Environmental Health Perspectives
Vol. 81, pp. 169–188, 1989

Multistage Models for Carcinogenesis

by David A. Freedman* and William C. Navidi†

The multistage model is tested on several human and animal data sets. It fits in some cases but not in others. With human lung cancer data, there is a drop in risk for ex-smokers quite different from the predictions of the model. The results are not conclusive but are compatible with the view that the multistage model provides a family of curves that often fit cancer incidence data, but may not capture the underlying biological reality.

Introduction

The Armitage-Doll multistage model says in essence that a cell progresses to malignancy through the states of a Markov chain (*1*). This model is often used in cancer risk assessment for example, by the U.S. Environmental Protection Agency (*2*), and it is often cited in discussions of the biological mechanisms of cancer, for example, by the International Agency for Research on Cancer (*3*). It therefore seems worthwhile to review the model and assess its fit to some of the main available data sets, which is the object of the present paper.

To state the model a bit more carefully (*4–6*): A normal cell goes through a definite sequence of stages until it becomes cancerous. Absent carcinogenic exposure, waiting times in the various stages are assumed to be independent, exponential random variables. So, there is a background rate of progression through each stage, which may be different for the different stages.

An animal or a human tissue is a collection of cells and fails (gets cancer) when the first cell in the collection fails. Thus, the failure time for the tissue is the minimum of the failure times of its component cells. Different cells are assumed to be independent with identically distributed failure times.

The next assumption: If a subject is exposed to a carcinogen such as tobacco smoke, the rate of progression through the various stages increases in proportion to dose; the constant of proportionality depends on the stage. For the insensitive stages, this constant is zero; for the sensitive stages, the constant of proportionality is positive. Stages that may be estimated as sensitive or insensitive, depending on how the data turn out, will be termed "potentially sensitive."

*Statistics Department, University of California, Berkeley, CA 94720.
†Mathematics Department, MC 1113, University of Southern California, Los Angeles, CA 90089-1113.
Address reprint requests to D. A. Freedman, Statistics Department, University of California, Berkeley, CA 94720.

In order for the usual approximations to work, it is also necessary to assume that the time to pass through a stage tends to be much larger than the lifetime of the animal (*7*). The rates of progression through the various stages are assumed to be the same for all cells and all subjects. In risk assessment, constancy of certain rates is assumed even across species.

With a final assumption, independence of competing risks, the model can be used to generate a likelihood function for data; parameters can be estimated by maximum likelihood. Then the adequacy of the fit can be assessed by a chi-squared test. The intent was to follow this strategy rigorously, but we ended up making some approximations for mathematical convenience, others to get numerical algorithms to converge, and still others to accommodate experimental designs.

In a strict modeling approach, the details become quite irritating. Perhaps as a result, most published efforts to assess the fit of the model tend to involve simple approximations to the likelihood function, and goodness-of-fit tests are seldom made (*8–11*). On the other hand, the statistical strategy followed here is similar to that in Brown and Hoel (*12–13*).

There are many variations on the model, for example, allowing a latency period between malignancy and the clinically observable end point. Polynomial dose response at each stage has also been considered and transitions from higher order to lower order states. Dose thresholds are sometimes used or nonlinear transformations of time. Random parameters are another option. There is little doubt that, given a data set, one variation on the model or another can be made to fit. Our question runs the other way round: Given a version of the model, will it fit a variety of data sets? For that purpose, we elected to start with the version described here, which is relatively simple and in general circulation.

Versions of the model are widely used in risk assessment, although their biological basis is more than a little obscure. In particular, despite remarkable progress on the

Table 1. Summary of results.

Data set	Species	End point	Results
British doctors	Humans	Lung cancer	Fits; 5–8 stages; one or two sensitive; heavy smokers included
Veterans	Humans	Lung cancer	Does not fit: p = 5/100,000
ACS	Human males	Lung cancer	Marginal fit; p = 1% 6 stages with 1st sensitive, or 7 stages with 6th sensitive fails on cross-validation
	Human females	Lung cancer	Fits; 3 to 8 stages; one sensitive
Mega-mouse	Mice	Bladder tumors	Does not fit
		Liver tumors	Fits; 7 stages, 2 sensitive
Peto et al.	Mice	Skin cancer	Does not fit

identification of DNA lesions in tumor cells, there is still no experimental verification of the assumptions in the multistage model at the level of the cell. For example, there is no general biological definition for the stages in the model; in most cases, these remain purely hypothetical constructs (*5,14–18*). Striking recent papers on the genetics of cancer include Bodmer et al. (*19*), Naylor et al. (*20*), and Solomon et al. (*21*).

Results

Overview

This section will review the data sets considered and summarize conclusions (Table 1). The most carefully studied application of the model is to lung cancer, which is considered first; then comes experimental data on animals.

Other Findings on Lung Cancer. For the British doctors, the Dorn veterans, and the ACS males, the model overpredicts around the edges of the data set. Models fitted to continuing smokers do not predict the risk for ex-smokers at all well: The models predict that excess risk will continue to increase (or stay constant) after quitting, while the data show a decrease.

Human Lung Cancer

British Doctors. Doll and Peto report on smoking and lung cancer in their seminal cohort study of British doctors (*9*). The data quality is considered to be excellent; dose was ascertained on three separate occasions. One drawback is the absence of information on age at start of smoking; following Doll and Peto, this is imputed as 22.5 years (including some allowance for the time from malignancy to death). Furthermore, although the study lasted 20 years with about 34,000 subjects, the number of events (lung cancer cases) is relatively small.

The data set used here, reported in Doll and Peto (*9*), selects only subjects who smoked at a nearly constant rate; only 215 events out of 571 are kept. The published data on ex-smokers are not in usable form and the unpublished data do not appear to be available.

Data on nonsmokers or current smokers are summarized. For this cohort, there is a paradoxical drop in risk (events per 1000 person years) for the highest dose group and at the highest ages. This is more readily seen from Doll and Peto's table than from our aggregation (Tables 2 and 3).

A variety of models fit quite well, with five to eight stages; models with four or nine stages did not fit. Previous work suggests five or six as the number of stages. (The plural "models" refers to special cases of the singular "multistage model.")

Is the dose response linear or quadratic? In the multistage model, this comes down to asking whether there is one sensitive stage or two. This question has been much debated. In our models, with five, six, or seven stages, the first and next-to-last appear to be sensitive. With eight stages, the first need not be sensitive. We view the data on the British doctors as compatible with a linear or a quadratic dose response; the latter provides a significantly better fit. (For the veterans or the ACS volunteers, a linear dose response fits about as well, or as badly, as quadratic.)

We allow either the first and next-to-last stages, or the first and last, to be sensitive. Both types of models give similar fits on current smokers, but make very different predictions for ex-smokers. The conventional choice is to allow the first and next-to-last stages to be sensitive (*22*).

When such models are fitted to the continuing smokers, the first stage usually turns out to be sensitive, and then the model predicts that the excess risk will continue to rise after cessation of smoking (Eq. 3). The rate of rise is quite appreciable. It is commonly believed, however, that the excess risk remains constant after cessation of exposure. This belief and the model do not fit together. Furthermore, the data suggest that the risk starts to drop on cessation of smoking.

Coming back to the continuing smokers, Doll and Peto (*9*) found it necessary to exclude the heavy smokers and gave ingenious arguments based on measurement error to justify the exclusion. They fit an appealing (but nonmultistage) model. Working fairly directly from the likelihood function given by the model, the fit is good whether the heavy smokers are kept in or not. The tables include the heavy smokers and a broader age range than that allowed by Doll and Peto (25–84 years). On the other hand, different parts of the data set do look different: A model fitted to the data used by Doll and Peto (age 40–79, dose 40 or less) overpredicts risk for the rest.

Table 2. The British doctors: Events/person years.[a]

Age, years	Dose, cigarettes/day, range and (mean)									
	0	1-4 (2.7)	5-9 (6.6)	10-14 (11.3)	15-19 (16.0)	20-24 (20.4)	25-29 (25.4)	30-34 (30.2)	35-40 (38.0)	40+ (50.9)
20-24	0/378	0/19.5	0/38	0/91.5	0/91	0/57	0/7.5	0/2	0/2.5	0/0
25-29	0/5099.5	0/400	0/701.5	0/1529	0/1427	0/1424	0/304.5	0/153	0/46	0/10.5
30-34	0/10838	0/914	0/1762.5	0/3270	0/3343	0/3966.5	0/1042.5	0/582.5	0/224.5	0/32.5
35-39	1/15105	0/1156.5	0/2178	0/3819.5	0/4649.5	0/6003.5	0/1991.5	0/1108.5	0/545.5	0/110.5
40-44	0/17846.5	0/1216	0/2041.5	1/3795.5	0/4824	1/7046	0/2523	1/1715.5	0/892.5	0/234
45-49	0/15832.5	0/1000.5	0/1745	1/3205	1/3995	1/6460.5	2/2565.5	2/2123	0/1150	0/305.5
50-54	1/12226	0/853.5	0/1562.5	2/2727	4/3278.5	6/5583	3/2620	3/2226.5	3/1281	0/335.5
55-59	2/8905.5	1/625	0/1355	1/2288	0/2466.5	8/4357.5	5/2108.5	6/1923	4/1063	1/284
60-64	0/6248	1/509.5	1/1068	1/1714	2/1829.5	13/2863.5	4/1508.5	11/1362	7/826	1/183.5
65-69	0/4351	0/392.5	1/843.5	2/1214	2/1237	12/1930	5/974.5	9/763.5	9/515	1/120
70-74	1/2723.5	1/242	2/696.5	4/862	4/683.5	10/1055	7/527	2/317.5	5/233	2/52
75-79	2/1772	0/208.5	0/517.5	4/547	5/370.5	7/512	4/209.5	2/130	2/88.5	0/18.5
80-84	0/1185.5	0/173	0/281	1/314	1/180.5	1/188	0/81	2/37	1/36	0/2.5
85+	0/870.5	0/77.5	0/149	0/123.5	1/61.5	1/67.5	0/28	0/1	0/.5	0/0

[a]Reproduced from Doll and Peto (*9*). Doll and Peto use the data inside the smaller box; we use the data in the larger box.

Table 3. The British doctors: Events/person years, aggregation of the data from Table 2.[a]

Age	Dose, cigarettes per day					
	0	1-9	10-19	20-29	30-40	41 or more
54 or less	2/76,948	0/15,532	9/39,863	13/41,530	9/12,048	0/1,028
55 to 64	2/15,154	3/3,558	4/8,298	30/10,838	28/5,174	2/468
65 to 74	1/7,074	4/2,174	12/3,996	34/4,486	25/1,829	3/172
75 or more	2/2,958	0/1,180	11/1,412	12/990	7/292	0/21

[a]Age range 25 to 84 years; heavy smokers included.

Dorn Veterans. Kahn (*23*) reports on Dorn's cohort study of about 300,000 American veterans, begun in 1954; also see Rogot (*24*) and Whittemore (*25*). The data used here come from a tape supplied in 1981 by the National Cancer Institute (NCI) under the Freedom of Information Act. This tape combines the 1954 and 1957 cohorts, reports on follow-up through 1969, and has been edited by NCI personnel. Data on the tape therefore differ from tables in Kahn (*23*).

The data quality may be questioned; in particular, dose was ascertained only once. On the positive side, this data set is quite large (1266 events); it has information on age at start of smoking and it includes ex-smokers. The risk for current smokers increases with dose.

This data set is sometimes cited as supporting the model (*10*). However, as far as we can see, no version of the model fits the veterans data. For nonsmokers and current smokers, the best had six stages, with only the fifth being sensitive: $\chi^2 = 50$ on 17 degrees of freedom, $p \approx 5/100{,}000$. Residuals (observed − expected) were quite regular, tending to be negative at the lowest or highest age and dose groups, positive at intermediate groups.

When this model was used to predict the risk for ex-smokers as a function of years since quitting, the ratio of observed to expected decreased steadily: Indeed, the excess risk in the model remains constant, while the data show a drop in risk.

