references [3, 11], and used, for example in reference [4]. For a scheme of the underlying Markov process see Figure 1; details are relegated to the Appendix. The so-called exact stochastic version is employed, with explicit solutions for piecewise constant parameters, as given in reference [11]. This model provides an expression for the survival probability $S(t)$ that there is no malignant cell in a lung at age t, depending on the exposure history of the lung. The model can also be formulated for the hazard $h_m(t) = -S'(t)/S(t)$ of the occurrence of the first malignant cell. Each of the biological parameters in Figure 1 can depend on the smoking rate d. A linear dependence of the mutational parameters on d is assumed, and both a stepwise plus a linear dependence of the clonal expansion rate is allowed. A model is fitted which depends on seven identifiable numbers.

The first malignant cell cannot be observed, but a cancer can at diagnosis. Therefore a lag-time is used which is distributed between 2 and 8 years. The hazard of lung cancer incidence $h(t)$ is calculated from the hazard for a malignant cell $h_m(t)$ as the average hazard weighted by the survival probability

$$h(t) = \frac{\sum_{l=2}^{8} S(t-l) h_m(t-l)}{\sum_{l=2}^{8} S(t-l)} \tag{1}$$

This formula can also be obtained from a convolution of densities. Some other lag-time distributions were tried and this one was selected for its simplicity.

The relative risk functions are $R_r(t) = h(t)/h_0(t)$, where $h_0(t)$ is the spontaneous hazard function with the same distribution of lag-time as above but without exposure to cigarettes.

3.2. Empirical models

For comparison two empirical models were fitted; one model with

$$R_r(t) = r(\bar{d}) \exp ct^{(s)} \tag{2}$$

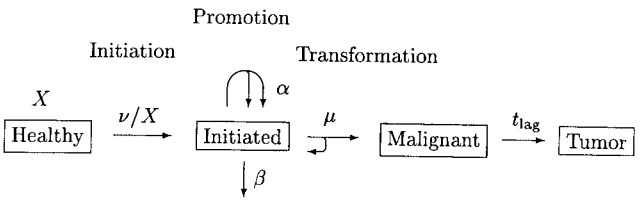

Figure 1. Scheme of the TSCE model. In the pool of X lung cells at risk, initiating mutations occur with rate ν/X. Thus intermediate cells are created by a Poisson process with rate ν. These cells divide with rate α, die or differentiate with rate β and divide in a second rate-limiting event to an intermediate cell and a malignant cell with rate μ. After a lag-time t_{lag} the malignant cell develops into a tumour at diagnosis.

is called modified relative risk (MRR) model in the following. The exposure $\bar{d}$ is the mean smoking rate over the period of smoking and $t^{(s)}$ is the time since end of smoking for ex-smokers, and zero for current smokers. The function $r(\bar{d})$ is piecewise linear with knots at 1, 20, 40 and 60 cigarettes per day; the last value is used for all larger dose rates. This model is motivated by the model of Doll and Peto [1] for current smokers, and by remarks of Richard Doll (private communication, Oxford 1995) for ex-smokers.

The other empirical model assumes that the contribution of smoking to relative risk at a certain time depends on the elapsed time between smoking and observation

$$R_r(t) = \sum_{t'} c(t - t')\mathrm{d}(t') \tag{3}$$

where $\mathrm{d}(t')$ is the smoking rate in the year t', and $c(t - t')$ is defined piecewise constant in the intervals $3-8, -20, > 20$ years, and zero for shorter time shifts. It is called the time since exposure (TE) model here; its motivation comes from models for radon risk [12].

When a hazard function is needed, the spontaneous risk function h_0 of the TSCE model equation (A1) is multiplied to these heuristic models for the relative risk, $h = R_r h_0$. For consistency the same smearing of lag-time as in equation (1) is used.

3.3. Unconditional likelihood

Prentice and Pyke [13] introduced the analysis of frequency matched case-control studies using the so-called unconditional likelihood, which is also known from prospective studies. In deriving their results they formalize the probabilities of cases and controls, θ_{1s} and θ_{0s} in stratum s, say, $s = 1, \ldots, k$, to be selected for the case-control study. These probabilities capture the biased sampling inherent in the case-control design. They enter the likelihood via the quotients $\xi_s = \theta_{1s}/\theta_{0s}$, $s = 1, \ldots, k$. Usually, the selection probabilities are unknown. Furthermore, it is obvious that a case-control study alone cannot yield information on the effect of age, say, when cases and controls are matched for age.

This approach of Prentice and Pyke is used here, but estimates for the ξ are supplied from external mortality data in order to estimate the age-effect from smoking on lung cancer risk. This approach has already been suggested for the linear logistic regression model, see Manski and McFadden [14] and Fears and Brown [15]. Breslow and Cain [16] and Breslow and Zhao [17] comment on this work and show that the additional information from the external rates leads to a lower variance of the estimators compared to a logistic regression with an offset that is unrelated to the matching variables. They present formulae for these variances. Similar considerations will apply in the non-linear setting here. To our knowledge the correct estimation of the variances of the estimators remains an open problem.

A given person in the study area with hazard function $h(t, d)$ develops a lung cancer during the observation interval Δt of 5 years with a probability of about $h(t, d)\Delta t$. The probability of developing the disease in that interval, conditional on being selected for the case-control study and not having developed the disease earlier equals [5]

$$P_s(t, d) = \frac{\xi_s h(t, d)\Delta t}{\xi_s h(t, d)\Delta t + (1 - h(t, d)\Delta t)} \tag{4}$$

For $h(t, d)$ the lung cancer hazard at the interview is used. It depends on age and the individual exposure history. Every case contributes a factor P and each control $1 - P$ to the likelihood

and therefore the log-likelihood

$$\ln L = \sum_{\text{cases}} \ln \hat{P} + \sum_{\text{controls}} \ln(1 - \hat{P}) \tag{5}$$

is maximized [5].

The ξ_s are estimated as quotients of the frequencies of cases and controls of being selected for the study. These frequencies are obtained using the mortality data from the German Cancer Research Center, differentiated for region, 5-year age category and gender. Because of the high lethality of lung cancer, the mortality data are used as proxy for incidence data. These estimates were pooled over the matching regions, weighted by population size, to obtain more stable estimates than separate estimates for each age/region category.

This is done separately for east and west, and for the two sexes. No severe bias is expected from the fact that the mortality data are from an earlier time interval than the case-control data.

To assess the goodness-of-fit, groups of cases and controls with similar risks are formed, for example, non-smokers, heavy smokers etc. Then the expected numbers of cases and controls in each risk group can be estimated using the estimated probabilities $\hat{P}$ (see reference [18]):

$$\begin{aligned} \text{Expected cases} \quad & \sum_i \hat{P}(t_i, d_i) \\ \text{Expected controls} \quad & \sum_i (1 - \hat{P}(t_i, d_i)) \end{aligned} \tag{6}$$

where the sums are over all cases and controls in the risk group. Note that by construction the sum of expected cases and controls in each risk group is equal to the observed one.

3.4. Conditional likelihood

An alternative which does not need selection probabilities is the conditional likelihood [5, 19, 20]. In each stratum, the probability that the cases and controls are distributed as observed is used

$$L_s = \frac{\prod_{i=1}^{m} h(t_i, d_i)}{\sum [\prod_{i=1}^{m} h(t_{li}, d_{li})]} \tag{7}$$

The summation in the denominator is over the set $R(m,n)$ of all subsets $l = (l1, \ldots, lm)$ of size m from $\{1, \ldots, m+n\}$. The denominator has many terms, for example, for 60 cases and 120 controls $\approx 3.60913 \times 10^{48}$ terms. Thus a straightforward calculation is hopeless. However, in an Appendix to the original introduction of this likelihood (in another context), Susannah Howard gave recursion formulas which allow fast calculation [21] (see also reference [22]).

The likelihood for all strata is the product of the likelihoods L_s

$$L_c = \prod_s L_s \tag{8}$$

This likelihood allows the estimation of relative risks, but not spontaneous risks. The parameters that determine the spontaneous hazard of the TSCE model do enter in the relative risk function, but only in a minor way. Therefore they are estimated only crudely with this likelihood; however they cannot be ignored. A way out is to use additional data [5]. The smoking behaviour of the controls of the case-control study is used as a sample for the

smoking behaviour of the population in the year 1988 in the various age groups. From this the expected number Λ_i of cases in each age group i is calculated for the model. Poisson-regression is used to compare with the observed number n_i of mortality cases in the study area. In this calculation, the sample size of controls can be small. To correct for this, we re-scale the population of the study area in such a way that the observed cases are equal to the number of controls. Poisson regression maximizes the likelihood [23]

$$L_p = \sum_i \left(n_i - \Lambda_i + n_i \ln \frac{\Lambda_i}{n_i} \right) \tag{9}$$

The parameters of the various models are obtained by maximizing

$$(\ln L_c + c_p \ln L_p) \tag{10}$$

where L_c is the conditional likelihood equation (8) for the case-control study, L_p is the Poisson likelihood equation (9), and c_p is a weighting coefficient. In order to allow for additional uncertainties in the Poisson-part $c_p = 0.5$ is used. The result should not depend on the precise choice of c_p.

3.5. Parameter estimates

The likelihood functions equation (5) and equation (10) depend on the model parameters in a non-linear way. The software package MINUIT [24] is used to estimate the values of the parameters that maximize the likelihood. Standard errors are estimated from the Fisher information matrix.

4. APPLICATION TO THE DATA ON LUNG CANCER MORTALITY

The parameter values for the three models in Section 3 are estimated, using the two likelihoods given in Sections 3.3 and 3.4, for the four data sets of combinations of east/west and male/female. The deviances of the unconditional likelihood of the various models and the four data sets are given in Table II. It can be seen that the TSCE model compares well with the heuristic models, except for eastern males, where the TE model gives the best fit. Among the heuristic models, the TE model fits consistently better than the MRR model.