ACS Volunteers. This study is described by Hammond (*26*). L. Garfinkel of the ACS provided a table of person-years and events for current smokers over the period July 1960 to June 1965, by age, age at start of smoking, dose, and sex. (The table differs in some respects from published data.) This is the only study with substantial numbers of women. Because of the large number of subjects (about 440,000 men and 570,000 women) there are a lot of events: 1542 for the male smokers, 164 for the females. The men smoked more heavily than the women and had higher cancer rates even controlling for smoking.

The data quality seems good. The risk for smokers goes up with dose. There is some deficit in events beyond age 79. This can be detected in the original data, but gets lost in the aggregated Table 10. The increased risk for ages 75 to 79 swamps the decrease in ages 80 and beyond, where the number of person-years is relatively small.

For current male smokers, models with 3 to 10 stages were tried. The best-fitting had 6 stages ($\chi^2 = 28$ on 13 degrees of freedom, $p = 1\%$); the estimated sensitivity for the fifth stage was negligible. Compared to 1542 events, the fit seems good. However, the pattern of residuals was as in the veterans study. Rates for the nonsmokers, estimated from the current smokers through the model, were much higher than the observed rates. Largely by accident, we got the data on never-smokers or ex-smokers only after fitting the models, so there was a genuine opportunity for cross-validation.

For current female smokers, a variety of models fit, with three to eight stages; the best was five stages; almost any pattern of sensitivity is obtainable. Estimated back-

ground rates are unreasonably low; indeed, such rates can be constrained to zero without any trouble.

Animal Data Sets

With human lung cancer data, dosimetry is problematic; the accuracy of diagnosis is open to question too. In principle, experimental data on animals should be better. Of course, animal experiments have problems of their own (*16*). Perhaps surprisingly, it is not so easy to get animal data suitable for testing multistage models. In particular, data on times of tumors are seldom published. (The NCI/NTP bioassays might be a good data source, if properly pooled.)

Mega-Mouse Study. This experiment is described by Staffa and Mehlman (*27*); see especially Littlefield et al. (*28*). The experiment involved about 24,000 mice; the carcinogen was 2-acetylaminofluorene (2-AAF); bladder tumors and liver tumors were the two end points.

A serial sacrifice design was employed. One group of animals was on continuous exposure. For a second group, exposure ceased at predetermined times before sacrifice. For all animals other than controls, exposure started at birth; this eliminates one interesting variable from the multistage model and reduces power in testing it.

Like other authors (*12–13*), we could not fit a multistage model to the bladder cancer data. With liver cancer, the model does fit a substantial part of the data, but then extrapolations to the rest of the data set are not so successful (*27–31*).

Peto Mice. The object of the beautiful experiment described in Peto et al. (*10*) was to demonstrate that duration rather than aging *per se* affects cancer risk. The carcinogenic agent was benzpyrene painted on the skin. The end point was malignant skin cancer. No control group was provided, probably because the tumor has no spontaneous incidence. Only one dose level was used, limiting the power of statistical tests. Painting was started at 10, 25, 40, or 55 weeks of age. The point is that incidence depends on duration of exposure not age at start, but Peto et al. do not really test the fit of the multistage model to the data. Further arguments are given by Peto, Parish, and Gray (*32*). Peto et al. report a strong, if paradoxical, effect for age at start of dosing with NDEA (*33*); also see Drew et al. on vinyl chloride (*34*), or Ames more generally (*35*).

Collapsing the data seemed advisable to improve the asymptotics and the power (*6*). We tried models with five through nine stages, the first and next-to-last being potentially sensitive. The best model had six stages, and $\chi^2 = 67$ on 37 degrees of freedom, $p = 2/1000$.

As will be argued, even setting aside the question of whether the model fits the data, the experiment cannot really separate age from duration within the multistage framework, because the stages in the model are statistical constructs, with no biological definition.

Simulation Studies

In the present context, simulation studies (*6*) show that maximum likelihood estimates and χ^2 tests perform quite well although difficulties are created by sparse cells and positivity constraints. Parameter estimates are far from normally distributed, due to end point effects so Wald's analog of the t-test does not perform well. In the present context, the χ^2 statistic is preferred to the likelihood ratio statistic. Differencing the χ^2's to test constraints is reasonably effective and agrees with results from the score test; in general, the latter may be preferred. References on the theory are Lehmann (*36,37*), Kalbfleisch and Prentice on failure-time data (*38*), and Rao on the score test (*39*).

Discussion

With lung cancer, there is substantial conflict among the various data sets as to the sensitivity of the stages; projected risks for nonsmokers or ex-smokers are inconsistent with observations. Likewise, in the mega-mouse study, there is some difficulty in extrapolating from one part of the data set to another (high risk to low risk, or continuous exposure to ceased exposure); bladder cancer does not fit at all. Such discrepancies make it less likely that the model is correctly describing the mechanisms of carcinogenesis and tend to undercut the reliability of the model in risk assessment. However, the findings are consistent with the view of the model as a family of curves that more or less fit various data sets without necessarily capturing the underlying biological reality.

Of course, testing the multistage model on data raises questions not only about the model but also about the data. The model could be wrong, or the data, or both. For example, take the British doctors. The model predicts too many cases among the heavy smokers and the older cohort members. This may reflect a failure of the model or flaws in the data. Thus, Doll and Peto (*9*) argue that the heavy smokers have overreported their smoking habit, or in the alternative that such smokers inhale less. For older persons, diagnostics may be poorer or these persons may be more cancer-proof; the latter idea goes back to Pearl (*40*).

Other possible explanations for lack of fit in such cases include individual differences in model parameters, perhaps due to genetic variation; dependence of competing risks; relatively longer times from malignancy to death for younger cohort members; and underreporting of dose by light smokers. Changes in the composition of cigarettes over time are another complication. Later age at start of smoking among older cohort members should also be considered, as in Stevens and Moolgavkar (*41*); however, this does not fit so well with the veterans or ACS data, where controlling for age at start of smoking makes little impact on the deficit in events at old age.

Up to a point, judgment calls in fitting may be in order, especially if there is some corroborating evidence. On the other hand the data can almost always be censored or adjusted so a multistage model fits, or the model can be tuned a little to fit the remaining data. Moreover, factors that affect those portions of the data where the model seems not to fit may also affect the region where the fit

seems good, so the fit can be just as much an artifact as the lack of fit. Ultimately, censoring the data or tuning the model to the data blunt the force of empirical conclusions.

Some readers may find our approach of fitting the model and testing by chi-squared or making extrapolations and checking them too mechanical. The model does provide a rich and loosely defined class of polynomials for describing data, a heuristic for suggesting hypotheses about biological mechanisms, a demonstration that the power law for incidence rates is compatible with a series of discrete cellular changes, and a source of beautiful mathematical puzzles. If those were the only virtues attributed to the model, our critical approach might be out of line. However, quite literal and dogmatic inferences are sometimes drawn from the model, particularly in the field of risk assessment. A strict approach to testing such a model may be in order.

Other readers may be concerned, and rightly so, about the sample size issue: With a large enough sample, any model may be rejected. Our results do suggest that the multistage model will be accepted when the number of events is relatively small and rejected when the number is relatively large. On the other hand, one conventional argument for the statistical version of the multistage model is that it fits the data. While failure to fit may not prove the model to be wrong, it cannot show the model to be right. Patterns of error in the fit, discrepancies among cohorts, and systematic errors in prediction seem relevant in assessing the merits of the conventional evidence for the model.

Our view is that on the whole, fitting the multistage model to cancer incidence data in humans or in bioassays does not seem likely to yield much new understanding about the mechanisms of cancer, unless the modeling results can be rigorously checked against observable phenomena, in the lab and in human populations. Reliable procedures for estimating cancer risks seem to be a long way off, barring some breakthrough in the biological understanding. Some of the alternative models are worth exploring (*18*).

Cook, Doll, and Fellingham show that while many cancer incidence data sets fit the model, many others do not, and problems with adjustments are discussed (*8*). Doll and Peto felt that the multistage model was a promising avenue to explore "even if current knowledge is too sparse for such models to be tested critically" (*9*). Peto reviews the biological evidence (*42*); Doll and Peto cannot be described as enthusiastic about dose-response models in risk assessment (*43*). Also see Wald and Doll (*44*). Armitage says that "Until and unless we obtain direct evidence about the presence and nature of intermediate stages, any statistical theory is likely to remain largely unfalsifiable, particularly if it is allowed to be modified with the flexibility to which we have become accustomed" (*45*).

For recent somewhat critical reviews, see Freedman and Zeisel (*16*) and Kaldor and Day (*17*). On the positive side see Lave (*46*), Vouk et al. (*47*), and Zeise et al. (*48*). Proponents of risk assessment have suggested reading the Food and Drug Administration report on saccharin as an example of what can be done (*49*). For critical comment on those risk assessments, see the National Academy of Sciences report on saccharin (*50*).

Detailed Results for Lung Cancer

Introduction

This section will report details of the modeling results on the three main lung cancer data sets: the British doctors, the Dorn veterans, and the ACS volunteers. Lung cancer data are usually modeled with one early stage and one late stage allowed to be sensitive; the first and next-to-last are the conventional choices. Dose will be measured in cigarettes per day; T_0 denotes the age at start of smoking; for ex-smokers, T_1 denotes the age at quitting.

Consider the hazard rate h(t) given by a multistage model, with the following interpretation: A person who survives to age t has chance h(t)dt of contracting lung cancer in the time interval (t,t + dt). The formulas for h(t) are derived in Whittemore and Keller (*4*); Kalbfleisch et al. (*5*); and Freedman and Navidi (*6*); the relevant ones are presented here. Equations 1–3 describes an n-stage model, with the 1st and n–1st potentially sensitive. Nonsmokers are covered by Equation 1, current smokers by Equation 2, and ex-smokers by Equation 3. The dose rate is assumed constant over the period of smoking:

$$A t^{n-1} \qquad [1]$$

$$A t^{n-1} + B\,\mathrm{dose}\,(t^{n-1} - T_0{}^{n-1}) + C\,\mathrm{dose}\,(t - T_0)^{n-1} + D\,\mathrm{dose}^2 (t - T_0)^{n-1} \qquad [2]$$

$$A t^{n-1} + B\,\mathrm{dose}\,(T_1{}^{n-1} - T_0{}^{n-1}) + C\,\mathrm{dose}\,[(t - T_0)^{n-1} - (t - T_1)^{n-1}] + D\,\mathrm{dose}^2 (T_1 - T_0)^{n-1} \qquad [3]$$

In Equations 4–6, stages 1 and n are potentially sensitive; again, Equation 4 is for nonsmokers, Equation 5 is for current smokers and Equation 6 for ex-smokers. Equation 4 makes sense for $n \geq 3$; Equation 5, for $n \geq 2$.

$$A t^{n-1} \qquad [4]$$

$$A t^{n-1} + B\,\mathrm{dose}\, t^{n-1} + C\,\mathrm{dose}\,(t - T_0)^{n-1} + D\,\mathrm{dose}^2 (t - T_0)^{n-1} \qquad [5]$$

$$A t^{n-1} + C\,\mathrm{dose}\,[(t - T_0)^{n-1} - (t - T_1)^{n-1}] \qquad [6]$$

The number of events in each cell of the basic cross tab (Table 2 for the British doctors) is taken to be Poisson, and independent from cell to cell. The expected number in a cell is the hazard rate times the number of person years. The latter is treated as constant in the modeling, even though it is slightly random. This last approximation seems to be quite good in the present context (*6*); for asymptotic theory, see Aalen (*51*) and Jacobsen (*52*). The independence of competing risks is needed to compute the expected value.