The presentation concentrates on the largest data set, that is, western males. Table III shows the expected number of cases and controls as calculated with equation (6). Similarly,

Table II. Deviances of the various models, as calculated with the unconditional likelihood. All numbers are rounded to integers. The second column gives the number of estimated parameters.

Model	Parameters	wm	em	wf	ef
TSCE	7	4230	3499	1341	636
MRR	8	4308	3492	1352	647
TE	6	4258	3474	1348	644

Table III. Observed and expected number of cases and controls for various groups of smoking behaviour, for western males and the TSCE model. By construction, the sum of observed controls and cases is equal to the sum of expected controls and cases. The first block contains non-smokers, persons who started to smoke after age 30, or smoked very little. The remaining persons are grouped in the other blocks in current smokers (including persons who stopped smoking up to 2 years before the interview), and ex-smokers, grouped in years since end of smoking.

Group	Controls		Cases	
	Observed	Expected	Observed	Expected
Non-smokers	473	465.1	30	37.9
$t_a > 30$ years	42	46.3	29	24.7
< 1 pack-year	45	44.4	4	4.6
Current smokers				
1–20 pack-year	101	98.1	96	98.9
20–40 pack-years	242	242.1	519	518.9
> 40 pack-years	176	173.7	643	645.3
Ex-smokers				
3–10 years	195	197.1	265	262.9
11–20 years	256	270.6	188	173.4
> 20 years	469	461.8	154	161.2

Table IV. Observed and expected number of cases and controls for age groups for western males and the TSCE model. Persons older than 75 years were not included in the study. By construction, the sum of observed controls and cases is equal to the sum of expected controls and cases.

Age	Controls		Cases	
[years]	Observed	Expected	Observed	Expected
All	1999	1999.1	1928	1927.9
−50	201	204.7	204	200.3
50–55	277	260.8	236	252.2
55–60	369	388.9	340	320.1
60–65	451	457.2	448	441.8
65–70	446	431.7	437	451.3
70–75	255	255.8	263	262.2

Table IV shows the quality of fit for age groups. The estimated parameters are given in Table V. The parameter y_1 is not significantly different from zero; no effect of smoking on initiation is found. There are strong effects on promotion and transformation.

More fits are done, with data from eastern Germany and among females, with the conditional likelihood and the heuristic models. Tables V, and VI give the resulting parameter estimates. The two likelihoods give a high agreement between the estimated parameter values. The estimated standard errors are quite similar. The effect of smoking on initiation, promotion and transformation is consistent among the data sets. For the males from the east, the step

Table V. Parameters of the mechanistic models. The smoking exposure is given in packs per day (p_d), that is, about 20 cigarettes/day. Standard errors are estimated from the Fisher information matrix. The unconditional likelihood is used except where noted; 'wm' are western males.

Parameter	wm	wm cond.	em	wf
$y_0[10^7 \mathrm{y}^{-2}]$	0.11 ± 0.04	0.13 ± 0.04	0.08 ± 0.04	0.16 ± 0.07
$y_1[p_d^{-1}]$	0 ± 0.07	0 ± 0.08	0.4 ± 0.7	0 ± 0.8
$\gamma_0[\mathrm{y}^{-1}]$	0.134 ± 0.008	0.128 ± 0.007	0.148 ± 0.010	0.117 ± 0.010
$\gamma_s[\mathrm{y}^{-1}]$	0.062 ± 0.007	0.065 ± 0.007	0.025 ± 0.008	0.005 ± 0.011
$\gamma_l[\mathrm{y}^{-1} p_d^{-1}]$	0.027 ± 0.006	0.024 ± 0.006	0.075 ± 0.011	0.039 ± 0.015
$m_1[p_d^{-1}]$	2.0 ± 0.4	1.9 ± 0.4	2.3 ± 0.6	3.9 ± 1.4
$q_0[10^6 \mathrm{y}^{-1}]$	1.2 ± 0.4	1.3 ± 0.5	1.2 ± 0.5	4.3 ± 2.5

Table VI. Parameters of the heuristic models for the relative risk functions. The parameters for the spontaneous risk function are not shown. Standard errors are estimated from the Fisher information matrix.

MRR model	
$c[\mathrm{y}^{-1}]$	-0.093 ± 0.007
$r(0.05[p_d])$	1.4 ± 1.4
$r(1.0[p_d])$	28 ± 4
$r(2.0[p_d])$	46 ± 9
$r(3.0[p_d])$	17 ± 15
TE model	
$t[y]$	$c(t)[y^{-1} p_d^{-1}]$
$3 \leqslant t < 8$	3.7 ± 0.7
$8 \leqslant t < 20$	0.42 ± 0.12
$t \geqslant 20$	0.21 ± 0.05

in promotion is smaller, but the linear effect larger, than for the western males. This may have to do with different reporting of low smoking rates. For the females, the model gives no significant step in promotion, but the linear term is significant; also the transformation term may be larger than for males.

The MRR model gives no significant relative risk at very low exposures, but large increasing risks up to two packs per day. For even higher smoking rates, the risk is going down, as it was also observed (and discussed) in reference [1]. The risk of ex-smokers is halved in about 7.5 years. According to the TE model, a cigarette is most dangerous up to about 8 years after smoking, but has much smaller effect later. Even after more than 20 years it does contribute to the excess hazard.

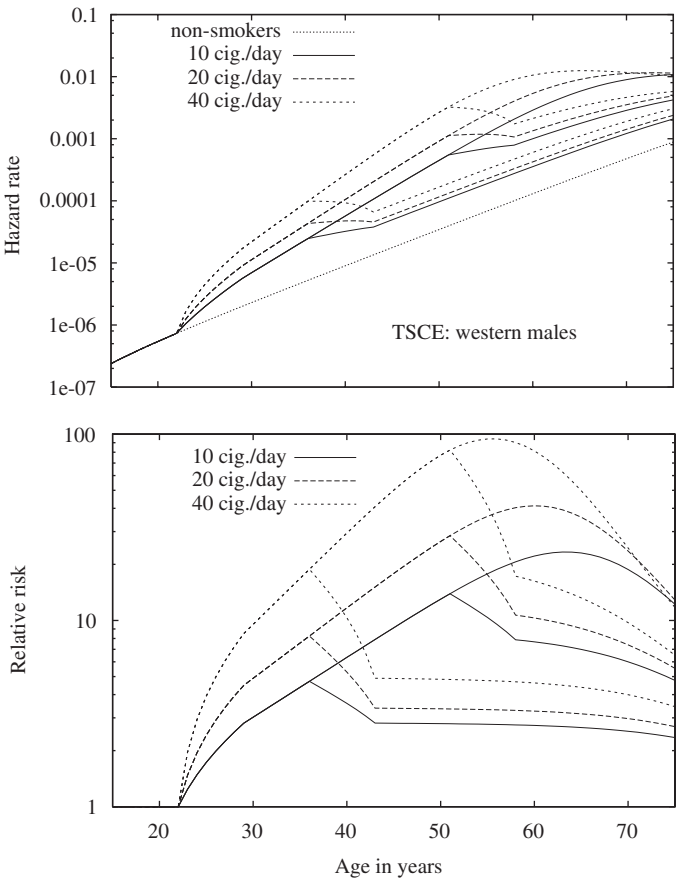

Figure 2. Risk for non-smokers, smokers and ex-smokers, starting smoking at age 20 for life, or until age 34 and 49 years. In each case, the smoking rate is 10, 20 and 40 cigarettes/day (0.5, 1, 2 py). The parameter are for western males, using the TSCE model and the unconditional likelihood.

The models are rather complex and therefore several plots are presented for selected exposure conditions. Figure 2 shows the fitted hazard and the relative risk of the TSCE model for western males. The relative risk reaches a maximum between 50 and 60 years of age, and decreases at higher ages. In the model the fast decrease of the hazard of ex-smokers is due to the transforming effect of cigarette smoke.

Figures 3 to 6 present the fitted hazard risks for other scenarios, which can be compared with Figure 2. Figure 3 shows the result for the conditional likelihood, which is almost indistinguishable from the result using the unconditional likelihood. Figure 4 gives the result for the eastern males. The maximum in relative risk of heavy smokers is a little earlier, and for the high exposure a little higher than for western males. Owing to these differences the two data sets were not pooled. The hazard and the relative risk of females from the west is plotted in Figure 5. While the general shape is similar, the estimated spontaneous hazard of males at high age is about 40 per cent higher than for females; also the relative risks of males

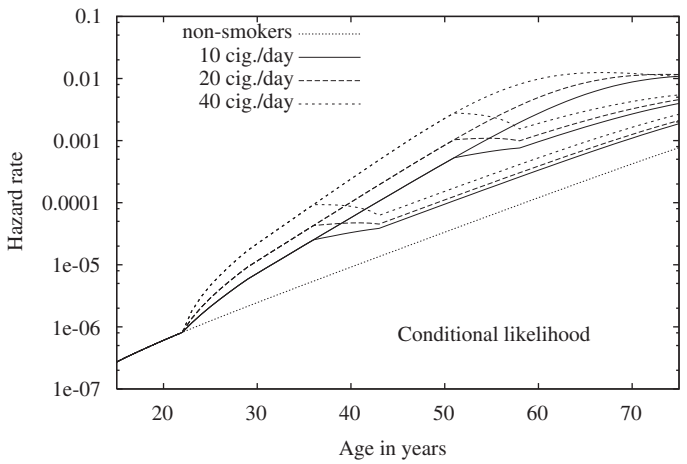

Figure 3. Absolute risk for western males, using the TSCE model and the conditional likelihood.