After suitable aggregation of the data, the coefficients in models Equations 1–3 and Equations 4–6 can be estimated by maximum likelihood, and then the goodness-of-fit can be assessed by the χ^2 statistic:

$$\chi^2 = \Sigma(\text{obs} - \text{exp})^2/\text{exp} \qquad [7]$$

We also considered using the Neyman-Pearson likelihood ratio statistic (or Wilks statistic):

$$\text{lrs}(X) = 2\{\sup_{\Theta \varepsilon G} \log L(X|\Theta) - \sup_{\Theta \varepsilon N} \log L(X|\Theta)\} \qquad [8]$$

Here, L is the likelihood function, X the data, and θ the parameter vector, for example, the 24 Poisson means for the British doctors (Table 3). The first sup is over the set G of all parameter vectors, namely, the saturated Poisson model. The second sup is over N, the set of θ's corresponding to multistage models. Simulation results (*6*) suggest that χ^2 has close to its asymptotic chi-squared distribution, while lrs is a little too big.

The coefficients A,B,C,D in Equations 1–3 and 4–6 must be nonnegative, and satisfy the constraint AD = BC. The coefficient A reflects background rates only; B includes the sensitivity of the late stage; C, the sensitivity of the early stage; D, the sensitivity of both stages. If B = D = 0, then the late stage is insensitive; if C = D = 0, the early stage is insensitive (*6*).

If stages 1 and n are sensitive, ex-smokers show an abrupt drop in predicted risk: As t increases from just below T_1 to just above T_1, the hazard rate jumps down, because terms involving the sensitivity of that stage drop out (compare Eq. 3 with Eq. 6 at $t = T_1$). In other words, the hazard reverts to that of an n-stage model with only the first stage sensitive. This discontinuity is a well-known feature of last-stage sensitivity and is an argument against such models. The British doctors' data are too thin to reject implausible models. Equally, these data cannot provide strong evidence in favor of preferred models.

British Doctors

Tables 4 and 5 show the empirical results for the nonsmokers and current smokers among the British doctors,

Table 4. Modeling results for the British doctors.[a]

n	A	B	C	D	χ^2	Comment
3	0.21 0.09	0.0055 0.0056	0.50 0.18	0.013	101	Does not fit
4	0.41 0.17	0.011 0.011	1.46 0.49	0.038	43	Does not fit, p = 3/1000
5	0.80 0.31	0.035 0.033	2.76 1.25	0.12	21	Fits: 1st and 4th stages sensitive
5	0.77 0.31	b b	5.78 0.43	b	27	Fits, but χ^2_1 = 5.5[c]; 1st stage only sensitive
6	1.50 0.54	0.45 0.22	1.16 0.88	0.35	20	Fits: 1st and 5th stages sensitive
6	1.32 0.53	1.47 0.11	b b	b	42	Does not fit; 5th stage only sensitive
7	2.16 0.85	1.44 0.33	0.73 0.51	0.49	23	Fits; 1st and 6th stages sensitive
7	2.07 0.83	2.36 0.18	b b	b	30	Fits, but χ^2_1 = 7.5; 6th stage only sensitive
8	3.13 1.28	3.21 0.52	0.35 0.42	0.36	30	Fits; 1st and 7th stages sensitive
8	3.11 1.27	3.69 0.28	b b	b	31	Fits; 7th stage only sensitive
9	4.52 1.86	5.61 0.79	0.00 0.56	0.00	43	Does not fit
10	6.39 2.66	8.33 1.17	0.00 1.08	0.00	67	Does not fit

[a]Current smokers of cigarettes only at constant dose and nonsmokers. Dose is in cigarettes per day. There are n stages; 1 and n-1 are allowed to be sensitive. SEs are shown below estimates. Estimates are constrained to be nonnegative. Coefficients estimated as 0 are reported as 0.00. The estimates and SEs should be divided by 10^{2n+1}. The age range is 25–84, heavy smokers included. There are 21 degrees of freedom for χ^2. A variety of models fit.

[b]Estimate has been forced to 0.

[c]χ^2_1 is the difference between the χ^2s for the restricted and unrestricted models.

Table 5. Modeling results for the British doctors.[a]

n	A	B	C	D	χ^2	Comment
3	0.20 0.09	0.0046 0.0048	0.53 0.18	0.012	102	Does not fit
4	0.40 0.17	0.0010 0.0010	1.40 0.49	0.036	43	Does not fit
5	0.80 0.31	0.034 0.032	2.78 1.24	0.12	22	Fits; 1st and 5th stages sensitive
5	0.77 0.31	b b	5.78 0.43	b	27	Fits, but $\chi_1^2 = 5.5$[c]; 1st stage only sensitive
6	1.51 0.54	0.44 0.21	1.23 0.92	0.35	20	Fits; 1st and 6th stages sensitive
6	1.32 0.53	1.47 0.11	b b	b	42	Does not fit; 6th stage only sensitive
7	2.16 0.85	1.42 0.33	0.75 0.52	0.49	23	Fits; 1st and 7th stages sensitive
7	2.07 0.84	2.36 0.18	b b	b	31	Fits, but $\chi_1^2 = 7.5$; 7th stage only sensitive
8	3.13 1.28	3.20 0.52	0.36 0.42	0.37	30	Fits; 1st and 8th stages sensitive
8	3.11 1.27	3.68 0.28	b b	b	31	Fits; 7th stage only sensitive
9	4.52 1.86	5.61 0.79	0.00 0.56	0.00	43	Does not fit
10	6.39 2.66	8.33 1.17	0.00 1.08	0.00	67	Does not fit

[a]Current smokers of cigarettes only at constant dose and nonsmokers. Dose is in cigarettes per day. There are n stages; 1 and n are allowed to be sensitive. SEs are shown below estimates. Estimates are constrained to be nonnegative. Coefficients estimated as 0 are reported as 0.00. The estimates and SEs should be divided by 10^{2n+1}. The age range is 25–84, heavy smokers included. There are 21 degrees of freedom for χ^2. A variety of models fit.

[b]Estimate has been forced to 0.

[c]χ_1^2 is the difference between the χ^2's for the restricted and unrestricted models.

with Equations 1 and 2 and Equations 4 and 5, respectively; heavy smokers are included in both tables, and the age range is 25 to 84 years.

The tables are remarkably similar because the only difference between Equations 2 and 5 is in the second term. In that term, t^{n-1} dwarfs T_0^{n-1} when t is upwards of 50 or so, where all the lung cancer cases are. So the two formulas are virtually identical for current smokers. In either case, a variety of models fit the data. There are only about 200 events, and that is not enough to pin things down.

Standard errors shown in the tables are computed from the Fisher information matrix, and the usual asymptotics do not apply when estimated coefficients are close to 0. There could be a similar issue for the chi-squared test. Searching over n creates another problem. However, simulation studies (*6*) suggest that in the present context, these problems are not serious, except the distribution of some of the estimates is quite asymmetric and long-tailed. That may be the reason why in the six-stage model, constraining C to 0 makes a big difference in χ^2, although the estimate is only 1.3 times the standard error. The Wald *t*-test is not appropriate here (*6,53,54*).

Doll and Peto fit their model on the age range 40 to 79 and dose range 0 to 40. As a final test, we fit the six-stage model on this portion of the data ($\chi^2 = 12$ on 20 degrees of freedom) and use it to predict the rest of the data (ages 25–84 and any dose). The data beyond age 85 are eliminated in view of the deficit in events among the oldest persons; the data on ages 20 to 24 are eliminated too, since age at start of smoking was imputed as 22.5.

The predictions are systematically too high. In total, there are 32 events predicted and 12 observed; by simulation, $p = 3/1000$ (*6*). Even for those aged 80 to 84 and smoking 40 cigarettes a day or less the predicted is too high: 16, with 6 observed. The predicted number is also too high among those aged 39 or less and smoking 40 cigarettes a day or less, but that could easily be a matter of chance.

The model fitted by Doll and Peto to the non-smokers and current smokers is not exactly a multistage model:

$$(a + b\,\mathrm{dose})^2 (t - 22.5)^{n-1} \qquad [8a]$$

This fits into the multistage framework only by having smokers and nonsmokers alike start on the progression to malignancy at age 22.5; compare Equations 11 and 12. However, for the British doctors, the A-term in Equation 2 is rather small, and the B-term can more or less be approximated by a multiple of $(t-22.5)^{n-1}$.

Doll and Peto take $T_0 = 22.5$; they report an average age of 19 at start of smoking and add 3.5 years for latency. We redid the tables with $T_0 = 19$; the fit was worse for n = 3,4,5, but very similar for larger n.

Dorn Veterans

The modeling results for the Dorn veterans are reported in Table 6. The best model for the current smokers has six stages, but does not fit (p = 5/100,000). In this model, only the late stage is sensitive. (By comparison, the six-stage model for the doctors has both stages coming in.) With only the late stage sensitive, the model predicts almost no response to age at start of smoking; those who start early and those who start late will have nearly the same risks. In reality, of course, starting early causes a huge increase in risk (*55*).

Next, Table 7 shows the observeds and person-years in the veterans data. Table 8 shows the residuals from the six-stage model in Table 6. On the whole, the residuals seem to be negative around the edges of the table and positive in the middle. For persons aged 75 or more, some of the discrepancies may be practically significant as well as statistically significant.

ACS Volunteers

Table 9 gives results for the ACS men. The best-fitting model has six stages (p = 1%); the effect of the late stage is insignificant. When used to predict the risk for nonsmokers, this model predicts 500 ± 60, with 99 observed, p < 1/1,000,000. (This sort of cross-validation has two advantages: to some extent, it corrects for data snooping; and to some extent, it picks up heterogeneity in the data.)

The data for the ACS men are shown in Table 10. The residuals in Table 11 resemble the veterans data in pattern of signs. For all three data sets (the doctors, the veterans, and the ACS men), the best-fitting models overpredict risk around the edges. Of course, the model could be right and this pattern could be artifactual.

Many models fit the data for ACS female current smokers including those with no background rate (Table 12): extrapolating on that basis from smokers to nonsmokers shows the latter will not get lung cancer. Such models would not be making good predictions, as ACS female nonsmokers get lung cancer at fairly high rates. Even the best-fitting model predicts 100 ± 150 events, with 229 observed. The trouble is that the ACS women smokers only have 164 events. As for the British doctors there are not enough events to pin things down.

Ex-Smokers

Models fitted to current smokers do not predict well for ex-smokers, as noted above. The discussion is continued in this section. There seems to be general agreement that when smokers quit, their excess risk freezes (*22,42,55,56*). Absolute risk (background + excess) must therefore increase as a function of time since quitting. However, the data show a drop in risk on cessation of smoking.

To illustrate the predictions, Table 13 computes risks from three multistage models: the ones that best fit the nonsmokers or current smokers among the British doctors, the American veterans, and the ACS men (six stages with first and fifth allowed to be sensitive; for the ACS men, the model is fitted to current smokers only). The risks are computed for three groups of men: NON, the nonsmokers; CS, the current smokers with age at start 22.5 and constant dose of a pack a day; EX, the ex-smokers, who started smoking like CS but quit at age 50.

To compute the table, the models given by Equations 1 and 2 are fitted to the nonsmokers and current smokers to estimate the coefficients A,B,C,D. Then Equation 1 is used to compute the risk for nonsmokers, 2 is used for the current smokers, and 3 is used to project the risk for ex-smokers.

One problem is that the projections of the three models are quite discrepant. For an extreme example, take the

Table 6. Modeling results for the Dorn veterans.[a]

n	A	B	C	D	χ^2	Comment
3	0.56 0.042	0.00 0.0016	0.58 0.05	0.00	130	Does not fit
4	0.83 0.06	0.00 0.002	1.27 0.11	0.00	69	Does not fit
5	1.24 0.09	0.00 0.003	2.66 0.23	0.00	58	Does not fit
6	1.75 0.13	1.01 0.072	0.00 0.02	0.00	50	Best fit, p = 5/100,000 5th stage only sensitive
7	2.51 0.19	1.50 0.11	0.00 0.04	0.00	70	Does not fit
8	3.55 0.27	2.21 0.16	0.00 0.07	0.00	118	Does not fit
9	4.99 0.38	3.21 0.23	0.00 0.14	0.00	196	Does not fit
10	6.96 0.53	4.64 0.32	0.00 0.24	0.00	309	Does not fit

[a]Current smokers of cigarettes only and nonsmokers. Dose is in cigarettes per day. There are n stages; 1 and n-1 are allowed to be sensitive. SEs are shown below the estimates. Estimates are constrained to be nonnegative. Coefficients estimated as 0 are reported as 0.00. The estimates and SEs should be divided by 10^{2n+1}. There are 17 degrees of freedom for χ^2. No model fits.

Table 7. The Dorn veterans: Events/person-years, current smokers of cigarettes only and nonsmokers.

	Dose, cigarettes per day				
Age	0	1-9	10-20	21-39	40+
Below 55	2/84,401	0/9,366	19/72,002	23/46,532	3/4,728
55-64	25/212,448	18/16,417	111/67,593	129/48,307	37/9,860
65-74	119/363,152	25/27,460	294/94,061	258/53,278	72/10,072
75 or more	27/75,499	2/5,164	63/13,959	31/5,595	8/906

Table 8. Residuals for the Dorn veterans, from the best-fitting model in Table 6.[a]

	Residuals O/E Dose, cigarettes per day				
Age	0	1-9	10-20	21-39	40+
Below 55	−1	−1	−3	−5	−2
55-64	−6	9	19	5	−4
65-74	16	−5	44	−8	−9
75 or more	−12	−8	−4	−19	−5

[a]The model overpredicts around the edges and underpredicts in the middle, for current smokers of cigarettes only and nonsmokers.

Table 9. Results for the ACS male current smokers.[a]

n	A	B	C	D	χ^2	Comment
3	1.58 0.31	0.00 0.008	0.25 0.07	0.00	315	Does not fit
4	2.54 0.48	0.00 0.01	0.64 0.16	0.00	127	Does not fit
5	4.51 0.75	0.00 0.02	1.45 0.35	0.00	38	Does not fit
6	9.02 1.07	0.028 0.047	2.52 0.73	0.0077	28	p = 1%; 1st and 5th stages sensitive
6	8.59 1.01	b b	3.00 0.23	b	28	p = 1%; 1st stage only sensitive
7	11.15 1.82	0.99 0.16	0.00 0.22	0.00	30	p = 0.5 of 1%; 6th stage only sensitive
8	14.91 2.72	1.70 0.23	0.00 0.33	0.00	68	Does not fit
9	19.09 4.00	2.84 0.34	0.00 0.51	0.00	158	Does not fit
10	23.29 5.77	4.63 0.50	0.00 0.72	0.00	305	Does not fit

[a]Nonsmokers are excluded. There are n stages; 1 and n-1 are allowed to be sensitive. Dose is in cigarettes per day. SEs are shown below the estimates. Estimates are constrained to be nonnegative. Coefficients estimated as 0 are reported as 0.00. The estimates and SEs should be divided by 10^{2n+1}. There are 13 degrees of freedom for χ^2. The six-stage model barely fits.
[b]Estimate has been forced to 0.

Table 10. ACS men: Events/person-years, current smokers only.

	Dose, cigarettes per day			
Age[a]	1-9	10-19	20-39	40+
Below 55	24/62,581	62/127,178	355/445,541	98/96,678
55-64	38/32,915	116/58,558	403/147,636	102/25,885
65-74	23/13,249	76/19,981	189/32,233	24/3,516
75 or more	6/2,211	11/2,333	14/2,375	1/178

[a]Age is at the beginning of the study, in 1959.

saving in relative risk at age 75 from quitting at age 50, namely, 1−(EX/CS). From the model fitted to the doctors or the veterans, this is estimated as 80%. But from the ACS men, the estimate is only 7%.

The main point is that the predictions seem to be qualitatively wrong. The predicted excess risk (EX − NON) increases steadily for the doctors and the ACS men, rather than freezing: the reason is that in the model, the first stage is sensitive, so the C-term in Equation 3 is positive and increasing. The excess risk is predicted as constant for the veteran ex-smokers because only the late stage is sensitive, so the excess risk is the B-term in Equation 3. The predicted absolute risk for ex-smokers increases rapidly with time since quit for all three models, as in Table 13.

Table 14 shows the observed and expected number of lung cancer cases among the veteran ex-smokers. The expected are from the six-stage model of Table 6: Equations 1 and 2 are fitted to the nonsmokers and current smokers, then Equation 3 is used to predict the risk for the ex-smokers. The model underpredicts in the early years and then overpredicts. (Of course, part of the reason for ex-

Table 11. Residuals for the ACS men from the six-stage model with first and fifth stages sensitive in Table 9.[a]

	Residuals O/E Dose, cigarettes per day			
Age	1–9	10–19	20–39	40+
Below 55	−1	−6	16	−12
55–64	−0	17	29	5
65–74	−10	1	−1	−8
75 or more	−5	−7	−16	−3

[a]The model overpredicts around the edges and underpredicts in the middle for current smokers.

Table 12. Results for the ACS female current smokers.[a]

n	A	B	C	D	χ^2	Comment
3	0.14 0.13	0.032 0.025	0.0027 0.0095	0.00062	16	Fits
3	0.098 0.086	b b	0.10 0.014	b	15	Fits; 1st stage only sensitive[c]
3	b b	b b	0.11 0.0089	b	17	Fits; 1st stage only and no background
4	0.24 0.24	0.059 0.039	0.0064 0.024	0.0016	8	Fits; 1st and 3rd stage sensitive
4	0.19 0.16	b b	0.30 0.043	b	8	Fits; 1st stage only sensitive
4	b b	b b	0.35 0.027	b	10	Fits; 1st stage only and no background
5	0.34 0.43	0.12 0.062	0.0079 0.040	0.0027	6	Fits; 1st and 4th stages sensitive
5	0.26 0.30	0.13 0.019	b b	b	6	Fits; 4th stage only sensitive
5	b b	0.15 0.011	b b	b	7	Fits; 4th stage only and no background
6	0.43 0.75	0.24 0.10	0.0018 0.048	0.00097	8	Fits; 1st and 5th stages sensitive
6	0.41 0.53	0.24 0.033	b b	b	8	Fits; 5th stage only sensitive
6	b b	0.26 0.020	b b	b	9	Fits; 5th stage only and no background
7	0.61 1.27	0.43 0.17	0.00 0.09	0.00	15	Fits; 6th stage only sensitive
7	b b	0.46 0.036	b b	b	15	Fits; 6th stage only and no background
8	0.84 2.11	0.76 0.27	0.00 0.17	0.00	25	p = 2%
9	1.03 3.41	1.31 0.43	0.00 0.28	0.00	40	Does not fit
10	1.00 5.37	2.22 0.67	0.00 0.36	0.00	60	Does not fit

[a]Nonsmokers are excluded. There are n stages; 1 and n-1 are allowed to be sensitive. Dose is in cigarettes per day. SEs are shown below the estimates. Estimates are constrained to be nonnegative. Coefficients estimated as 0 are reported as 0.00. The estimates and SEs should be divided by 10^{2n+1}. There are 13 degrees of freedom for χ^2. A variety of models fit, including quite implausible ones. In these data, either B or C and A can be constrained to 0; not all possibilities are reported.

[b]Estimate has been forced to 0.

[c]The fit is by maximum likelihood, not minimum χ^2; occasionally, constraints will improve the χ^2.

Table 13. Predicted lung cancer deaths per 100,000 person years for nonsmokers, current smokers, and ex-smokers.[a]

	Doctors			Veterans			ACS men		
Age	NON	CS	EX	NON	CS	EX	NON	CS	EX
50	5	58	58	6	67	67	28	38	38
55	8	112	66	9	109	71	45	68	66
60	12	202	78	14	170	76	70	114	110
65	17	348	99	20	254	82	105	185	176
70	25	571	130	30	368	92	152	290	274
75	36	900	175	42	520	104	214	441	412

[a]Age at start of smoking is 22.5 years, and dose is a pack per day; the ex-smokers quit at age 50. The risks are predicted from multistage models fitted separately to three cohorts. The models are inconsistent with each other, and the results are incompatible with the freezing of excess risk on quitting. Non, nonsmokers; CS, current smokers; Ex, ex-smokers.

Table 14. Veterans ex-smokers: Observed lung cancer deaths and numbers expected from the six-stage model in Table 6.[a]

Years since quit[b]	Person-years	Observed	Observed per 100,000 person-years	Expected	O/E
0–4	15,693	26	166	20.4	1.27
5–9	33,633	45	134	41.1	1.09
10–14	41,786	52	124	46.9	1.11
15–19	35,008	25	71	33.1	0.76
20–24	27,878	11	39	17.4	0.63
25–29	21,844	6	27	10.8	0.55
30–34	13,426	4	30	6.6	0.61
35+	4,212	0	0	2.1	0.00

[a]The observed number of cases per 100,000 person years declines steadily, and so does the ratio of observed to expected. Persons who quit due to doctor's orders are excluded from the data.

[b]Years since quit at the beginning of the study is given by 5-year intervals; this is replaced by a truncated midpoint in the calculation. For example, a person who quit 5–9 years before 1954 is assumed to have quit in 1947.

cess events in Table 14 may be that sick people quit smoking.) Only the late stage in the model is sensitive; making the last stage sensitive instead of the next-to-last would barely affect Table 6; it would make the underprediction problem in Table 14 even worse but would partially correct the overprediction.

As far as we can tell, the excess risk in fact declines with years since quitting rather than freezing; even the absolute risk (background + excess) declines, quite contrary to the predictions of the model. Table 14 shows the decline for the veterans. These data are sparse, so cross-tabulation to control survivor bias does not seem advisable; however, controlling for the term dose $\times$ ($T_1^{n-1} - T_0^{n-1}$) in Equation 3, as an indicator of risk, does not change the picture very much. Nor does indirect standardization on the risk at time of quitting. The spike in risk at the time of quitting, however, is noticeable.

For the ACS men too, Table 15 shows that the absolute risk (events per 100,000 person years) declines steadily as a function of time since quitting. The first line in Table 15 may be an artifact (sick people quit smoking). The last line may be low due to the missing events for older persons. Even between lines 2 and 3, there may be a survivor bias: The men most at risk die early. However, controlling for age at quitting and dose (by cross-tabulation) makes little difference, so survivor bias does not seem to be a big problem.

The absolute risk does seem to drop with time since quitting for the veterans and ACS males; the rapid increase predicted by the model simply is not there. To state the point more sharply, constant excess risk is incompatible with the sensitivity of the first stage, needed so that age at start influences the response; decreasing excess risk is incompatible with any of the models fitted here. (The phenomenon can be incorporated by having random parameters or a long and variable latency period between malignancy and death.)

For the British doctors, data on ex-smokers are not available. However, the risk for ex-smokers seems to be less than their risk at time of quitting, until 20 years after quitting; the numbers are small, but in aggregate, the observed number of events for the ex-smokers is less than predicted from the risk at time of quitting (*57*). On the whole, our results are consistent with this finding [for other data and reviews, see (*58–61*)].

For a literature review on lung cancer, see the IARC monograph (*55*). With bladder cancer, the risk is considered to drop when exposure ceases (*55*). For experimental results on regression of lesions when carcinogenic in-

Table 15. Observed absolute risk for ACS male ex-smokers (number of lung cancer deaths per 100,000 person years) by years since quit smoking.

Years since quit[a]	Person-years	Number of events	Rate per 100,000 person-years[b]
<1	42,053	69	164
1–4	97,469	111	114
5–9	201,655	108	54
10+	66,566	6	9

[a]Years since quit is at the beginning of the study, in 1959.

[b]These are crude rates; standardization does not change the pattern (*54*).

sult stops, see McKenzie and Rous (*62*); Hennings et al. (*63*); and Farber and Sarma (*64*); in the other direction, see Littlefield et al. (*28*). There is further discussion of ex-smokers in Freedman and Navidi (*65*) and Gaffney and Altshuler (*66*). Gaffney and Altshuler focus on the British doctors. They find declining excess risk after cessation of smoking, and note the inconsistency with the Armitage-Doll model. They present an alternative model.

The tension between models for continuing smokers and ex-smokers seems to be well known. The resolution attempted in Brown and Chu (*22*) is not satisfying. The multistage model in that paper is fitted not to data but to output from logistic regressions, which are themselves inconsistent with the multistage model; the parameters of the fitted multistage model are allowed to depend on dose; age and duration of smoking are treated as extra parameters, constant across subjects, and estimated, even though the data are available.