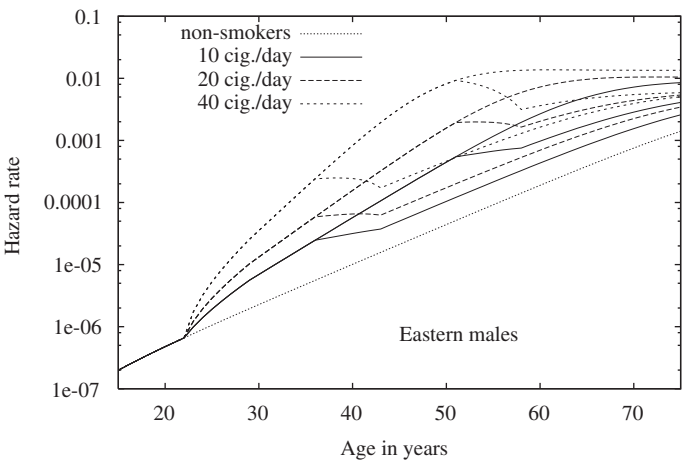

Figure 4. Absolute risk for eastern males, using the TSCE model and the unconditional likelihood.

above age 40 are higher. This is most pronounced for ex-smokers, where females approach the risk of non-smokers much closer than males. Again, this may be in part due to different reporting.

5. DISCUSSION

Dose-response models can be fitted well to case-control data. This applies to both mechanistic and heuristic models. The additional technical difficulties of the case-control design relative to the cohort design are limited. For application of the mechanistic TSCE model some additional

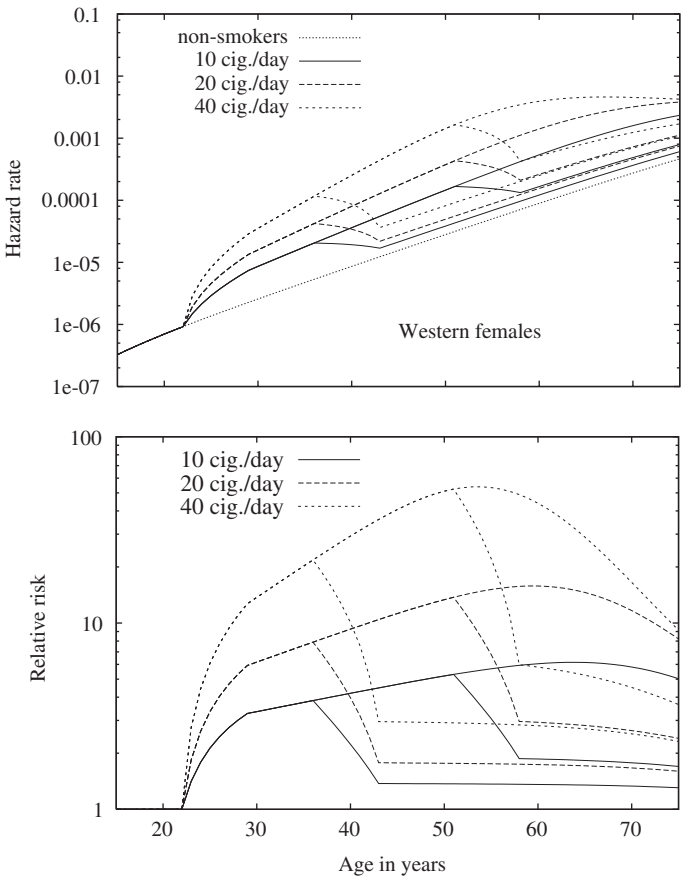

Figure 5. Risk for western females, using the TSCE model and the unconditional likelihood.

information is necessary which allows estimation of the age dependence of the spontaneous hazard function.

The unconditional and the conditional likelihood give comparable results. Knowledge of selection probabilities allows quality of fit tests via comparisons of observed and expected numbers. As these are also needed for the unconditional likelihood, that approach may be favoured. Use of the conditional likelihood requires non-standard techniques for estimating the age dependence of the spontaneous hazard.

Large case-control studies can be a valuable source for testing mechanistic models and for learning about the mechanism of cancer induction. The high risk of smokers, the large number of lifelong smokers, and ex-smokers make modelling of smoking risk a good testing ground for mechanistic models.

In general, confounding may be a cause of concern in case-control studies. In the TSCE models it can be addressed by incorporating potential confounding variables into the model in order to estimate their effect simultaneously, as was done, for example, in references [4, 25]. The data set used for this analysis did not contain any information about additional exposures,