Some Technical Issues

Latency

After a cell (or cluster of cells) becomes malignant, some period of time must elapse until the cancer becomes clinically detectable; and another period of time until death ensues. These periods of time are the latencies. The first waiting time is not empirically observable, almost by definition; so there is little direct evidence about its distribution: indirect evidence suggests this time may be appreciable (*67,68*). (Of course, other evidence might suggest this time is short.) The second waiting time has been studied for many human cancers; for lung cancer, it may be on the order of 18 months.

Latency complicates the modeling problem even further. Some authors treat latency as constant across subjects, to be estimated statistically along with the other parameters in the multistage model; others treat the latency as following some textbook distribution (like the Weibull), whose parameters are then estimated. Such assumptions are hard to defend empirically.

Unless noted otherwise, we cut the knot by setting latency to zero. This has the advantage of simplicity, but cannot be taken too literally. The problem is serious, because in the end the data are on times to a clinically detectable end point. If a large part of the distribution for that time is left unspecified, the model is poorly defined.

As a practical matter, allowing positive latency reduces the number of estimated stages, by increasing the rate of change of the fitted hazard function at the relevant time period.

Independence of Competing Risks

Let τ be the time to failure for the whole tissue in the multistage model, namely, the time for the first target cell to complete its progression through the n stages of the process. Then the conditional distribution of τ given $\tau > t$, is assumed equal to the distribution of τ given survival on test to time t. (In the latter event, we condition not only that $\tau > t$, but that all risks mature after time t.) The assumption of equality is a version of independence of competing risks, which allows the model to be used even when the data on waiting times are censored by death from other causes; and in the case of human subjects, by withdrawal from the study, data selection by the investigators, etc.

This assumption may not be verifiable from the data (*69*), and we can see only two possible defenses: a) It has been used since the time of Bernoulli; b) it is at present impossible to do risk modeling any other way.

Pooling

For estimation and testing, it is necessary to arrive at some definite aggregation of the data, which in our examples is usually presented in the form of a two- or three-dimensional cross-tab. It will be advantageous to pool cells, eliminating the sparse ones. This improves power (up to a point) and makes the null distribution of the χ^2 statistic closer to the asymptotic limit; for some empirical evidence, see Freedman and Navidi (*6*).

To make the asymptotics of the χ^2 test go through, the same aggregation must be used for both estimation and testing. For the lung cancer data sets, the aggregation was suggested by the age × dose table in Frome and Checkoway (*70*). The same four age groups were used for all cohorts: a) 54 or less; b) 55 to 64; c) 65 to 74; and d) 75 or more.

For the British doctors, we chose six dose groups (cigarettes per day): a) 0; b) 1 to 9; c) 10 to 19; d) 20 to 29; e) 30 to 40; and f) 41 or more. For the veterans and the ACS volunteers, there were five dose groups as in the raw data: a) 0; b) 1 to 9; c) 10 to 20; d) 21 to 39; and e) 40 or more. (For the volunteers, group c was 10–19 and group d was 20–39. Occasional smokers in the veterans cohort were excluded.)

For the British doctors, there are $4 \times 6 = 24$ cells in the aggregated cross-tab (Table 3), 4 parameters A,B,C,D, and 1 constraint (AD = BC) so the χ^2's have $24 - 4 + 1 = 21$ degrees of freedom. The veterans have 4 cells fewer and 17 degrees of freedom in the χ^2's. The ACS data has no nonsmokers and 13 degrees of freedom.

The next object is to explain how aggregate cross-tabs are derived from the raw data, with the current smokers in the British doctors as a first example. The original data are given in Table 2 and the aggregation in Table 3. To illustrate the arithmetic, the data for ages 55 to 64 and dose 30 to 40 are reproduced from Table 2, as Table 16.

In Table 3, there is a cell corresponding to ages 55 to 64 and dose 30 to 40. The number of events in that cell is obtained by adding the numbers of events in the basic

Table 16.

	Dose, cigarettes/day	
Age, years	30–34	35–40
55–59	6/1923	4/1063
60–64	11/1362	7/826

cells above: 6 + 4 + 11 + 7 = 28, and likewise for the person-years.

The second example is the veterans' data. How do we get from the raw data on the tape to the aggregate data in Table 7? The raw data can be used to make a three-dimensional cross-tab of basic cells:

age × age at start of smoking × dose

For nonsmokers, there is only one dimension of interest: age. For ex-smokers, there are four dimensions: age, age at start of smoking, age at quitting, dose. The tape reports year of birth and death; age at start or quit is only reported by 5-year groups.

For instance, one basic cell in the cross-tab corresponds to current age 55 to 59, age at start 20 to 24, and smokes 10 to 20 cigarettes per day. Until death, subjects may contribute person years to each basic cell in the three-dimensional cross-tab and an event if they die of lung cancer.

Each age × dose group in the aggregate cross-tab is the result of pooling over a set of basic cells. For example, consider the following cell in the aggregate cross-tab: current age 55 to 64, smoking 10 to 20 cigarettes per day. Person years, observeds and expecteds for that cell are obtained by adding up the numbers for the basic cells in the three-dimensional cross-tab, corresponding to current age 55 to 59 or 60 to 64; age at start of smoking 5 to 9, or 10 to 14, . . . , or 50 to 54; dose 10 to 20. (We took age at start "less than 10" as 5–9; likewise, "50 years or older" as 50–54.)

For the ACS volunteers, the data provided to us were already in the form of a cross-tab, with age and years since quit for ex-smokers at baseline in 1959. The study period was 1960 to 1965, so we added 4 years to get current age. For ex-smokers, age at start was not collected.

By convention, the number of person-years in each basic cell is treated as a constant in the modeling. From cell to cell, the numbers of events are taken as independent.

The expected number of events in a basic cell equals the number of person-years times the hazard rate for that cell. That is where independence of competing risks comes in. For example, if persons well along the progression toward lung cancer were more susceptible to heart attacks, survival to age t would change the hazard; however, by the independence assumption, the hazard does not change.

When computing the hazard function in Equations *1–6*, age and dose are taken as truncated mid-points; for example, age 55 to 59 becomes 57, age at start 20 to 24 becomes 22, and dose 10 to 19 becomes 15. A dose of 40 or more is taken a bit arbitrarily, as 50.

Our aggregation was chosen to avoid sparse cells and to treat all three cohorts in a similar way. However, the procedure puts little emphasis on the response to T_0, which is projected out. (To maximize the likelihood, only the sum of the expecteds needs to be approximately right, not their distribution over T_0.) If the model is right, the response to T_0 can be inferred from the response to age and dose. For the veterans, the model does not do this at all well (Table 17). Our aggregation may be criticized as leading to inefficient procedures; however, they are valid, and they evidently provide efficient-enough tests.

Table 17. Observed and expected numbers of events for the veterans by age at start of smoking.

Age at start	Person-years	Observed	Expected	O/E
Less than 10	2,839	11	7	1.6
10–14	32,929	114	79	1.4
15–19	231,146	562	488	1.2
20–24	145,376	308	334	0.9
25–29	37,312	63	96	0.7
30–34	16,070	22	43	0.5
35–39	8,138	6	22	0.3
40 or more	11,490	7	25	0.3

Models for Animal Data

Introduction

For experimental data on animals, doses are set high, so the hazard is high, too. The Poisson distribution must be replaced by the binomial and hazards converted to probabilities. The process will be illustrated on the mega-mouse data (*27*). and the Peto et al. results on skin cancer and aging (*10*) will be considered last.

Mathematical Preliminaries

Let h be the hazard in a multistage model. With the independence of competing risks, the probability of an animal getting cancer during the period t to t + s, conditional on having survived till the beginning of the period, is

$$1 - \exp\left\{-\int_t^{t+s} h(u)\,du\right\} \approx 1 - \exp\{-h(t)s\} \qquad [9]$$

For the current-exposure group in the mega-mouse study, exposure starts at $T_0 = 0$, and the formulas for h become simpler. Consider a model with n precancerous stages, of which m are potentially sensitive. Suppose $T_0 = 0$, and exposure is continuous. Let N be the number of target cells. Assume the rate of progression through stage i is $a_i + b_i d$, when the dose rate is d. The a's and b's must be nonnegative; stage i is potentially sensitive if b_i is allowed to be positive; the number of such stages is denoted m. The hazard rate at time t is essentially as follows (*71*):

$$N\left[\prod_{i=1}^{m}(a_i+b_i d)\right]\frac{t^{n-1}}{(n-1)!} \qquad [10]$$

Multiplying t^{n-1} in Equation 10, there is a polynomial in dose of degree m, with nonpositive roots; and only its m + 1 coefficients can be estimated. In general, data on the current-exposure group cannot determine which of the stages are the sensitive ones, although their number m can be estimated, and only certain products in the basic parameters N, a_i, b_i can be estimated. Identifiable parameters can be obtained by rewriting Equation 10 as

$$h(t) = (c_0 + c_1 d + \cdots + c_m d^m)\, t^{n-1} \qquad [11]$$

where d is dose and t is time. The c's are estimable; they should be nonnegative, and the polynomial should have m nonpositive roots—a constraint which is hard to impose.

To capture the constraints, it is possible to factor the dose polynomial, rewriting the hazard rate (Eq. 11) as

$$G\left[\prod_{j=1}^{m} (1 + r_j d)\right] t^{n-1} \qquad [12]$$

The lead constant G is identifiable; and so are the r_j's, if arranged in decreasing order. These r's must be nonnegative. For Equation 12 to make sense, the background rates (the a's in Eq. 10) must be positive.

With ceased exposure, it matters which states are sensitive. To model the mega-mouse data, we wanted to consider having two early stages sensitive. The calculus gets out of hand amazingly fast, but a direct computation is still feasible when the first and second stages are sensitive. If exposure starts at time $T_0 = 0$ and ends at time T_1, the hazard at a later time t is

$$\begin{aligned} &A t^6 \\ &+ B\,\text{dose}\,[t^6 - (t - T_1)^6 - 6T_1(t - T_1)^5] \\ &\qquad + C\,\text{dose}\,[t^6 - (t - T_1)^6] \\ &+ D\,\text{dose}^2\,[t^6 - (t - T_1)^6 - 6T_1(t - T_1)^5] \end{aligned} \qquad [13]$$

Here, A represents background; B, the sensitivity of the second stage; C, of the first stage; and D, both (*4,6*).

Mega-Mouse Study

The mega-mouse experiment did not focus on time-to-tumor, so the results do not fit naturally into the framework of the multistage model. To see the problem more sharply, take for example the 336 mice assigned to a dose group of 150 ppm with planned sacrifice at 24 months. Of these, 130 survived to 24 months and were sacrificed at that time; among the sacrificed animals, 100 had bladder tumors. The ratio 100/130 represents prevalence, not incidence. Indeed, it is not known when these tumors developed.

To model this data set, we entertained two polar assumptions: a) The counts represent incidence, that is, tumors which arose during the month of sacrifice; b) tumors are rarely fatal, so the counts represent nearly all the tumors that arose at or before the sacrifice time. Either assumption gives about the same likelihood function for the current exposure group, as will now be discussed.

The statistical analysis is performed by treating the number of survivors as constant. The counts in the various cells are taken to be independent binomials; the number of trials is the number at risk in the cell. If assumption *a* holds, the probability of a mouse getting liver cancer in month t is given by Equation 9, with s = 1 (the period is viewed as 1 month). If assumption *b* holds, the probability that an animal sacrificed at time t has cancer is

$$P\{\tau < t\} = 1 - \exp\left\{-\int_0^t h(u)\,du\right\} \qquad [14]$$

In effect, Equation 14 just increases the number of stages by 1. More specifically, if h is the hazard rate for, say, a six-stage model with the first and fifth sensitive, then $\int_0^t h(u)\,du$ is the hazard rate for a seven-stage model, again with stages one and five sensitive. This follows from the usual inductive construction of h (*4,6*). Thus, assumptions *a* and *b* turn out to differ only in the estimated number of stages (for the current exposure group anyway). We proceed, somewhat hesitantly, on the basis of assumption *a*.