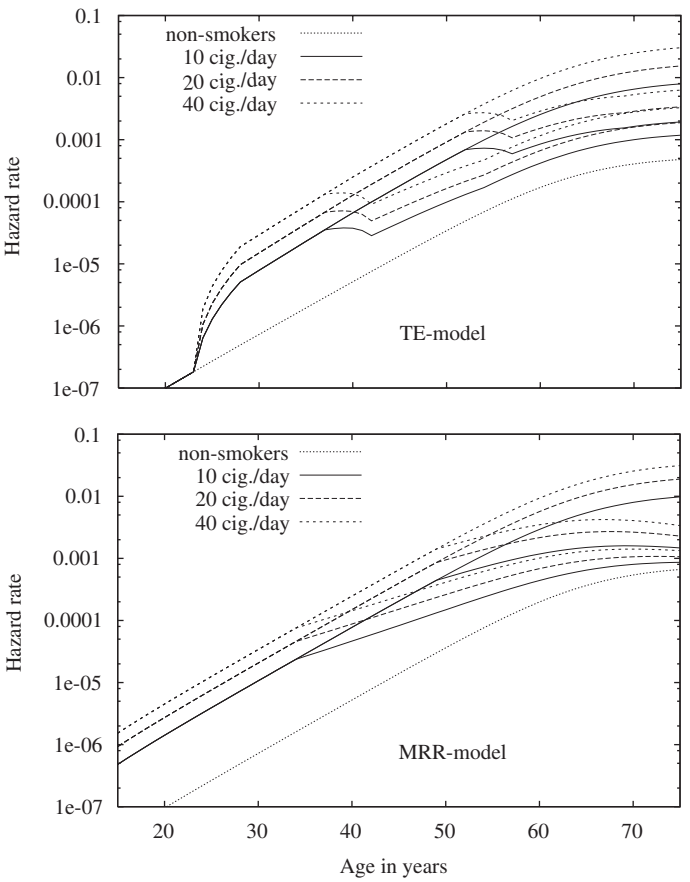

Figure 6. Absolute risk for western males, using the heuristic models and the unconditional likelihood.

and these are not available for all persons in the present data set. Also, for lung cancer, smoking is the dominant risk factor, far more important than indoor radon, for which the studies were designed; the analysis here gives relative risks due to smoking well beyond 10. The most exposed groups with more than 140 Bq/m^3 radon comprise about 8 per cent of the persons in the eastern study and about 3 per cent in the western study. The estimated odds ratios due to radon are consistently below 2 even for these groups [6, 7]. An indoor radon exposure to 100 Bq/m^3 of radon is expected to induce a relative risk of 1.09 [26]. Therefore inclusion of radon is not expected to alter the estimates of the effects of smoking substantially. An inclusion of other confounders, like, for example, profession, would be an interesting subject but lies beyond the scope of this work.

A greater effect on the detailed risk estimates is expected from the form of the mechanistic model. The TSCE model makes specific assumptions about the carcinogenic process which are surely oversimplifications of the true processes (see Appendix). Even within that framework, many versions of the model are possible, which differ, for example, in the assumed shape of the exposure dependence of promotion [27]. It rests with the persons who do the analysis

to choose a version of the model which does justice to the data. For that reason, heuristic models are used in parallel here, and quality of fit analyses as in Tables III and IV were done. Also statements about the action of smoking in the model are only made when they are well supported by the data.

The dose–response in the TSCE model can qualitatively be summarized by: smoking acts on promotion and transformation, not on initiation. A linear dependence of transformation on the smoking rate is consistent with the data, while for promotion, there may be a step-like contribution in addition.

The shape of the hazard functions as estimated with the TSCE model and presented in Figures 2 to 5 can be summarized as follows. While the hazard of non-smokers increases roughly exponentially up to age 75, it flattens out for lifelong smokers. Thus surprisingly the relative risk of lifelong smokers decreases beyond an age of about 60 years. Note, that Becker in a reanalysis of the published British doctor's data [28] also found a decreasing relative risk at high age in lifelong smokers. The hazard of ex-smokers reduces fast, but not to the hazard of non-smokers. It would be desirable to verify these results in prospective studies. The smoking information, especially on the early smoking habits of the older age groups, may be less reliable in this retrospective study.

The heuristic models in Figure 6 have quite different hazard functions, because the various models do limit the shape of the relative risk functions which can be obtained. The non-smoker hazard levels at high age in these models; the two heuristic models cannot decrease the relative risk of lifelong smokers, and therefore average the hazard at high ages. As a consequence, the spontaneous hazard rises more steeply for medium ages, and is smaller at low ages, where there are no data. This demonstrates one way how mechanistic models can supplement epidemiological models; even the conceptually simple TSCE model can produce highly non-linear dependence of risk on exposure. Some of the features of its risk functions may be used to improve heuristic models.

APPENDIX: THE TSCE MODEL

The TSCE model (see Figure 1) supposes the presence of $X(t)$ susceptible stem cells at age t. For the lung, these cells are not identified for sure, but candidates are basal cells and secretory cells, which both sit in the lining of the bronchial tubes. For the purpose here, the number of these cells is assumed to be constant over life. Each of these cells is under the risk of an initiating event, for example, a mutation, at a rate of $\mu_1(t)$. Thus initiated cells are created in a Poisson process at a rate $\nu(t) = X(t)\mu_1(t)$. Each initiated cell can divide into two initiated cells at a rate of $\alpha(t)$ and it can die or differentiate at a rate of $\beta(t)$. The net growth rate is called promotion. Each initiated cell can also divide into an initiated cell and a malignant cell with a rate of $\mu(t)$. This transforming event again is not necessarily one specific mutation but may be an approximation of a more complicated series of molecular events which is described by a rate-limiting rate. The TSCE model is a generalization of the model of Knudson for retinoblastoma [29] and can be thought of as a mathematical formalization of the initiation–promotion–transformation paradigm of carcinogenesis.

A stochastic treatment is necessary to describe the birth and death process of the initiated cells. The Kolmogorov equations of the Markov process are solved using standard techniques. The hard part is a Riccatti equation. For constant, and for piecewise constant parameter

values, it can be solved explicitly. The probabilities for any number of initiated cells and no malignant cells are then calculated as a function of age [30, 31]. From this both the hazard and the survival can be given in a closed form [11].