Data on the mice who died before planned sacrifice are available on tape, so there is an opportunity for cross-checking; also see Farmer et al. (*29*). Combining data from sacrifice and spontaneous deaths would require explicit modeling of latency, which has been estimated for bladder and liver cancer as being about 6 months (*72*). Such modeling requires introducing further assumptions, which seem as drastic in their own way as assumption *a*. For other views, see Kalbfleisch et al. (*5*) or Malani and van Ryzin (*73*).

For the current-exposure group in the mega-mouse experiment, the data are reported in a basic two-dimensional cross-tab for dose × sacrifice time (*28*). With the ceased-exposure group there is a third dimension, namely, the time at which exposure ended. The cell counts in the basic cross-tab are taken to be independent binomial variables; in each cell, the event probability is given by Equation 9 with s = 1, and the number of trials is the number of sacrificed animals.

To stabilize the χ^2, we wanted to avoid sparse cells; nor was a sum of binomials attractive. Therefore, some dose × sacrifice groups with low dose or early sacrifice were eliminated from the fitting. To some extent this choice was data-driven, but it was treated as deterministic in the statistical analysis. The impact of this move seems to be small (*6*). Of course, in Equation 7 for χ^2, the denominators are the binomial variances.

Bladder Tumors. The mega-mouse experiment had two end points, cancers of the bladder and the liver (transitional cell carcinomas of the urinary bladder and primary hepatocellular carcinomas). The bladder data are shown in Table 18.

There are a substantial number of responses only at high dose and after long exposure and the hazard increases more rapidly in dose than in time. Carlborg demonstrated this by fitting a Weibull hazard of the form $cd^a t^b$, with $a > b + 1$ (*30,31*). Just by way of illustration, consider the high risk current-exposure groups inside the marked region of Table 18. Fitting the Weibull hazard

Table 18. Results on bladder cancer from the mega-mouse study: Number of response/number sacrificed (*12,13*).

Dose, ppm	Sacrifice time, months								
	9	12	14	15	16	17	18	24	33
0	0/142	0/140	0/113	0/88	0/183	0/127	1/400	1/384	0/24
30							4/1573	0/900	1/92
35						1/389	1/796	2/638	0/45
45					1/271	1/264	1/383	1/445	0/12
60	0/279	2/268	0/224	1/181	0/265	0/206	3/269	3/415	1/11
75	0/139	0/137	0/110	1/94	0/175	0/134	1/267	3/311	4/12
100	1/142	1/138	0/117	0/89	1/90	4/67	5/131	25/160[a]	8/10[a]
150	0/140	9/141	21/114[a]	28/90[a]	36/85[a]	36/65[a]	62/121[a]	100/130[a]	

[a]Cell used in fitting.

there gives a = 5 and b = 3, roughly. If the rate of increase in dose is faster than the rate in time, Equation 10 cannot hold. On this basis, Carlborg prefers the Weibull to the multistage. However, the values for a and b are quite sensitive to the cells used in fitting. Furthermore, the Weibull did not fit even the handful of cells marked in Table 18, which represents a third cut at fitting progressively smaller portions of the data.

We also tried on the high-risk cells four- or five-stage models with all stages sensitive: a = 4 and b = 3, or a = 5 and b = 4. The former is better, but does not fit even the selected cells, having $\chi^2 = 22$ on 7 degrees of freedom. Also, the multistage model predicts 197 cancers in the censored cells (outside the marked region) with 55 observed; the Weibull model does about the same. In short, neither the Weibull nor the multistage fits.

Brown and Hoel (*12*) say a hazard is "factorable" if, like the multistage hazard in Equation 10, it is a function of dose × a function of time. They argue that no factorable hazard function will fit the whole data set. Transforming the dose scale does not affect factorability and therefore will not make the model fit. The sparseness of most of the cells in the table may not affect their analysis too much (*6*).

Liver Tumors. For liver tumors too, the current-exposure group seemed the natural starting point. To stabilize the behavior of the χ^2 statistic, it seemed advantageous to develop the models only on part of the data set, censoring the sparse cells with low dose or early sacrifice. (For each dose group, we started with the longest exposures, and cut back until cells with two events or fewer were encountered.)

The models had grotesque chi-squareds, and the residuals showed most of the problem to come from a dose of 60 ppm and sacrifice at 24 months. After looking more closely at the data source (*28*) we convinced ourselves that there was a misprint in that cell, which reports 7/415 events, and an incidence rate of 17.1%. The rate looks plausible, and we changed the numerator to 0.171 × 415 = 71, agreeing with (*13*).

Table 19 shows the corrected data; cells used in fitting are inside the marked region; 27/621 events are censored (outside the region). We went ahead on the corrected data, fitting models with hazard rates defined by Equation 11 without the constraint of nonpositive roots, or by Equation 12 with all constraints imposed. The constraints in Equation 11 are seldom imposed, so it is worth considering what happens without them. To illustrate, suppose n = 7 and m = 2 so there are seven stages of which two are sensitive. Equation 11 gives the model

$$155 + 588d + 896d^2, \quad \chi^2 = 26.3, \quad lrs = 26.2 \qquad [14a]$$

$$142(1 + 2.38d)^2, \quad \chi^2 = 26.7, \quad lrs = 26.5 \qquad [14b]$$

lrs is the Neyman-Pearson likelihood ratio statistic of Equation 8, while χ^2 is defined by Equation 7 with the binomial variances in the denominators. Equation 11 fits a little better but the polynomial has imaginary roots. The double root from Equation 12 seems unlikely; however, the discriminant of the quadratic has been prevented from going negative and an end point maximum occurs when the latter vanishes. In sum, fitting the polynomial without constraints does not lead to a proper multistage model.

Table 20 reports the results of fitting multistage models of the form of Equation 12 to the liver tumor data in the

Table 19. Results on liver cancer from the mega-mouse study: Number of response/number sacrificed (*12,13,27*).

Dose, ppm	Sacrifice time, months								
	9	12	14	15	16	17	18	24	33
0	0/142	0/140	0/113	0/88	1/183	0/128	1/401	9/383[a]	8/23[a]
30							17/1573[a]	55/900[a]	44/92[a]
35						2/389	7/792[a]	55/639[a]	20/45[a]
45					2/272	5/264[a]	7/383[a]	57/445[a]	5/12[a]
60	0/279	2/268	0/224	1/182	4/265[a]	6/206[a]	7/268[a]	71/415[a]	7/11[a]
75	0/139	0/137	0/110	1/94	5/174[a]	5/134[a]	6/267[a]	62/311[a]	8/12[a]
100	1/142	3/138	1/117	4/90	1/90	1/67	6/131[a]	47/160[a]	8/10[a]
150	0/140	1/141	1/114	1/90	3/86	0/65	7/121[a]	56/130[a]	

[a]Cell used in fitting.

Table 20. Results for the mega-mouse study: Continuous exposure to 2-AAF and resulting liver tumors.[a]

n	m	df	χ^2	G	r_1	r_2	r_3	Comment
6	1	25	72	27	11			Does not fit
6	2	24	53	37	2.2	2.2		Does not fit
6	3	23	53	37	2.2	2.2	0.00	Does not fit
7	1	25	48	104	12			$p < 5/1000$
7	2	24	27	142	2.4	2.4		Fits, $p = 30\%$
7	3	23	27	142	2.4	2.4	0.00	Fits, $p = 26\%$
8	1	25	64	386	14			Does not fit
8	2	24	39	518	2.6	2.6		Marginal fit, $p = 3\%$
8	3	23	39	518	2.6	2.6	0.00	Marginal fit, $p = 2\%$
9	1	25	133	1391	15			Does not fit
9	2	24	109	1789	3.0	3.0		Does not fit
9	3	23	109	1815	2.9	2.9	0.02	Does not fit r_4
9	4	22	109	1815	2.9	2.9	0.02	Does not fit $r_4 = 0$

[a]A seven-stage model fits the data, with two stages sensitive. The parameters express the hazard function (Eq. 12) as a function of d = (dose in ppm) /100 and t = (time in months) /100. Parameters estimated as 0 are recorded as 0.00.

current exposure group; reported cancers are treated as incident. The Fisher information matrix usually cannot be inverted for these models, so standard errors are not reported. For a simulation study on the accuracy of the MLE's and the minor impact of censoring, see Freedman and Navidi (*6*).

The data can be fitted by a multistage model with seven stages, of which two are sensitive; we saw no pattern in the residuals. Since exposure starts at birth, the data on the current-exposure group cannot determine which two stages out of the seven are the sensitive ones. There are ${}^7C_2 = 21$ models to explore, and it was easiest to begin with the most familiar: Equations 1, 2, and 3 with the first and sixth stages sensitive. This is connected with Equation 12 as follows:

$$A = G, \quad B = Gr_1, \quad C = Gr_2, \quad D = Gr_1 r_2 \qquad [15]$$

$$A = G, \quad B = Gr_2, \quad C = Gr_1, \quad D = Gr_1 r_2 \qquad [16]$$

In the present application, r_1 and r_2 are estimated as equal, so the two models coincide. On the ceased-exposure group, this model gave $\chi^2 = 100$ on 24 degrees of freedom. Examination of the residuals showed that most were positive; predicted risk was too low after exposure ended. This suggested trying a model with two early stages sensitive, rather than one early and one late: Equation 13.

The coefficients in Equation 13 can be obtained as above from Equations 15 and 16, giving A = 142, B = C = 338, and D = 803. This time, the model fits the ceased-exposure group: $\chi^2 = 30$ on 24 degrees of freedom ($p = 18\%$), although the residuals suggest that predicted risk is too low at longer sacrifice times. (Since the choice of model was data-driven, the p-values are no longer protected by cross-validation, and are biased toward accepting the model.) Almost as an afterthought, we went back to the omitted cells with low dose or early exposure (Table 19). There were 27 events in these cells, with a predicted 41. This looks bad, but the censored observations tend to be the smaller ones. The bootstrap assigns a p-value of 14% to a test based on the statistic "predicted-observed" for the censored cells (*6*).

We then looked at the eight-stage model with first and second stages sensitive. This was marginal on the main group of cells, and looked fine on the censored ones (26 predicted, 27 observed, although the bias is still there). The model did not fit the ceased-exposure group ($\chi^2 = 41$ on 24 degrees of freedom). Again, the residuals were too positive.

Brown and Hoel

Brown and Hoel (*13*) fit a multistage model to the liver data. We fully agree with their conclusion:

> The way in which dose is represented in the model may be very consequential, and [this] illustrates the basic difficulties one may encounter when attempting to conclude *with confidence* anything about the initiation/promotion mechanisms based on tumor count data.

In general terms, our results are consistent with theirs, but there are some points of disagreement: we like seven stages with two sensitive, they like six stages with one sensitive, or four with two sensitive. They seem to be following assumption *b* and Equation 14, so seven of our stages correspond to six of theirs; however, we cannot fit a seven-stage model with one sensitive, or a five-stage model with two sensitive.

One reason for the discrepancies seems to be Brown and Hoel's decision to eliminate the group sacrificed at 33 months, based on examination of residual plots; the plots may show the heterogeneity we picked up in cross-validation. Another reason is the choice of functional form: they include a constant latency parameter, make a nonstandard adjustment for background, and have a dose threshold effect.

We are using $\chi^2 = \Sigma\,(\text{obs-exp})^2/\text{var}$, Brown and Hoel are using lrs. However, by our reckoning, fitting the seven-stage model with two stages sensitive on all 56 cells gives lrs = 64 or $\chi^2 = 65$, with 53 degrees of freedom, so this may not matter. On the interpretation in the presence of sparse cells, see Freedman and Navidi (*6*). The estimated

r's from fitting to all 56 cells were quite different from those in Table 20, being 3.32 and 1.56, respectively; G was about the same, at 141.

Peto Mice

Peto et al, report on an experiment with four treatment groups, benzpyrene being applied starting at 10 weeks, 25 weeks, 40 weeks, and 55 weeks (*10*). Peto et al. chart the tumors fortnightly, and Table 21 reproduces some of the data from their Appendix for weeks of benzpyrene 40 through 58.

As we understand it, the first line in the table reports the number of tumors that developed during the weeks 39 and 40 of treatment and were charted at week 40; this line also reports the corresponding number of animals at risk. In group 1, there were no tumors and 130 animals at risk, and so forth. We did not use data prior to charting at week 40 or after week 90 when the cells get very sparse.

For group 3 and week 54, there seems to be a typographical error (*10*) which is noted in Table 21; presumably, the number of animals at risk is 166, not 116, and we have used 166 in the analysis below. The data are collapsed as shown in Table 22: for group 1, the number of tumors appearing at weeks 40, 42, . . . , 50 is 0 + t . . . + 2 = 8, and the number at risk is 130 + 128 + . . . + 121 = 753. The other entries in Table 22 are similar.

The counts in the pooled cells are modeled as binomial, the number of trials being the sum of the number of animals at risk in the basic cells, as shown in Table 22, and the probability of an event is the average of the probabilities given by Equation 9, weighted by the numbers at risk; time t is measured in weeks, and s = 2 (a fortnight is 2 weeks). This procedure seems to give a reasonable approximation to a sum of independent binomials, since the multistage probabilities do not change much from fortnight to fortnight. If X_i are independent binomials with N_i trials and event probability p_i, and the p_i are not too variable, then $X_1 + X_2$ is approximately binomial with $N_1 + N_2$ trials and event probability $p = (N_1p_1 + N_2p_2) / (N_1 + N_2)$.

In this study, there was only one dose group; the units are chosen so dose = 1, and then dose ^{2} = dose. The C and D terms in Equations 1–3 collapse, so only A, B, C + D can be estimated. Since AD = BC, the separate coefficients C and D can be determined if all are positive.

The results are shown in Table 23 and suggest A = B = 0, so C and D are not estimated separately. The numerical algorithm we used would not maximize the likelihood function with the constraint that all coefficients be nonnegative, so negative values are allowed. The parameter estimates are so close to the boundary that the Fisher standard errors are not reliable; however, standard errors can be obtained by the bootstrap.

The model has been fitted to only one dose group and predicts no background, for A and B are negligible. This is impressive. The interpretation of zeros: If A = B = D

Table 21. Data from Peto et al. (*10*) for weeks 40 to 58.

Weeks of benzpyrene	Group 1 (BP from age 10 weeks)	Group 2 (BP from age 25 weeks)	Group 3 (BP from age 40 weeks)	Group 4 (BP from age 55 weeks)
40	0/130	0/154	1/292	1/311
42	1/128	1/154	0/190	1/303
44	1/126	0/150	1/187	3/296
46	2/125	2/149	0/182	2/286
48	2/123	5/146	1/181	2/275
50	2/121	0/141	0/176	3/265
52	6/119	4/141	2/172	1/255
54	2/113	3/137	1/116[a]	6/239
56	3/111	4/130	4/162	9/222
58	10/108	4/126	4/154	10/203

[a]This is a probable typographical error: The number of mice at risk cannot drop from 172 to 116 and then go back up to 162.

Table 22. Peto mice: Observed skin cancers/numbers at risk.

Weeks of benzpyrene	Age at start of exposure, weeks			
	10	25	40	55
40-50	8/753	8/894	3/1108	12/1736
52-54	8/232	7/278	3/338[a]	7/494
56-58	13/219	8/256	8/316	19/425
60-62	13/187	16/229	12/286	26/345
64-66	12/162	15/188	20/239	18/243
68-70	16/132	27/150	16/184	21/175
72-74	15/95	14/96	15/139	13/118
76-78	5/68	13/67	16/93	7/76
80-82	7/51	6/43	7/65	8/36
84-88	9/31	11/36	12/58	2/33

[a]A probable error has been corrected.

Table 23. Results for the Peto mice.[a]

n	A	B	C+D	χ^2	Comment
5	−3	−4	251	77	Does not fit
6	14	−18	350	67	Does not fit, $p = 2/1000$
7	34	−35	481	81	Does not fit
8	68	−67	651	117	Does not fit

[a]There are 37 degrees of freedom in the χ^2. The hazard is per week, and the coefficients should be divided by 10^{2n+1}.

= 0 and C > 0, the first stage is the only sensitive one and the only one with no background rate. If A = B = C = 0 and D > 0, then the first and next-to-last stages are sensitive but have no background rates; the other stages have positive background rates. If A = B = 0 but C > 0 and D > 0, then the first and next-to-last stages are sensitive; all stages but the first have positive background rates (*6*). If the first and next-to-last stages are both sensitive, the dose response would be quadratic: Peto et al. were concerned whether their results are compatible with quadratic dose response (*10*).

Partly on biological grounds by mainly on statistical ones, Peto et al. adopted a 28-week latency period between the transition to the cancerous state for a cell and the appearance of a tumor. As a result, they fit a hazard rate of the form (duration − 28)^3, that is, a four-stage model. We do not use the lag, and find a six-stage model gives the best fit, with a $\chi^2 = 67$ on 37 degrees of freedom; the lagged model fits a little better with $\chi^2 = 60$. Results might be cross-validated on data in Lee and O'Neill (*74*).

The large values of χ^2 in Table 23 are mainly due to two or three cells, where the differences between observed and expected are substantial. This could have been an artifact of the aggregation. So we reaggregated, making some effort to eliminate the discrepancies. (The second pooling: 40–52, 54–56, 58–60, 62–64, 66–68, 70–72, 74–76, 78–80, 82–84, 86–88.) The estimated coefficients stayed about the same, but the χ^2 only dropped imperceptibly, from 67 to 64. The incidence rates are sufficiently irregular that we stopped trying to fit models.

Peto et al. are trying to show that cancer results from the duration of exposure to the carcinogen, rather than the effect of time *per se*. The experiment and the associated arguments are interesting but hardly conclusive, even setting aside the question of whether the model fits the data.

In the multistage framework, the rate of progression through the stages depends on dose but not time and that seems to be a critical point in the argument. If, for example, the rate of progression through a stage really was time-dependent, Peto et al. might agree that time *per se* plays a role in carcinogenesis. But the stages in the model are not experimentally identified; so we can produce a model where one stage has a time-dependent hazard rate, and the overall hazard rate is exactly the same as in the best multistage model. In short, if the stages are not identifiable, neither is the duration versus age issue. To finish the argument we put up the multistage model and the alternative.

The best-fitting multistage model (with no latency, A = B = ∅) is of the form

$$(\text{C dose} + \text{D dose}^2)\ \text{duration}^5$$

This corresponds to a six-stage model, in which the background rate for the first stage vanishes; the first stage and possibly the fifth are sensitive; hazard rates for the six stages do not depend on time. For now, call this the "C - D model."

Consider next an alternative model with four stages: The first stage has the same hazard rate as the first stage in the C - D model; the third stage has the same hazard rate as the fifth stage in the C - D model; the fourth stage has the same hazard rate as the sixth stage in the C - D model; the remaining second stage has a time-dependent hazard rate, h (t) = constant t^2.

In this alternative model, a cell starts to age only after it has been moved out of stage one by the benzpyrene, which is not so far-fetched, given that Peto et al. are studying a tumor that does not occur spontaneously. The hazard rate in the second stage is time-dependent, so time *per se* plays a role in carcinogenesis. In short, no argument about the effect of time itself seems likely to succeed until the stages are better defined.

The alternative model may seem artificial, but no more so than the multistage model itself. The construction may also prompt the question, Why should the hazard rate be time-dependent? However, insisting on multistage models with hazard rates be depending only on dose is simply to decide the question of age versus duration on an *a priori* basis.

Technical Details

Fisher Standard Errors

Let L $(X|\Theta)$ be the likelihood of the data X given the vector of parameters Θ. The Fisher information matrix I(Θ) is defined as

$$I(\Theta) = -E\left\{\frac{\partial^2 \log L(X \mid \Theta)}{\partial \Theta^2}\right\} \qquad [16a]$$

The expected value is computed with respect to Θ. Asymptotically, the inverse of I(Θ) gives the variance-covariance matrix of the MLE, at Θ. Ordinarily, the MLE $\hat{\Theta}$ will be substituted for Θ, to get a sample-based estimate. The Fisher SEs are the square roots of the diagonal elements of $I(\Theta)^{-1}$. In particular, these can be computed from the data; the unknown parameter is not involved. Observed information may also be used.

Regularity conditions are given in Lehmann (*36*), for example, and exclude cases where Θ falls on the boundary, corresponding to 0's for B or C in the present application. For some positive results on the boundary, see Freedman and Navidi (*6*).

The matrix I(Θ) is nonnegative definite, but not necessarily positive definite. With the mega-mouse liver data, at the MLE $\hat{\Theta}$, we found $I(\hat{\Theta})$ to be rank-deficient, suggesting a singular distribution for $\hat{\Theta}$.

Computational Details

Most of the computer work was done in FORTRAN on a VAX 750, with many of the calculations replicated somewhat independently in True BASIC on an IBM PC-XT. A few were replicated quite independently by Duncan Thomas at USC, but this does not imply that he agrees (or disagrees) with our conclusions.

To find the maximum of the log likelihood function, we used a computer routine written by NAG (Numerical Algorithm Group). This starts searching from a given initial point; it either reports failure to converge or finds the maxium. Usually, as best we can tell, it does find the global maximum; occasionally, it is fooled by a local maximum.

The algorithm was started from several points to see if there were multiple maxima and derivatives of the likelihood function were checked at each reported value to make sure this was at least a local maximum. In almost all the data sets described above, the algorithm found only one value, which we believe to be the global maximum. There was an exception: in fitting all 56 cells of the megamouse liver data, NAG's first pick was a saddle point on the line $r_1 = r_2$.

We thank K. Brown, J. Cairns, M. Eaton, B. Efron, L. Garfinkel, P. Lee, S. Moolgavkar, D. Petitti, M. Pike, J. Robins, S. Swan, D. Thomas, and A. Whittemore for their help. The American Cancer Society, the National Cancer Institute, and the National Center for Toxicological Research kindly provided data. Our research was partially supported by NSF grant DMS 86-01634.