When estimating parameters from incidence data, there is a complication known as the identifiability problem. For example, the parameters $\mu_1(t)$ and $X(t)$ cannot be determined separately from incidence data, as only the product appears in the hazard function. One can either use identifiable parameters or fix one of the parameters, for example, $X(t)$, to some plausible value. Identifiable parameters are used here. In one set of such parameters, the hazard function for constant parameters is [32]

$$h_0(t) = \frac{y_0(e^{(\gamma_0+2q_0)t} - 1)}{\gamma_0 + q_0(e^{(\gamma_0+2q_0)t} + 1)} \tag{A1}$$

The three parameters y_0, γ_0 and q_0 are estimated. They are functions of the biological parameters ν, μ, α, β, which are given below in a more general context. The identifiable parameters are selected such that each of them influences a particular part of the hazard function, and therefore they can be estimated reliably from data. Specifically, the hazard starts out linearly with coefficient y_0, goes over to an exponential growth with the effective clonal expansion rate γ_0, until it asymptotically becomes a constant y_0/q_0.

The exposure pattern of smokers in the data set considered here can be brought in the form of piecewise constant parameters: the rate of cigarette smoke is collected in age intervals. The various biological parameters of the model are assumed to depend on that rate during the respective age intervals. For the identifiable combinations used here it has been shown [11] that in principle the dependence on the exposure rate can be extracted from sufficiently powerful incidence data. In practice it is necessary to make some assumption on these functions in order to limit the number of estimated parameters. Next the choices for these dose–effect relations are motivated.

Initiation occurs spontaneously, or due to some external influences, like smoking, chemicals or radiation. The same applies to transformation. Both of these parameter functions are expected to depend roughly linearly on the smoking rate, as it is expected for mutations. Promotion is an average growth advantage of initiated cells. Its dependence on the smoking rate is not known from biological experience and therefore has to be estimated from fits to the cancer data. Based on earlier experience, a step with smoking and a linear dependence on smoking rate is allowed.

The same identifiable parameters as in reference [33] are used here. They are selected such that the biological effects of initiation, promotion and transformation are clearly separated. The left equality sign in each of the equations defines the identifiable parameter in terms of the biological ones; the right equality sign gives the assumed shape of the dependence on smoking rate d:

$$\begin{aligned}
\text{Initiation} \quad & Y(d) = (\nu(0)\mu(0))\left(\frac{\nu(d)}{\nu(0)}\right) = y_0(1 + y_1 d) \\
\text{Transformation} \quad & m(d) = \mu(d)/\mu(0) = 1 + m_1 d \\
\text{Promotion} \quad & \gamma(d) = \alpha(d) - \beta(d) - \mu(d) = \gamma_0 + \gamma_s(\text{if } d > 0) + \gamma_1 d \\
& q(d) = \tfrac{1}{2}[\sqrt{\{\gamma^2(d) + 4\alpha(d)\mu(d)\}} - \gamma(d)] = (1 + m_1 d)q_0
\end{aligned} \tag{A2}$$

The parameter q describes the stochastic effect which causes the hazard to level asymptotically. For lifelong constant exposure, the hazard approaches a constant value $Y(d)m(d)/q(d)$ for advanced ages. The quantity q is approximately $\mu/(1-\beta/\alpha)$ [11]. When $q(d))$ is made proportional to $\mu(d)$, as done here, then the quotient β/α is approximately independent of exposure.

The hazard for at least one malignant lung cell is calculated for each person using the explicit recursion formulae over the age intervals given in reference [11]. The recursion can be converted in a straightforward way into about 15 lines of fast computer code.

ACKNOWLEDGEMENTS

We thank Dr Suresh Moolgavkar, Dr Georg Luebeck and Dr Wolfgang Jacobi for discussions. We further thank Dr Nikolaus Becker for providing data from the German Cancer Research Center. This work was in part (WFH and PJ) supported by the EU under contract number FIGH-CT1999-00005.