REFERENCES

1. Armitage, P., and Doll, R. Stochastic models for carcinogenesis. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV (L. LeCam and J. Neyman, Eds.), University of California Press, Berkeley, CA 1961, pp. 19–38.
2. U.S. Environmental Protection Agency. Guidelines for carcinogenic risk assessment. Fed. Reg. 51: 33992–34003 (1986).
3. Borszonyi, M., Day, N. E., Lapis, K., and Yamasaki, H., Eds. Models, Mechanisms and Etiology of Tumour Promotion. International Agency for Research on Cancer, Scientific Publication, No. 56, Lyon, France, 1984.
4. Whittemore, A., and Keller, J. Quantitative theories of carcinogenesis. SIAM Rev. 20: 1–30 (1978).
5. Kalbfleisch, J. D., Krewski, D. R., and van Ryzin, J. Dose-response models for time-to-response toxicity data. Can. J. Stat. 11: 25–45 (1983).
6. Freedman, D., and Navidi, W. On the Multistage Model for Carcinogenesis. Technical Report No. 97, Department of Statistics, University of California, Berkeley, CA, 1987.
7. Moolgavkar, S. H. The multistage theory of carcinogenesis and the age distribution of cancer in man. J. Natl. Cancer Inst. 61: 49–52 (1978).
8. Cook, P. J., Doll, R., and Fellingham, S. A. A mathematical model for the age distribution of cancer in man. Intl. J. Cancer 4: 93–112 (1969).
9. Doll, R., and Peto, R. Cigarette smoking and bronchial carcinoma among regular smokers and lifelong non-smokers. J. Epidemiol. and Commun. Health 32: 303–313 (1978).
10. Peto, R., Roe, F. J. C., Lee, P. N., Levy, L., and Clack, J. Cancer and ageing in mice and men. Br. J. Cancer 32: 411–426 (1975).
11. Peto, J., Seidmann, H., and Selikoff, I. J. Mesothelioma mortality in asbestos workers: Implications for models of carcinogenesis and risk assessment. Br. J. Cancer 45: 124–135 (1982).
12. Brown, K. G., and Hoel, D. G. Modeling time-to-tumor data: Analysis of the ED01 study. Fundam. Appl. Toxicol. 3: 458–469 (1983).
13. Brown, K. G., and Hoel, D. G. Multistage prediction of cancer in serially dosed mice with application to the ED01 study. Fundam. Appl. Toxicol. 3: 470–477 (1983).
14. Ames, B. N., Magaw R., and Gold, L. Ranking possible carcinogenic hazards. Science 236: 271–279 (1987).
15. Bishop, J. M. The molecular genetics of cancer. Science 235: 305–311 (1987).
16. Freedman, D., and Zeisel, H. From mouse to man: The quantitative assessment of cancer risks. Statist. Sci. 3: 3–56 (1988).
17. Kaldor, J. M., and Day, N. E. Interpretation of Epidemiological Studies in the Context of the Multistage Model of Carcinogenesis. Technical Report, International Agency for Research on Cancer, Lyon, 1986.
18. Moolgavkar, S. H. Carcinogenesis modeling: From molecular biology to epidemiology. Annu. Rev. Public Health 7: 151–169 (1986).
19. Bodmer, W. F., Bailey, C. J., Bodmer, J., Bussey, H. J. R., Ellis, A., Gorman, P., Lucibello, F. C., Murday, V. A., Rider, S. H., Scambler, P., Sheer, D., Solomon, E., and Spurr, N. K. Localization of the gene for familial, adenomatous polyposis on chromosome 5. Nature 328: 614–616 (1987).
20. Naylor, S. L., Johnson, B. E., Minna, J. D., and Sakaguchi, A. Y. Loss of heterozygosity of chromosome 3p markers in small-cell lung cancer. Nature 329: 451–454 (1987).
21. Solomon, E., Voss, R., Hall, V., Bodmer, W. F., Jass, J. R., Jeffreys, A. J., Lucibello, F. C., Patel, I., and Rider, S. H. Chromosome 5 allele loss in human colorectal carcinomas. Nature 328: 616–619 (1987).
22. Brown, C. C., and Chu, K. Use of multistage models to infer stage affected by carcinogenic exposure: Example of lung cancer and cigarette smoking. J. Chron. Dis. 40 (suppl.) 2: 171S–179S (1987).
23. Kahn, H. A. The Dorn study of smoking and mortality among US veterans: Report on eight and one-half years of observation. In: Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases (W. Haenszel, Ed.), National Cancer Institute Monograph No. 19, Washington, D.C. 1969, pp. 1–126.
24. Rogot, E. Smoking and mortality among US veterans. J. Chron. Dis. 27: 189–203 (1974).
25. Whittemore, A. Effect of cigarette smoking in epidemiological studies of lung cancer. Stat. Med. 7: 223–238 (1988).
26. Hammond, E. C. Smoking in relation to the death rates of one million men and women. In: Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases (W. Haenszel, Ed.), National Cancer Institute Monograph No. 19, Washington, D. C. 1969, pp. 127–204.
27. Staffa, J. A., and Mehlman, M. A., Eds. Innovations in Cancer Risk Assessment (ED01 study). J. Environ. Pathol. Toxicol. 3: 1–246 (1980).
28. Littlefield, N. A., Farmer, J. H., Gaylor, D. W., and Sheldon, W. G. Effects of dose and time in a long-term, low dose carcinogenic study. J. Environ. Pathol. Toxicol. 3: 17–34 (1980).
29. Farmer, J. H., Kodell, R. L., Greenman, D. L., and Shaw, G. W. Dose and time response models for the incidence of bladder and liver neoplasms in mice fed 2-acetylaminofluorene continuously. J. Environ. Pathol. Toxicol. 3: 55–68 (1980).
30. Carlborg, F. W. Multistage dose response models in carcinogenesis. Fd. Cosmet. Toxicol. 19: 361–365 (1981).
31. Carlborg, F. W. 2-Acetylaminofluorene and the Weibull model. Fd. Cosmet. Toxicol. 19: 367–371 (1981).
32. Peto, R., Parish, S. E., and Gray, R. G. There is no such thing as ageing, and cancer is not related to it. In: Age-Related Factors in Carcinogenesis (A. Likhachev, V. Ansimov, and R. Montesano, Eds.), International Agency for Research on Cancer, Scientific Publication No. 58, Lyon, France, 1986, pp. 43–53.
33. Peto, R., Gray, R., Brantom, P., and Grasso, P. Nitrosamine carcinogenesis in 5120 rodents. In: N-Nitroso Compounds: Occurrence, Biological Effects and Relevance to Human Cancer (J. K. O'Neill, R. C. von Borstel, C. T. Miller, J. Long, and H. Bartsch, Eds.), International Agency for Research on Cancer, Scientific Publication No. 57, Lyon, 1984, pp. 627–665.
34. Drew, R. T., Boorman, G. A., Haseman, J. K., McConnell, E. E., Bosey, W. M., and Moore, J. A. The effect of age and exposure duration on cancer induction by a known carcinogen in rats, mice, and

hamsters. Toxicol. Appl. Pharmacol. 68: 120–130 (1983).
35. Ames, B. N. Measuring Oxidative Damage in Humans: Relation to Cancer and Aging. Technical Report, Department of Biochemistry, University of California, Berkeley, CA.
36. Lehmann, E. Theory of Point Estimation. Wiley and Sons, New York, 1983.
37. Lehmann, E. Testing Statistical Hypotheses, 2nd ed. Wiley and Sons, New York, 1986.
38. Kalbfleisch, J. D., and Prentice, R. L. The Statistical Analysis of Failure Time Data. Wiley and Sons, New York, 1980.
39. Rao, C. R. Linear Statistical Inference and its Applications, 2nd ed. Wiley and Sons, New York, 1973.
40. Pearl, R. Tobacco smoking and longevity. Science 87: 216–217 (1938).
41. Stevens, R. G., and Moolgavkar, S. H. A cohort analysis of lung cancer and smoking in British males. Am. J. Epidemiol. 119: 624–641 (1984).
42. Peto, R. Epidemiology, multistage models, and short-term mutagenicity tests. In: Origins of Human Cancer, Book C (H. H. Hiatt, J. D. Watson, and J. A. Winsten, Eds.), Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1977, pp. 1403–1428.
43. Doll, R., and Peto, R. The causes of cancer. J. Natl. Cancer Inst. 66: 1191–1309 (1981).
44. Wald, N.J., and Doll, R., Eds. Interpretation of Negative Epidemiological Evidence for Carcinogenicity. International Agency for Research On Cancer Scientific Publication No. 65, Lyon, 1985.
45. Armitage, P. Multistage models of carcinogenesis. Environ. Health Perspect. 63: 195–201 (1985).
46. Lave, L. B. Health and safety risk analyses: Information for better decisions. Science 236: 291–295 (1987).
47. Vouk, V. B., Hoel, D. G., Butler, G. C., Peakall, D. B., Eds. Methods for Estimating Risk of Chemical Injury: Human and Non-Human Biota and Ecosystems. Wiley and Sons, New York, 1985.
48. Zeise, L., Wilson, R., and Crouch, E. A. C. Dose response relationships for carcinogens: A review. Environ. Health Perspect. 73: 259–308 (1987).
49. U.S. Public Health Service. Saccharin and its salts: Proposed rule making. Report of the Food and Drug Administration. Fed. Reg. 42: 19996–20010 (1977).
50. National Academy of Sciences. Saccharin: Technical Assessment of Risks and Benefits. Washington, DC, 1978.
51. Aalen, O. O. Nonparametric inference for a family of counting processes. Ann. Stat. 6: 701–726 (1978).
52. Jacobsen, M. Statistical Analysis of Counting Processes. Springer-Verlag, New York, 1982.
53. Moolgavkar, S. H., and Venzon, D. J. Confidence regions in curved exponential families: Applications to matched case-control and survival studies with general relative risk function. Ann. Stat. 15: 346–359 (1987).
54. Moolgavkar, S. H., and Venzon, D. J. Curvature, Bias and Diagnostics: Application to Relative Risk Regression Models. Technical Report, Fred Hutchinson Cancer Research Center, Seattle, WA, 1987.
55. International Agency for Research on Cancer. Tobacco Smoking. Monograph 38. International Agency for Research on Cancer, Lyon, France, 1986.
56. Peto, J. Early- and late-stage carcinogenesis in mouse skin and man. In: Models, Mechanisms and Etiology of Tumour Promotion (M. Borszonyi, N. E. Day, K. Lapis, and H. Yamasaki, Eds.), International Agency for Research on Cancer, Scientific Publication No. 56, Lyon, France, 1984, pp. 359–372.
57. Doll, R. Practical steps towards the prevention of bronchial carcinoma. Scot. Med. J. 15: 433–447 (1970).
58. U.S. Public Health Service. Smoking and Health. Report of the advisory committee to the Surgeon General of the Public Health Service. Department of Health, Education and Welfare, Washington, DC, 1979.
59. Wynder, E. L., and Hecht, S., Eds. Lung Cancer. UICC Technical Report Series, Vol. 25, Geneva, 1976.
60. Doll, R., and Peto, R. Mortality in relation to smoking: 20 years observations on male British doctors. Br. Med. J. 2: 1525–1536 (1976).
61. Lubin, J. H., Blot, W. J., Berrino, F., Flamant, R., Gillis, C. R., Kunze, M., Schmahl, D., and Visco, G. Modifying risk of developing lung cancer by changing habits of cigarette smoking. Br. Med. J. 288: 1953–1956 (1984).
62. McKenzie, I., and Rous, P. The experimental disclosure of latent neoplastic changes in tarred skin. J. Exp. Med. 73: 391–416 (1941).
63. Hennings, H., Shores, R., Wenk, M. L., Spangler, E. F., Tarone, R., and Yuspa, S. H. Malignant conversion of mouse skin tumours is increased by tumour initiators and unaffected by tumour promoters. Nature 304: 67–68 (1983).
64. Farber, E., and Sarma, D. S. R. Hepatocarcinogenesis: A dynamic cellular perspective. Lab. Invest. 56: 4–22 (1987).
65. Freedman, D., and Navidi, W. Ex-Smoker and the Multistage Model for Lung Cancer. Technical Report No. 135, Department of Statistics, University of California, Berkeley, CA 1988.
66. Gaffney, M., and Altshuler, B. An examination of the role of cigarette smoke in lung carcinogenesis using multistage models. Technical report. JNCI 80: 925–931 (1988).
67. Geddes, D. M. The natural history of lung cancer: A review based on rates of tumor growth. Br. J. Dis. Chest 73: 1–17 (1979).
68. Weiss, W. Implications of tumor growth for the natural history of lung cancer. J. Occup. Med. 26: 345–352 (1984).
69. Tsiatis, A. A nonidentifiability aspect of the problem of competing risks. Proc. Natl. Acad. Sci. (U.S.) 72: 20–22 (1975).
70. Frome, G., and Checkoway, H. Use of Poisson regression models in estimating incidence rates and ratios. Am. J. Epidemiol. 121: 309–23 (1985).
71. Day, N. E., and Brown, C. C. Multistage models and primary prevention of cancer. J. Natl. Cancer Inst. 64: 977–89 (1980).
72. Kodell, R., Farmer, J. H., and Greenman, D. L. Estimation of distribution of time to appearance of tumor and time to death from tumor after appearance of tumor in mice fed 2-acetylaminofluorene. J. Environ. Pathol. Toxicol. 3: 89–102 (1980).
73. Malani, H., and van Ryzin, J. Nonparametric Estimates of the Tumor Incidence Rate, the Tumor Lethality Rate and the Mortality Rate in Absence of Tumors in Animal Carcinogenicity Experiments. Technical Report No. B56. Division of Biostatistics, Columbia University, New York, 1986.
74. Lee, P. N., and O'Neill, J. A. The effect both of time and dose applied on tumour incidence rate in benzopyrene skin painting experiments. Br. J. Cancer 25: 759–770 (1971).