REFERENCES

1. Doll R, Peto R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. *Journal of Epidemiology and Community Health* 1978; **32**:303–313.
2. Moolgavkar SH, Dewanji A, Luebeck G. Cigarette smoking and lung cancer: reanalysis of the British doctor's data. *Journal of the National Cancer Institute* 1989; **81**:415–420.
3. Heidenreich W, Jacob P, Paretzke H. Solutions of the clonal expansion model and their application to the tumor incidence of the atomic bomb survivors. *Radiation Environment Biophysics* 1997; **36**:45–58.
4. Luebeck E, Heidenreich W, Hazelton W, Paretzke H, Moolgavkar S. Biologically-based analysis of the Colorado uranium miners cohort data: age, dose and dose-rate effects. *Radiation Research* 1999; **152**:339–351.
5. Moolgavkar SH. When and how to combine results from multiple epidemiological studies in risk assessment. In *The Role of Epidemiology in Regulatory Risk Assessment*, Graham JD (ed.). Elsevier: Amsterdam, 1995; 77–90.
6. Kreienbrock L, Kreuzer M, Gerken M, Dingerkus G, Wellmann J, Keller G, Wichmann HE. Case-control study on lung cancer and residential radon in Western Germany. *American Journal of Epidemiology* 2001; **153**(1): 42–52.
7. Wichmann H, Gerken M, Wellmann J, Kreuzer M, Kreienbrock L, Keller G, Wölke G, Heinrich J. Lungenkrebsrisiko durch Radon in der Bundesrepublik Deutschland (Ost)—Thüringen und Sachsen. In *Fortschritte in der Umweltmedizin*, Wichmann von HE, Schlipkötter HW, Füllgraf G (eds.). Ecomed Verlagsgesellschaft: Landsberg, 1999.
8. Wichmann H, Kreienbrock L, Kreuzer M, Gerken M, Dingerkus G, Wölke G. Lungenkrebsrisiko durch Radon in der Bundesrepublik Deutschland. Erste Risikoanalysen in West- und Ostdeutschland. In *Radon-Statusgespräch 1998*, *Berichte der SSK*. G. Fischer: Stuttgart, 1998; Chapter 17, 103–121.
9. Wichmann H, Kreienbrock L, Kreuzer M, Gerken M, Dingerkus G, Wellmann J, Keller G. Lungenkrebsrisiko durch Radon in der Bundesrepublik Deutschland (West). *Fortschritte in der Umweltmedizin*. Ecomed Verlagsgesellschaft: Landsberg, 1998.
10. Becker N, Wahrendorf J. *Atlas of Cancer Mortality in the Federal Republic of Germany*. Springer: Berlin, 1997.
11. Heidenreich W, Luebeck EG, Moolgavkar SH. Some properties of the hazard function of the two-mutation clonal expansion model. *Risk Analysis* 1997; **17**:391–399.
12. National Research Council. Committee on Health Risks of Exposure to Radon (BEIR VI). *Health Effects of Exposure to Radon*. National Academy Press: Washington, D.C., 1999.
13. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**: 403–411.
14. Manski CF, McFadden D. Alternative estimators and sample designs for discrete choice analysis. In *Structural Analysis of Discrete Data with Econometric Applications*, Manski CF, McFadden D (eds). MIT Press, Cambridge: Massachusetts, 1981; 2–50.
15. Fears TR, Brown CC. Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* 1986; **42**(4):955–960.
16. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika* 1988; **75**:11–20.
17. Breslow NE, Zhao LP. Logistic regression for stratified case-control studies. *Biometrics* 1988; **44**(3):891–899.

18. Breslow N, Day N. *Statistical Methods in Cancer Research. Volume I—The Analysis of Case-control Studies*. IARC Scientific Publications No. 32: Lyon, 1980.
19. Prentice R, Breslow N. Retrospective studies and failure time models. *Biometrika* 1978; **65**:153–158.
20. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
21. Howard S. Contribution to the discussion of a paper by DR Cox: Regression models and life-tables. *Journal of the Royal Society*, *Series B* 1972; **34**:210–211.
22. Gail MH, Lubin JH, Rubinstein LV. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* 1981; **81**:703–707.
23. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd edn. Chapman & Hall: London, 1991.
24. James F. *Minuit function minimization and error analysis*, *version 94.1*. CERN: Geneva, 1994.
25. Hazelton WD, Luebeck EG, Heidenreich WF, Moolgavkar SH. Analysis of a historical cohort of Chinese tin miners with arsenic, cigarette, and pipe smoke exposures using the biologically based two-stage clonal expansion model. *Radiation Research* 2001; **156**:78–94.
26. Lubin JH, Boice JD. Lung cancer risk from residential radon: meta-analysis of eight epidemiological studies. *Journal of the National Cancer Institute* 1997; **89**:49–57.
27. Heidenreich W, Atkinson M, Paretzke H. Radiation induced cell inactivation can increase the cancer risk. *Radiation Research* 2001; **155**:870–872.
28. Becker N. Cigarette smoking and lung cancer: a reconsideration of the British doctors' data with cumulative damage models. *Epidemiology* 1994; **5**:27–34.
29. Moolgavkar SH, Knudson, Jr AG. Mutation and cancer: a model for human carcinogenesis. *Journal of the National Academy of Sciences* (*USA*) 1981; **66**:1037–1052.
30. Moolgavkar S, Luebeck G. Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. *Risk Analysis* 1990; **10**:323–341.
31. Moolgavkar SH. Stochastic models of carcinogenesis. In *Handbook of Statistics 8: Statistical Methods in Biological and Medical Sciences*, Rao R, Chakraborty R (eds). Elsevier: Amsterdam, 1991; 373–393.
32. Heidenreich W. On the parameters of the clonal expansion model. *Radiation Environment Biophysics* 1996; **35**:127–129.
33. Heidenreich W, Jacob P, Paretzke H, Cross F, Dagle G. Two step model for fatal and incidental lung tumor risk in rats exposed to radon. *Radiation Research* 1999; **151**:209–217.

